Proceedings of the 33rd Pacific Asia Conference on Language, Information and Computation

Edited by
Ryo Otoguro, Mamoru Komachi and Tomoko Ohkuma

September 13–15, 2019
Future University Hakodate

ISSN 2619-7782

Published by Waseda Institute for the Study of Language and Information, Waseda University, Tokyo, Japan

Acknowledgments

PACLIC 33 is hosted by Future University Hakodate in conjunction with The Japan Association for the Study of Logic, Language and Information and supported by Waseda University Comprehensive Research Organization, The Association for Natural Language Processing, The Japan Society for Artificial Intelligence and Hakodate City.

Foreword

It is our great pleasure and honor to hold the 33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33) at Future University Hakodate, Japan. "Open space, Open mind" is the underlying philosophy of Future University Hakodate, which makes it exactly the right place to host PACLIC. Following the long tradition of PACLIC, PACLIC 33 also emphasizes the synergy of theoretical analysis and processing of natural language. PACLIC 33 aims to enhance the interaction between researchers working in different fields of language study in the Asia-Pacific region as well as around the world.

We received 132 submissions from around the world including Algeria, Brazil, Chile, China, Czechia, France, Hong Kong, India, Indonesia, Iran, Italy, Japan, Macao, Mexico, Philippines, Singapore, South Korea, Taiwan, Thailand, Tunisia, Ukraine, the UK, the US and Vietnam. Out of 132 papers, 37 were accepted for oral presentations and 26 for poster presentations. The acceptance rate for oral presentations and poster presentations are 28% and 20% respectively.

In addition to oral and poster presentations, the conference highlights four keynote talks, one special invited talk and one satellite workshop. We are grateful to Justine Cassell from Carnegie Mellon University, Mary Dalrymple from University of Oxford, Yuji Matsumoto from Nara Institute of Science and Technology/Riken AIP and Junichi Tsujii from National Institute of Advanced Industrial Science and Technology for accepting to give a keynote talk. Jong-Bok Kim from Kyung Hee University has kindly agreed to give an invited talk in commemoration of the Humboldt Research Award given to him in 2019. We also thank Yasunari Harada, Chu-Ren Huang, Jong-Bok Kim, Yasuhiro Katagiri and Miwa Morishita for organizing the 27th Joint Workshop on Linguistics and Lanuage Processing during the conference and Rachel Edita O. Roxas and Manolito V. Octaviano for giving an invited talk in the workshop.

PACLIC 33 would not be made possible without the support from many people. We would like to express our sincere gratitude toward program committee members and sub-reviewers whose professional reviews allowed us to maintain the high quality standard of PACLIC. We are also deeply indebted to the local organizing committee at Future University Hakodate: Yasuhiro Katagiri, Hitoshi Matsubara, Hajime Murai, Asuka Terai, Misako Nambu, Kaoru Sumi and Ayahiko Niimi as well as student staff members. We would also like to thank Waseda University Comprehensive Research Organization, The Association for Natural Language Processing, The Japan Society for Artificial Intelligence and Hakodate City for their generous financial support for the conference.

Ryo Otoguro
Mamoru Komachi
Tomoko Ohkuma

PACLIC 33 Program Committee Chairs

Organizers

Steering Committee Standing Members

Chu-Ren Huang, The Hong Kong Polytechnic University, Hong Kong
Jong-Bok Kim, Kyung Hee University, Seoul
Ryo Otoguro, Waseda University, Tokyo
Rachel Edita O. Roxas, National University, Manila
Maosong Sun, Tsinghua University, Beijing
Benjamin T'sou, City University of Hong Kong, Hong Kong
Min Zhang, Soochow University, Suzhou

Organizing Committee

Ryo Otoguro, Waseda University (Chair)
Yasunari Harada, Waseda University (Co-chair)
Yasuhiro Katagiri, Future University Hakodate (Honorary chair)

Local Organizing Committee

Yasuhiro Katagiri, Future University Hakodate (Chair)
Hitoshi Matsubara, Future University Hakodate (Co-chair)
Hajime Murai, Future University Hakodate
Asuka Terai, Future University Hakodate
Misako Nambu, Future University Hakodate
Kaoru Sumi, Future University Hakodate
Ayahiko Niimi, Future University Hakodate

Program Committee

Chairs

Ryo Otoguro, Waseda University
Mamoru Komachi, Tokyo Metropolitan University
Tomoko Ohkuma, Fuji Xerox Co.

Members

Laurence Anthony
Alice Mae Arbon
Masayuki Asahara
Qian Chen
Doris Chen
Charibeth Cheng
Emmanuele Chersoni
Sung-Kwon Choi
Jin-Woo Chung
Li Dong
Helena Gao
Yasunari Harada
Hitomi Hirayama
Jeffrey J. Holliday
Munpyo Hong
Shu-Kai Hsieh
Jiangping Huang
Suyeon Im
Tomoyuki Kajiwara
Daisuke Kawahara
Jong-Bok Kim
Ji-Hye Kim
Kanako Komiya
Valia Kordoni
Yusuke Kubota
Oi Yee Kwong
Huei-Ling Lai
Olivia Lam
Yong-Hun Lee
Albert Lee
Sang-Im Lee-Kim
Baoli Li
Wei-Wen Liao
Dongsik Lim
Jingxia Lin
Te-Hsin Liu
Bingquan Liu
Yunfei Long
Lu Lu
Chen Lyu
Erlyn Manguilimotan
Yuji Matsumoto
Koji Mineshima
Yasuhide Miura
Ponrudee Netisopakul
Takashi Ninomiya
Hitoshi Nishikawa
Nathaniel Oco
Kenji Oda
Ethel Ong
Chutamanee Onsuwan
Yohei Oseki
David Yoshikazu Oshima
Jong C. Park
Nattama Pongpairoj
Haoliang Qi
Tao Qian
Rodolfo Jr Raga
Yafeng Ren
Hiroyuki Shinnou
Shu-Ing Shyu
Melanie Siegel
Pornsiri Singhapreecha
Leif Romeritch Syliongka
Zhiyang Teng
Yuen-Hsien Tseng
Takehito Utsuro
Zhongqing Wang
Tak-Sum Wong
Hongzhi Xu
Yun Xue
Jie Yang
Cheng-Zen Yang
Satoru Yokoyama
Minoru Yoshida
Liang-Chih Yu
Longtu Zhang
Meishan Zhang
Jiajun Zhang
Chengzhi Zhang

Additional reviewers

Kristine Mae Adlaon
Yoshihiko Asao
Eun Jin Chun
Masahiro Kaneko
Wakako Kashino
Yuan Ling
Rui Liu
Shutian Ma
Hiroki Narita
Mizuki Sango
Helen Villanueva
Hiroaki Yamada
Hayahide Yamagishi
Qingqing Zhou

Table of Contents

Poster papers

A Gold Standard Dependency Treebank for Indonesian

Ika Alfina, Arawinda Dinakaramani,
Mohamad Ivan Fanany, and Heru Suhartanto
Faculty of Computer Science, Universitas Indonesia
Depok, Indonesia
ika.alfina@cs.ui.ac.id, arawinda.dinakaramani@ui.ac.id,
{ivan, heru}@cs.ui.ac.id

Abstract

Resources for syntactic parsing for Indonesian are very limited, as there are only two dependency treebanks publicly available and both are small in size. Not only that, we found out that the word segmentation method used by both treebanks needs improvement. Therefore, in this work we proposed a revision for one of these treebanks, Indonesian Parallel Universal Dependencies treebank. Besides improving word segmentation, we also improved POS tagging and syntactic annotations. Because in Indonesian grammar there are some special structures, we also proposed how to adjust UDv2 annotation guidelines with those Indonesian grammar rules. To evaluate the quality of the new treebank, we built Indonesian dependency parser model using Parsito (UDPipe) parser. Using ten-fold cross-validation, the model that built using the revised treebank had UAS of 83.33% and LAS of 79.39%, over the original treebank with UAS of 73.32% and LAS of 65.98%.

1 Introduction

Indonesian is a language spoken by more than 260 million people in 2019, but its resources for Natural Language Processing (NLP) research are still limited. Especially for the syntactic parsing studies, the availability of syntactic corpora (treebank) is scarce.

As far as we know, there are only two dependency treebanks for Indonesian that available publicly. Both are provided by the Universal Dependencies (UD)[1]. The first one is UD Indonesian-GSD (McDonald et al., 2013) that consists of 5,593 sentences, and the second one is UD Indonesian-PUD (Zeman et al., 2018) that consists of 1,000 sentences.

Unfortunately, after conducting reviews to the quality of both treebanks, we found out major flaws, especially in the word segmentation that does not comply with Indonesian grammar. In UD Indonesian-GSD, all clitics are not separated from the main words. While in UD Indonesian-PUD, words with clitics are always be split in any context and the reduplicated or hyphenated words that occur frequently in Indonesian are always separated into multiple tokens.

In this work, we proposed a revision to the UD Indonesian-PUD since its size is smaller than the UD Indonesian-GSD. Meanwhile, we also observerd there are some characteristics of Indonesian grammar that needs special treatments. UD created guidelines for cross-linguistically grammatical annotation. The current version of the annotation guidelines is named Universal Dependencies v2 (UDv2). To address special constructions in Indonesian grammar, we proposed some adjustments to UDv2 guidelines.

The contributions of our work are two folds. First, we proposed some adjustments to UDv2 annotation guidelines to build dependency treebank for Indonesian. Specifically, we proposed special treatments in word segmentation and POS tagging process for Indonesian and proposed the use of some dependency relations for certain language constructions in Indonesian grammar. Second, we proposed a revision to the UD Indonesian-PUD treebank of 1,000 sentences, resulting in a better gold standard depen-

[1] https://universaldependencies.org/

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 1-9, Hakodate, Japan, September 13-15, 2019

dency treebank for Indonesian. This revised treebank had been made public[2].

The rest of this paper is organized as follows: Section 2 addresses the Indonesian grammar; Section 3 describes the Indonesian PUD treebank; Section 4 explains the proposed annotation guidelines; Section 5 describes the annotation procedure and the statistics of the revised treebank; Section 6 discusses the experiments and results, and finally, Section 7 presents the conclusions and future work.

2 Indonesian Grammar

In this section, we discuss some Indonesian grammar rules that are relevant to our work in revising the UD Indonesian-PUD treebank.

2.1 Reduplicated Words

Some words in Indonesian are formed using reduplication (Sneddon et al., 2010). For example, the plural nouns such as *anak-anak (children)*, singular nouns such as *arak-arakan (procession)*, verbs such as *merobek-robek (shredding)*, adjectives such as *hiruk-pikuk (noisy)*, and adverbs such as *terus-menerus (continuously)*.

This characteristic implies that in the word segmentation process, this kind of words should not be split into multiple tokens.

2.2 Indonesian Clitics

A clitic is "*a morpheme in morphology and syntax that has syntactic characteristics of a word, but depends phonologically on another word or phrase*"[3]. Indonesian has two kinds of clitic: 1) As a personal pronoun; and 2) as a particle (Alwi et al., 1998). The clitics of personal pronoun are *ku- (I), -ku (me/my), kau- (you), -mu (you/your)*, and *-nya (him/her/it)*, and of the particle are *-lah, -kah*, and *-tah*.

As a clitic has a syntactic role, in the word segmentation process, we need to separate them from the main word. Furthermore, Larasati (2012) reported that by handling the clitic, the accuracy of an Indonesian to English machine translation system was improved.

Most of Indonesian clitics of personal pronoun have an ambiguous nature, especially *-nya*. Table 1 shows examples of words that ended with *-nya*. On this table, we use the UDv2 part-of-speech (POS) label[4] to abbreviate the word class. For each word, the syntax of how the word is formed and the actual POS are presented.

Word	Syntax	Actual POS
bukunya (her/his book)	NOUN + *-nya*	NOUN + DET
bukunya (the book)	NOUN + *-nya*	NOUN + DET
akhirnya (finally)	NOUN + *-nya*	ADV
khususnya (especially)	ADJ + *-nya*	ADV
jauhnya (the distance)	ADJ + *-nya*	NOUN
cantiknya (very beautiful)	ADJ + *-nya*	ADJ + ADV
sebenarnya (actually)	se + ADJ + *-nya*	ADV
dibukanya (open by him/her)	VERB + *-nya*	VERB + PRON
dibukanya (the opening)	VERB + *-nya*	NOUN

Table 1: Examples of the ambiguity of *-nya*.

We can see that the syntax of "NOUN + *-nya*" has three possible interpretations: "*-nya*" as the possessive determiner; "*-nya*" as the determiner; and "*-nya*" changes a NOUN into an ADV. While the syntax of "ADJ + *-nya*" has three possible meaning: "*-nya*" could change an ADJ into an ADV or a NOUN, or the meaning of *-nya* becomes "very" (ADV). The syntax of "*se* + ADJ + *-nya*" forms an ADV. The last syntax, "VERB + *-nya*" has two possible interpretations: "*-nya*" as the PRON following the VERB, or "*-nya*" changes a VERB into a NOUN. The use of "*-nya*" to change an ADJ or VERB into a NOUN is called the predicate nominalization (Sneddon et al., 2010, p311).

This shows how challenging it is to decide whether a token that ended with "*-nya*" should be split or not in the word segmentation process. It requires the information of POS tags of some words around the token with "*-nya*".

2.3 Compound Words

A compound is *"a combination of two simple words which come together to form a complex word"* (Sneddon et al., 2010). In Indonesian there are three ways to write the compound words: 1) as a single token, such as *kacamata (eyeglasses), matahari (sun)*; 2) hyphenated, such as *pemuda-pemudi*

[2]https://github.com/ialfina/revised-id-pud/

[3]https://en.wikipedia.org/wiki/Clitic

[4]https://universaldependencies.org/u/pos/index.html

(youngsters); and 3) as two tokens, such as *sapu tangan (handkerchief)*. Beside of noun compound, there are also compound words of verb, adjective, and so on. Table 2 shows some examples of Indonesian compound words.

POS	Examples
NOUN	*tanggung jawab (responsibility)*
VERB	*bertanggung jawab (to be responsible)*
ADJ	*luar biasa (excellent)*
ADV	*sering kali (often), kadang kala (sometime)*
NUM	*salah satu (one, a/an)*
DET	*salah seorang (a/an, for person)*
SCONJ	*di mana (where)*

Table 2: Examples of compound words in Indonesian.

Since a compound word has a syntactic role, we suggest that the compound words that have already written as a single token or hyphenated do not need to be split in the word segmentation process. While compound words written as two words need special treatment so that the relation between those two words is retained.

2.4 Noun Phrases in Indonesian

A noun phrase is *"a sequence of words which functions in the same way as a noun"* (Sneddon et al., 2010). A noun phrase always contains a noun as its head. There are two kinds of dependency direction, either head-initial or head-final (Hawkins, 1990). While English usually uses head-final direction for noun phrase, Indonesian mostly uses head-initial with some exceptions (Alwi et al., 1998). Table 3 shows some syntactic constructions of Indonesian noun phrases and their respective head directionality.

Type	Direction
NOUN + Demonstrative DET	head-initial
Quantity DET + NOUN	head-final
NOUN + Possessive DET	head-initial
NOUN + ADJ	head-initial
NOUN/PROPN + NOUN/PROPN	head-initial

Table 3: Head directionality of several types of noun phrase in Indonesian.

The following are some examples of Indonesian noun phrases:
1) *buku ini (this book)*
2) *dua mahasiswa (two students)*
3) *beberapa masalah (some problems)*
4) *rumah baru (new house)*
5) *rumah sakit (hospital)*
6) *rumahku (my house)*
7) *ekor anjing (the dog's tail/tail of the dog)*
8) *rumah Ika (Ika's house)*
9) *pemilik toko (a store owner/the owner of store)*
10) *sepatu Nike (the Nike shoes)*

Example 1-3 are noun phrases with a determiner (DET). Example 1 uses a demonstrative determiner. While in English this kind of determiner is placed before the noun, in Indonesian it is written after the noun. Example 2 dan 3 use quantity determiner that is either a number or words that describe number such as *beberapa (some/several), semua (all)*. This kind of noun phrase has the same syntax with English, a determiner is written before the noun.

Example 4 is a noun phrase with an adjective as the second word that describes the first word. Example 5 is a noun compound that we discuss in Subsection 2.3 with the syntax of "NOUN + NOUN".

Example 6-8 are noun phrases that show ownership. Example 6 uses *-ku* clitic as the possessive pronoun. Example 7-8 use "NOUN + NOUN/PROPN" syntax where the second token is the owner of the first token. While in English the ownership is marked by the *'s* clitic or using the *of* preposition, there's no such syntax in Indonesian.

Example 9-10 are noun phrases with the same syntax with Example 7-8, but with different semantics. In these phrases, the second word is not the owner of the first word, but only describes it. To differentiate between those two kinds of "NOUN + NOUN/PROPN" phrases require the knowledge of whether it is possible for the second word to 'own' the first word, that makes this task challenging.

3 Indonesian Parallel Universal Dependencies (PUD) Treebank

In this section, we discuss the treebank being revised, the UD Indonesian-PUD.

3.1 Universal Dependencies

Universal Dependencies (UD) is a framework for cross-linguistically grammatical annotation for dependency treebank. Initially, de Marneffe et al. (2006) designed type dependency for English that later was called Stanford Dependencies. Stanford

Dependencies scheme was designed to represent English grammatical relations between words in a sentence (de Marneffe and Manning, 2008). This representation was later adopted by de Marneffe et al. (2014) to create universal dependencies that can be applied to other languages to support cross-linguistically parsing. This new scheme was named Universal Dependencies (UD).

The first version of UD annotation guidelines was called UDv1 (Nivre et al., 2016). The recent version of the annotation guidelines is UDv2, that has tagset of 17 POS tags and has 37 dependency relations plus some dependency relation subtypes to be used by certain languages to adapt to UD.

3.2 Parallel Multilingual Treebanks

Parallel Universal Dependencies (PUD) treebanks created for *CoNLL 2017 share task for Multilingual Parsing from Raw Text to Universal Dependencies* (Zeman et al., 2018). They created parallel treebanks for 18 different languages. Each treebank consists of 1,000 sentences, in the same order. The sources of the sentences are from news domain and Wikipedia. The original language of the first 750 sentences is English, and the rest are German, French, Italian and Spanish.

3.3 Indonesian-PUD Treebank

UD Indonesian-PUD (hereinafter referred to as ID-PUD) is part of PUD. We observed that the ID-PUD has some problems, where the major flaws are its word segmentation and POS tagging.

For the word segmentation, the list of error is as follows: 1) The words with reduplication are always be split into multiple tokens; 2) Other hyphenated words are also always separated into multiple tokens; 3) The clitic *-nya* is always separated from its parent word, despite its context; 4) Many tokens that are composed of two base words such as *ketidaksesuaian (the discordance)* were separated into two tokens, while it should remain as one token.

For the POS tagging, a lot of tokens were incorrectly labeled, either a verb was labeled as a noun or a noun labeled as a verb. We suspect this problem happened because the tool determined the POS tag based on the base-word of the verb. If the base-word is a noun, then the verb is labeled as a noun, which is incorrect.

4 Adjusting UDv2 for Indonesian

This section presents our proposed annotation guidelines for the specific characteristics of Indonesian grammar.

4.1 Word Segmentation and POS Tagging as an Inseparable Task

For most word segmentation cases, all clitics should be separated from its parent word. What makes this task difficult is that for the clitic of *-nya* there are cases when it should not be separated, as explained in Subsection 2.2. For example, we have two sentences contains the word *dibukanya*:
a) ***Dibukanya*** *toko itu menimbulkan kemacetan. (The opening of the store caused a traffic jam.)*
b) *Paket itu* ***dibukanya*** *dengan hati-hati. (The package was opened (by her/him) carefully.*
For sentence (a), *dibukanya* should not be separated since it has a role as a NOUN, and for sentence (b), the token *dibukanya* should be split since it contains two syntactic token, *dibuka (was opened)* as a VERB and *-nya (him/her)* as a PRON.

To decide whether we will split a token ended with *-nya*, we proposed this general approach: 1) Split the token ended with *-nya* into two parts, the parent token and *-nya* ; 2) Determine the POS tag of the parent token; 3) Use Table 1 as the reference to solve the ambiguity by using the POS tag of tokens before or after the examined token; and 4) Finally, if the final POS tag of the examined token is a NOUN or an ADV, re-merge the parent token and *-nya*. We leave the details of this approach for future work.

Thus, the word segmentation task needs POS tags information to decide whether to split tokens ended with *-nya* or not. That's why word segmentation and POS tagging should become an inseparable task.

4.2 Adjusting Dependency Relations to Indonesian Grammar

UDv2 defined 37 dependency relation labels plus some subtypes of dependency relations to comply with special characteristics of certain languages. For current Indonesian treebanks in UD, 13 subtypes are used as shown in Table 4.

After analyzing those 13 subtypes, we proposed to retain subtypes No. 1-8 and to remove the remaining 5 subtypes. Also, we adopted 7 subtypes used by

No	Deprel	Description
1	acl:relcl	for relative clause
2	cc:preconj	for pre-conjunction
3	csubj:pass	subject clause of passive
4	nsubj:pass	subject of passive sentence
5	dep:prt	for clitic of particle
6	nmod:poss	for phrase of ownership
7	obl:tmod	noun phrase of time
8	flat:name	for named entities
9	compound:plur	for reduplicated words
10	obl:poss	for phrase of ownership
11	compound:n	for noun compound
12	compound:v	for verb compound
13	compound:a	for adjective compound

Table 4: List of subtypes used by current Indonesian treebanks in UD.

other languages and proposed the use of a new subtype. In total, we proposed the use of 16 subtypes for annotating Indonesian dependency treebank.

4.2.1 Removing five subtypes

The following is the explanation of why we propose not to use subtypes of *compound:plur*, *obl:poss*, *compound:n*, *compound:v*, and *compound:a*.

In the original ID-PUD, the reduplicated words are split into three tokens. For example, *anak-anak (children)* was split into *anak*, -, and *anak*. Subtype of *compound:plur* was created to link the third token to the first one. Since we opted not to split the reduplicated words, we no longer need this subtype.

The subtype of *obl:poss* most likely was created due to incorrect POS tagging of some nouns that labeled as verbs. For example, in ID-PUD noun phrase of *kehidupan kita (our life)* has POS tags of "VERB + PRON", while the correct POS tags should be "NOUN + DET". The correct relation between *kita (our)* to *kehidupan (life)* should be *nmod:poss*. There is no need to define *obl:poss* subtype since that case, the noun phrase with syntax of "VERB + PRON" for ownership, never exist.

UDv2 has *compound* label for noun phrases with syntax of "NOUN/PROPN + NOUN/PROPN". Since in the original ID-PUD there are noun phrases with syntax of "VERB + NOUN" such as in *bela diri (self-defense)* or "NOUN + ADJ" such as in *rumah sakit (hospital)*, a new subtype of *compound:n* was created. Since all noun phrases should have syntax shown by Table 3, we suggest to solve this problem by improving the quality of POS tagger, instead of introducing this subtype.

The subtypes of *compound:v* and *compound:a* were used for verb and adjective compound in the original ID-PUD. Table 2.3 shows that besides these two types of compound words, in Indonesian grammar there are also compound of adverb, number, determiner, and subordinating conjunction.

Because the number of compound words other than nouns is limited, we proposed that the compound words of verb, adjective, adverb, number, and determiner to be represented by only one single label. Since in English treebank ***compound:prt*** subtype was used for verb compound, we proposed the used of that label for those five compound word types in Indonesian grammar. As for the compound word of subordinating conjunction (SCONJ) that can be regarded as the function word, we proposed to use *fixed* label as suggested by UDv2 guidelines.

4.2.2 Adopting other six subtypes from treebanks of other languages

Besides adopting *compound:prt* for compound words, we also proposed the adoption of other six subtypes defined for other languages in UDv2: 1) *flat:foreign*; 2) *flat:range*; 3) *nmod:npmod*; 4) *nmod:tmod*; 5) *obl:agent*; and 6) *obl:mod*.

In UDv2 guidelines about *flat* subtype, *flat:foreign* is used to annotate a foreign phrase that cannot be given a compositional analysis. Subtype of *flat:range* was used by Ukranian PUD treebank to label the dependent of noun phrase like *"2018-2019"* or *"8 until 10"*. We considered this annotation scheme better than the current *nummod* subtype used in the original ID-PUD for this case, since *nummod* was initially designed for noun phrase with quantity determiner, such as in *5 buku (five books)*.

In ID-PUD, noun phrases with the syntax of "NOUN/PROPN + NOUN/PROPN" are labeled as a *compound*, even if the semantics of the phrase is far away from the definition of compound discussed in Subsection 2.3. We propose to use *compound* label only for noun phrases with syntax of "NOUN + NOUN". For "NOUN + PROPN" or "PROPN + NOUN" we proposed the use of *nmod:npmod* subtype instead. For example, for phrase *ibukota Indonesia*, the word *Indonesia* was given label of *nmod:npmod*. As for phrases of

"PROPN + PROPN" the *flat* label should be used as suggested by UDv2 guidelines.

Since *obl:tmod* label has been used for noun/noun phrase related to the time that describes the predicate, we propose to also use *nmod:tmod* subtype for noun/noun phrase related to time that describes the noun, such as in *laporan 2019* (the report of 2019) where 2019 (the year) describes the noun of *laporan (report)*.

In Alwi et al. (1998), it stated the passive sentence does not have an object but could have a noun/noun phrase that represents the agent. For this purpose, there are two possible labels. If it is a noun phrase with preposition we used *obl* label, but if the agent is written without preposition, *obl:agent* subtype will be used.

Subtype of *obl:mod* is initially used by French treebank for nominal adjunct of a predicate. We want to adopt this label for noun/noun phrase without preposition that describes the predicate but not the object nor the agent of the predicate. For example in the sentence *"Bunga bank naik 1%" (Bank interest rose 5%)*, the token % will be given *obl:mod* label.

4.2.3 Proposing a new subtype

We observed that some adverbs in Indonesian can be formed with the syntax of "*secara/dengan (with)* + ADJ/VERB/NOUN". Examples of such adverbs are *secara bijaksana (wisely), dengan bersemangat (excitedly), dengan setara (equally)*.

According to UDv2, since *secara* or *dengan* are the prepositions, their POS tag is ADP if followed by a noun phrase or SCONJ if followed by a clause. In syntactic parsing, the token with ADP tag will be labeled with *case* label and the token with SCONJ tag with *mark* label.

On the other hand, the syntax of "*secara/dengan* + ADJ/VERB/NOUN" in forming an adverb in Indonesian grammar needs special treatment. We proposed a new label named ***case:adv*** for *secara/dengan* and for the ADJ/VERB/NOUN following them, its POS tag need to be changed to ADV so that we can label them as *advmod*. It will be the responsibility of the POS tagger to identify this kind of adverb in sentences and to modify the POS tag of the related words.

Figure 1 shows an example of how a dependency tree has changed. A reduplicated word *saudara-saudara (folks)* was split into three tokens in the original treebank, but remain as one token in the revised treebank. Additionally, we also revised the POS tag of word *yang (that)* and changed the subject of this sentence.

5 Revising the Indonesian PUD Treebank

In this section, we present the annotation procedure and the statistics of the revised treebank.

5.1 Annotation Procedure

The revision was done in 2 stages: 1) Revising the word segmentation and POS tags; 2) Revising the dependencies. Both stages were done by two annotators, with the background of computer science and Indonesian linguistics. The total time for learning the UDv2 annotation guidelines, proposing the adjustment for Indonesian grammar and conducting the revision of ID-PUD treebank was six months.

For each stage, the revising was done in two phases: the learning phase and the revision phase. On the learning phase, each annotator was given 50 first sentences of ID-PUD to be analyzed. On the meeting, the annotators discussed what should be done in revising the treebank by referring to UDv2 guidelines and references of Indonesian grammar. After both annotators agree on all issues, the revision phase was started. The process was done iteratively. If there was a new case found in the revision phase, the annotators were back to the learning phase and update the guidelines. After that, the revision phase was resumed.

5.2 Statistics of the Revised Treebank

Table 5 shows the comparison of token distribution in the original and revised treebanks, as the effect of changes in the word segmentation process. Since in the revised UD-PUD, we did not split the reduplicated words and a lot of other hyphenated words, the number of tokens is smaller compared to the original ID-PUD. Likewise, the average number of tokens in the sentence becomes smaller, and the number of unique tokens increased.

Table 6 shows the comparison of UPOS distribution in the original and revised treebank. It shows that major revision had been done in the POS tagging process. For example, there is no SCONJ and

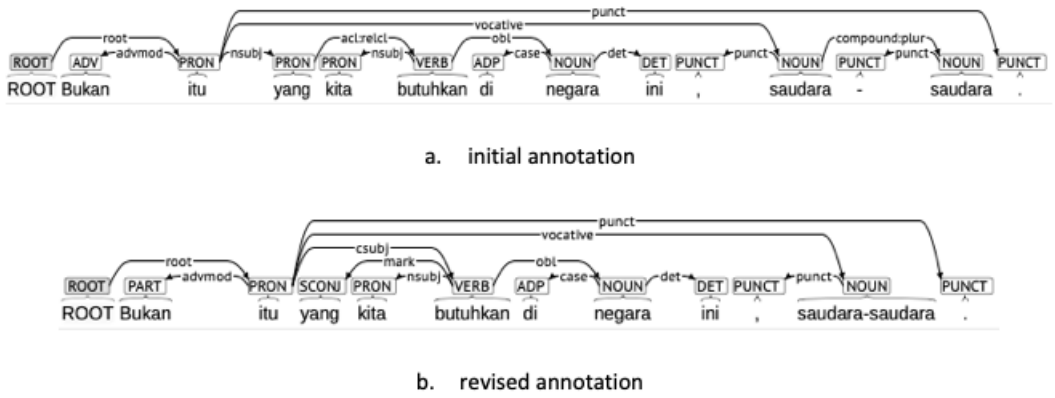


Figure 1: The initial and revised annotation for a sentence of *"Bukan itu yang kita butuhkan di negara ini, saudara-saudara." (That's not what we need in our country, folks.)*

Description	Original	Revised
Number of tokens	19,900	19,401
Avg number of tokens in sentence	19.9	19.4
Number of unique tokens	4,692	4,732

Table 5: Statistics of number of tokens.

INTJ in the original treebank, but in the revised one we have 487 occurrences of SCONJ and 4 occurrences of INTJ.

UPOS	Ori	Rev	UPOS	Ori	Rev
ADJ	1358	962	PART	57	276
ADP	2832	1901	PRON	989	1049
ADV	1049	623	PROPN	1456	2217
AUX	211	424	PUNCT	2579	2384
CCONJ	612	595	SCONJ	0	487
DET	522	940	SYM	37	39
INTJ	0	4	VERB	1965	2359
NOUN	5578	4618	X	86	17
NUM	569	506			

Table 6: The UPOS distribution.

As for the dependency relation labels, in the new treebank, we only use 32 of 37 UDv2's main labels and 15 of 16 subtypes described in Subsection 4.2. UDv2's main labels that were not used are *clf, dep, expl, list*, and *reparandum*, while the subtype not used is *flat:name*. In total, the revised ID-PUD used 47 labels.

We decided not to use *flat:name* for names and used *flat* for all proper noun instead since it's still not clear for us which names are suitable for *flat:name*. Once we know how to differentiate between *flat* and *flat:names* we will revise the treebank.

6 Experiments and Results

To evaluate the quality of the revised ID-PUD, we built the Indonesian parser model using Parsito (UDPipe) that built by Straka et al. (2015). Parsito is a transition-based parser that utilized neural network classifier for prediction and requires no feature engineering. We used this parser with default parameter.

Accuracy was evaluated using the ten-fold cross-validation method. The performance measurements used are UAS (Unlabeled Attachment Scores) and LAS (Labeled Attachment Score) (Kübler et al., 2009). Table 7 shows the comparison of accuracy between the original and revised ID-PUD treebank.

Treebank	UAS	LAS
Original	73.32%	65.98%
Revised	**83.33%**	**79.39%**

Table 7: Experiment results.

The result shows that the model built by our revised treebank has higher UAS and LAS than the original one, with a margin of 10% for UAS and around 13% for LAS. It shows that the revised treebank has better consistency in annotation so that the learning algorithms can learn the pattern better than when using the original one.

To find out which labels had achieved good accuracy, we used MaltEval (Nilsson and Nivre, 2008) to compute the F1-score of 47 labels used in the revised treebank and shown the result in Figure 2.

We had a hypothesis that there is a correlation between F1-score and the number of occurrences of

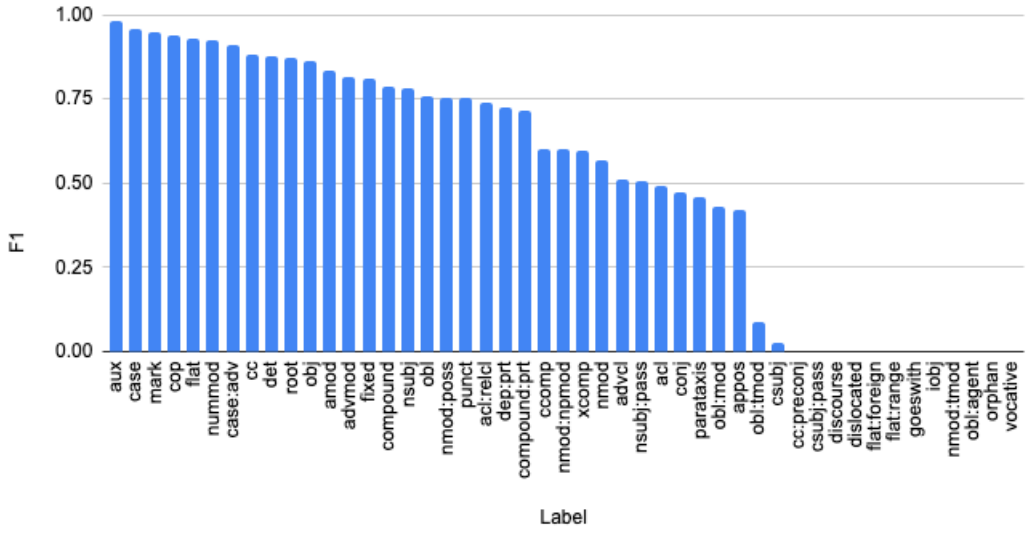


Figure 2: The F1-score of dependency labels in the revised ID-PUD.

each label, but that isn't true because the correlation coefficient is only 56%. For example, *case:adv* that occurs only 57 times has F1-score of 91%, while *conj* which occurs 666 times, has F1-score of 47%. We suggest that the low F-score was caused due to the lack of consistent patterns for those labels.

However, for those labels that occurs only 10 times or less, we believed that the F1-score can be improved by increasing the size of the treebank and adding more examples with those labels.

To improve the accuracy of the Indonesian parser model, we have three suggestions: 1) to revisit the choices of the dependency labels so that each label was designed with distinct characteristics; 2) to revisit the annotation whether the rules had been applied consistently; and 3) to employ additional morphology features that have not been added to this revised ID-PUD treebank.

7 Conclusions and Future Work

We proposed a revision to an existing dependency treebank in Indonesian, named ID-PUD that consists of 1,000 sentences. The annotation was done manually, refers to UDv2 annotation guidelines and the references of Indonesian grammar. Besides, we also proposed how to conduct word segmentation and POS tagging for Indonesian sentences, especially related to the handling of *-nya*. Some changes in dependency labels for Indonesian dependency treebank are also proposed, resulting in the 16 subtypes to adjust to Indonesian grammar rules.

To evaluate the quality of the new treebank, we used Parsito (UDPipe) parser to build the parser model using 10-fold cross-validation method. The results show that the model built using the revised treebank has a higher UAS and LAS with the margin of more than 10% than the original ID-PUD. This shows that the new treebank has a better label consistency compared to the original one.

This manual revision of ID-PUD took so much time and efforts. In future work, we want to build tools to automate the word segmentation and POS tagging described in this paper. Besides that, adding morphology features needs to be done so that this treebank has the same attributes with other parallel treebanks in PUD.

Acknowledgments

This work was supported by the research grant of PTUPT (Penelitian Terapan Unggulan Perguruan Tinggi) No. NKB-1693/UN2.R3.1/HKP.05.00/2019 from the Ministry of Research and Technology, Republic of Indonesia.

References

Hasan Alwi, Soenjono Dardjowidjojo, Hans Lapoliwa, and Anton M. Moeliono. 1998. *Tata Bahasa Baku Bahasa Indonesia*.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The Stanford typed dependencies representation. In *Coling 2008: Proceedings of the workshop on Cross-Framework and Cross-Domain Parser Evaluation - CrossParser '08*.

Marie-Catherine de Marneffe, Bill MacCartney, and Christopher D. Manning. 2006. Generating Typed Dependency Parses from Phrase Structure Parses. In *LREC 2006*.

Marie-Catherine de Marneffe, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D Manning. 2014. Universal Stanford Dependencies : A cross-linguistic typology. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*.

John A Hawkins. 1990. A Parsing Theory of Word Order Universals. *Linguistic Inquiry*, 21(2):223–261.

Sandra Kübler, Ryan McDonald, and Joakim Nivre. 2009. *Dependency Parsing*. Morgan & Claypool.

Septina Dian Larasati. 2012. Handling Indonesian Clitics: A Dataset Comparison for an Indonesian-English Statistical Machine Translation System. In *26th Pacific Asia Conference on Language, Information and Computation*, pages 146–152.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbachbrundage, Yoav Goldberg, Dipanjan Das, Kusman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Tackstrom, Claudia Bedini, Nuria Bertomeu Castello, and Jungmee Lee. 2013. Universal Dependency Annotation for Multilingual Parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 92–97.

Jens Nilsson and Joakim Nivre. 2008. MaltEval : An Evaluation and Visualization Tool for Dependency Parsing. In *LREC 2008*, pages 161–166.

Joakim Nivre, Marie-catherine De Marneffe Filip, Ginter Yoav, Jan Haji, D Manning Ryan, Mcdonald Slav, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. pages 1659–1666.

James Neil Sneddon, Alexander Adelaar, Dwi Noverini Djenar, and Michael C. Ewing. 2010. *Indonesian Reference Grammar*. A&U Academic.

Milan Straka, Jan Hajič, Jana Straková, and Jr. Jan Hajič. 2015. Parsing Universal Dependency Treebanks using Neural Networks and Search-Based Oracle. In *14th International Workshop on Treebanks and Linguistics Theories (TLT 14)*, pages 208–220.

Daniel Zeman, Jan Haji, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 Shared Task : Multilingual Parsing from Raw Text to Universal Dependencies. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–21.

Investigating an Effective Character-level Embedding in Korean Sentence Classification

Won Ik Cho, Seok Min Kim, and Nam Soo Kim
Human Interface Laboratory
Department of Electrical and Computer Engineering and INMC
Seoul National University
1 Gwanak-ro, Gwanak-gu, Seoul, Korea, 08826
{wicho,smkim}@hi.snu.ac.kr, nkim@snu.ac.kr

Abstract

Different from the writing systems of many Romance and Germanic languages, some languages or language families show complex conjunct forms in character composition. For such cases where the conjuncts consist of the components representing consonant(s) and vowel, various character encoding schemes can be adopted beyond merely making up a one-hot vector. However, there has been little work done on intra-language comparison regarding performances using each representation. In this study, utilizing the Korean language which is character-rich and agglutinative, we investigate an encoding scheme that is the most effective among *Jamo*[1]-level one-hot, character-level one-hot, character-level dense, and character-level multi-hot. Classification performance with each scheme is evaluated on two corpora: one on binary sentiment analysis of movie reviews, and the other on multi-class identification of intention types. The result displays that the character-level features show higher performance in general, although the *Jamo*-level features may show compatibility with the attention-based models if guaranteed adequate parameter set size.

1 Introduction

Ever since an early approach exploiting the character features for the neural network-based natural language processing (NLP) (Zhang et al., 2015), character-level embedding[2] has been widely used for many tasks such as machine translation (Ling et al., 2015), noisy document representation (Dhingra et al., 2016), language correction (Xie et al., 2016), and word segmentation (Cho et al., 2018a). However, little consideration was done for intra-language performance comparison regarding variant representation types. Unlike English, a Germanic language written with an alphabet comprising 26 characters, many languages used in East Asia are written with scripts whose characters can be further decomposed into sub-parts representing individual consonants or vowels. This conveys that (sub-)character-level representation for such languages has the potential to be managed with more than just a simple one-hot encoding.

In this paper, a comparative experiment is conducted on Korean, a representative language with a featural writing system (Daniels and Bright, 1996). To be specific, the Korean alphabet *Hangul* consists of the letters *Jamo* denoting consonants and vowels. The letters comprise a morpho-syllabic block that refers to *character*, which is resultingly equivalent to the phonetic unit *syllable* in terms of Korean morpho-phonology. The conjunct form of a character is {Syllable: CV(C)}; this notation implies that there should be at least one consonant (namely *cho-seng*, the first sound) and one vowel (namely *cwung-seng*, the second sound) in a character. An additional consonant (namely *cong-seng*, the third sound) is auxiliary. However, in decomposition of the characters, three slots are fully held to guarantee a space for each component; an empty cell comes in place of the third entry if there is no auxiliary consonant. The number of possible sub-characters, or (compos-

[1] Letters of Korean alphabet *Hangul*.

[2] Throughout this paper, the terms *embedding* and *encoding* are parallelly used depending on the context.

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 10-18, Hakodate, Japan, September 13-15, 2019

ite) consonants/vowels, that can come for each slot is 19, 21, and 27. For instance, in a syllable '간 (*kan*)', the three clock-wisely arranged characters ㄱ, ㅏ, and ㄴ, which sound *kiyek* (stands for *k*; among 19 candidates), *ah* (stands for *a*; among 21 candidates), and *niun* (stands for *n*; among 27 candidates), refers to the *first*, the *second* and the *third* sound respectively.

To compare five different *Jamo*/character-level embedding methodologies that are possible in Korean, we first review the related studies and the previous approaches. Then, two datasets are introduced, namely binary sentiment classification and multi-class intention identification, to investigate the performance of each representation under recurrent neural network (RNN)-based analysis. After searching for an effective encoding scheme, we demonstrate how the result can be adopted in combating other tasks and discuss if a similar approach can be applied to the languages with the complex writing system.

2 Related Work

Inter-language comparison with word and character embedding was thoroughly investigated in Zhang and LeCun (2017), for Chinese, Japanese, Korean, and English. The paper investigates the languages via representations including *character, byte, romanized character, morpheme*, and *romanized morpheme*. The observation of tendency for Korean suggests that adopting the raw characters outperforms utilizing the romanized character-level features, and moreover both the performance are far beyond the morpheme-level features. However, to be specific on the experiment, decomposition of the morpho-syllabic blocks was not conducted, and the experiment did not make use of the dense embedding methodologies which can project the distributive semantics onto the representation. We concluded that more attention is to be paid to different character embedding methodologies of Korean. Here, to reduce ambiguity, we denote a morpho-syllabic block which consists of consonant(s) and a vowel by *character*, and the individual components by *Jamo*. A *Jamo* sequence is spread in the order of the *first* to the *third* sound if a *character* is decomposed.

There has been little study done on an effective text encoding scheme for Korean, a language that has distinguished character structure which can be decomposed into sub-characters. A comparative study on the hierarchical constituents of Korean morpho-syntax was first organized in Lee and Sohn (2016), in the way of comparing the performance of *Jamo, character, morpheme*, and *eojeol (word)*-level embeddings for the task of text reconstruction. In the quantitative analysis using edit distance and accuracy, the *Jamo*-level feature showed a slightly better result than the character-level one. The (sub-)character-level representations presented the outcome far better than the morpheme or *eojeol*-level cases, as in the classification task of Zhang and LeCun (2017). The results show the task-independent competitiveness of the character-level features.

In a more comprehensive viewpoint, Stratos (2017) showed that *Jamo*-level features combined with word and character-level ones display better performance with the parsing task. With more elaborate character processing, especially involving *Jamo*s, Shin et al. (2017) and Cho et al. (2018c) made progress recently in the classification tasks. Song et al. (2018) aggregated the sparse features into multi-hot representation successfully, enhancing the output within the task of error correction. In a slightly different manner, Cho et al. (2018a) applied dense vectors for the representation of the characters, obtained by skip-gram (Mikolov et al., 2013), improving the naturalness of word segmentation for noisy Korean text. To figure out the tendency, we implement the aforementioned *Jamo*/character-level features and discuss the result concerning classification tasks. The details on each approach are to be described in the following section.

3 Experiment

In this section, we demonstrate the featurization of five (sub-)character embedding methodologies, namely (i) ***Jamo*** (Shin et al., 2017; Stratos, 2017) (ii) **modified *Jamo*** (Cho et al., 2018c), (iii) **sparse *character* vectors**, (iv) **dense *character* vectors** (Cho et al., 2018a) trained based on fastText (Bojanowski et al., 2016), and (v) **multi-hot *character* vectors** (Song et al., 2018). We featurize only *Jamo*/character and no other symbols such as numbers and special letters is taken into account.

	Representation	Property	Dimension	Feature type
(i) *Shin2017*	ㄱ ··· ㅎ / ㅏ ··· ㅢ / ㄱ ··· ㉥	*Jamo*-level	67	one-hot
(ii) *Cho2018c*	(i) + ㄱ ··· ㅏ ··· ㉥	*Jamo*-level	118	one-hot
(iii) *Cho2018a*-Sparse	··· 간 ··· 밤 ··· 햇 ···	character-level	2,534	one-hot
(iv) *Cho2018a*-Dense	··· 간 ··· 밤 ··· 햇 ···	character-level	100	dense
(v) *Song2018*	··· 간 ··· 밤 ··· 햇 ··· + α	character-level	67	multi-hot

Table 1: A description on the *Jamo*/character-level features (i-v).

For (i), we used one-hot vectors of dimension 67 (= 19 + 21 + 27), which is smaller in width than the ones suggested in Shin et al. (2017) and Stratos (2017), due to the omission of special symbols. Similarly, for (ii), 118-dim one-hot vectors are constructed. The different point of (ii) regarding (i) is that it considers the cases that *Jamo* is used in the form of single (or composite) consonant or vowel, as frequently observed in the social media text. The cases make up an additional array of dimension 51.

For (iii) and (iv), we adopted a vector set that is distributed publicly in Cho et al. (2018a), reported to be extracted from a drama script corpus of size 2M. Constructing the vectors of (iii) is intuitive; for N characters in the corpus, a N-dimensional one-hot vector is assigned for each. Case of (iv) can be considered awkward in the sense of using characters as a meaningful token, but we concluded that the Korean characters can be handled as a word piece[3] (Sennrich et al., 2015) or subword n-gram (Bojanowski et al., 2016) concerning the nature of their composition. All the characters are reported to be treated as a separate token (subword) in the training phase that uses skip-gram (Mikolov et al., 2013). Although the number of possible character combinations in Korean is precisely 11,172 (= 19 * 21 * 28), the number of ones that are used in real-life reaches about 2,500 (Kwon et al., 1995). Since the description says that the corpus is removed with punctuation and consists of around 2,500 Korean syllables, we exploited the dictionary of 100-dim fastText-embedded vectors which is provided in the paper, and extracted the list of the characters to construct a one-hot vector dictionary[4].

[3]The word piece models were not investigated in this study since here we concentrate on the (sub-)character-level embeddings.

[4]Two types of embeddings were omitted, namely the *Jamo*-based fastText and the 11,172-dim one-hot vectors; the former was considered inefficient since there are only 118 symbols at most and the latter was assumed to require a huge computation.

(v) is a hybrid of *Jamo* and character-level features; three vectors indicating the first to the third sound of a character, namely the ones with dimension 19, 21, and 27 each, are concatenated into a single multi-hot vector. This preserves the conciseness of the *Jamo*-level one-hot encodings and also maintains the characteristics of conjunct forms. In summary, (i) utilizes 67-dim one-hot vectors, (ii) 118-dim one-hot vectors, (iii) 2,534-dim one-hot vectors, (iv) 100-dim dense vectors, and (v) 67-dim multi-hot vectors (Table 1).

3.1 Task description

For evaluation, we employed two classification tasks that can be conducted with the character-level embeddings. Due to a shortage of reliable open source data for Korean, we selected the datasets that show a clear description of the annotation. One, a sentiment analysis corpus, is expected to display how well each character-level encoding scheme conveys the information regarding lexical semantics. The other, an intention analysis corpus, is expected to show how comprehensively each character-level encoding scheme deals with the syntax-semantic task that concerns sentence form and content. The details on each corpus are stated below.

3.1.1 Naver sentiment movie corpus

The corpus NSMC[5] is a widely used benchmark for evaluation of Korean language models. The annotation follows Maas et al. (2011) and incorporates 150K:50K documents for the train and test set each. The authors assign a positive label for the reviews with a score > 8 and negative for the ones with a score < 5 (in 10-scale), adopting a binary labeling system. To prevent confusion that comes from gray-zone data, neutral reviews were removed. The posi-

[5]https://github.com/e9t/nsmc

tive and negative instances are equally distributed in both train and test set.

3.1.2 Intonation-aided intention identification for Korean

The corpus 3i4K[6] (Cho et al., 2018b) is a recently distributed open-source data for multi-class intention identification. The labels, in total seven, include *fragment* and five clear-cut cases (*statement, question, command, rhetorical question (RQ), rhetorical command (RC)*). The remaining class is for the intonation-dependent utterances whose intention mainly depends on the prosody assigned to underspecified sentence enders, considering head-finality of the Korean language. Since the labels are elaborately defined and the corpus is largely hand-labeled (or hand-generated), the corpus size is relatively small (total 61K) and some classes possess a tiny volume (e.g., about 1.7K for RQs and 1.1K for RCs). However, such challenging factors of the dataset can show the aspects of the evaluation that can be overlooked. The train-test ratio is 9:1.

3.2 Feature engineering

In the first task, to manage with the document size, the length of *Jamo* or character sequence was fixed to the maximum of (i-ii) 420 and (iii-v) 140[7]. Similarly, in the second task, (i-ii) 240 and (iii-v) 80[8]. The length regarding (i-ii) being three times as long as that of (iii-v) comes from the spreadings of the sub-characters for each character.

For both tasks, the document was numericalized in the way that the tokens are placed on the right end of the feature, to preserve *Jamo*s or characters which may incorporate syntactically influential components of the phrases in a head-final language. For example, in a sentence "배고파 (pay-ko-pha, *I'm hungry*)", a vector sequence is arranged in the form of [0 0 $\cdots$ 0 0 $v1$ $v2$ $v3$], where $v1$, $v2$, and $v3$ each denotes the vector embeddings of the characters *pay*, *ko*, and *pha*. Here, *pha* encompasses the head of the phrase with the highest hierarchy in the sentence, which assigns the sentence a speech act of *statement*. The spaces between *eojeol*s were represented as zero vector(s)[9].

To look into the content of the corpora, the first dataset (NSMC) contains many incomplete characters such as solely used sub-characters (e.g., ㅋ ㅋ, ㄲㄲ) and non-Korean symbols (e.g., Chinese characters, special symbols, punctuations). The former ones were treated as characters, whereas the latter ones were ignored in all features. Although (i, iii, iv) do not represent the symbols regarding the former as non-zero vector while (ii, v) do so, we concluded that this does not threaten the fairness of the evaluation, since a wider range of representation is own advantage of each feature. The second dataset (3i4K) contains only the full characters. Thus no disturbance or biasedness was induced in the featurization.

3.3 Implementation

The implementation was done with Hangul Toolkit[10], fastText[11], and Keras (Chollet and others, 2015), which were used for character decomposition, dense vector embedding and RNN-based training, respectively. For RNN models, bidirectional long short-term memory (BiLSTM) (Schuster and Paliwal, 1997) and self-attentive sentence embedding (BiLSTM-SA) (Lin et al., 2017) were applied.

In vanilla BiLSTM, an autoregressive system that is representatively utilized for time-series analysis, a fully connected layer (FC) is connected to the last hidden layer of BiLSTM, finally inferring the output with a softmax activation. In BiLSTM with a self-attentive embedding, the context vector whose length equals to that of the hidden layers of the BiLSTM, is jointly trained along with the network so that it can provide the weight assigned to each hidden layer. The weight is obtained by making up an attention vector via a column-wise multiplication of the context vector and the hidden layers. The model specification is provided as supplementary material.

3.4 Result

For both tasks, we split the train set into 9:1 to have a separate validation set. As a result, we achieved

[6] https://github.com/warnikchow/3i4k

[7] The data description says the maximum volume of the input characters is 140.

[8] The number of the utterances with the length longer than 80 were under 40 ($< 0.07\%$).

[9] *Eojeol* denotes the unit of spacing in the written Korean.

[10] https://github.com/bluedisk/hangul-toolkit

[11] https://pypi.org/project/fasttext/

Accuracy (F1-score)	NSMC		3i4K	
	BiLSTM	*BiLSTM-SA*	*BiLSTM*	*BiLSTM-SA*
(i) ***Shin2017***	0.8203	0.8316	0.8694 (0.7443)	**0.8769** (0.7692)
(ii) ***Cho2018c***	0.7895	0.7973	0.8688 (0.7488)	0.8728 (0.7727)
(iii) ***Cho2018a*-Sparse**	**0.8271**	**0.8321**	0.8694 (0.7763)	0.8722 (0.7741)
(iv) ***Cho2018a*-Dense**	**0.8312**	**0.8382**	**0.8799** (0.7887)	**0.8844** (0.7963)
(v) ***Song2018***	**0.8271**	0.8314	**0.8696** (0.7713)	0.8761 (0.7828)

Table 2: Performance comparison. Only the accuracy is provided for NSMC since the labels are equally distributed. Two best models regarding accuracy (and F1-score for 3i4K) are bold (and underlined) for both tasks, with each architecture (BiLSTM and BiLSTM-SA).

135K instances for the training of NSMC (15K for the validation) and 50K for the training of 3i4K (5K for the validation).

3.4.1 Performance

The evaluation phase displays the pros and cons of the conventional methodologies (Table 2). In both tasks, (iv) showed significant performance. It is assumed that the result comes from the distinguished property of (iv); it does not break up the syllabic blocks and at the same time provides the distributional semantics to the models, in the way of employing skip-gram (Mikolov et al., 2013). (v) also performs in a similar way, by using a multi-hot encoding that assigns own role to each vector representation, displaying a compatible performance using BiLSTM in both tasks.

(iii) preserves the blocks as well, but one-hot encoding hardly gives any information on each character. It is assumable that such representation can be powerful for the dataset with a rich and balanced resource, as in NSMC, but is weak if the class volume is imbalanced, which led to an insignificant result for 3i4K. Although some compatible performance was achieved with BiLSTM, the models regarding (iii) reached saturation fast and displayed overfitting afterward, while the models with the other features showed a gradual increase in accuracy. The reason for fast saturation seems to be the limited flexibility coming from the vast parameter set size.

The unexpected point is that the models utilizing additional letters (ii) showed significant performance degeneration in NSMC task, where the solely used sub-characters (as ㅋ ㅋ implying joy or ㅠ ㅠ implying sadness) were expected to be aggregated into the featurization and yield a positive outcome. In the pilot research executed without validation set (that the model performing best with the test set was searched greedily), a comparable result as in (i) was shown. Thus, the reason for the degeneration seems to be the limitation of using a validation set, where the cutback in the training resource is inevitable[12]. Also, some solely used sub-characters might have caused the disorder in the inference of the sentiment, since not all the users employ the sentiment-related sub-characters in the same way. Supporting this observation, feature (ii) shows much less difference with (i) in 3i4K, where only the full characters are adopted.

3.4.2 Using self-attentive embedding

The advantage of using self-attentive embedding was the most emphasized in *Jamo*-level feature (i) for both tasks, and the least in (iii) (Table 2). We assume that relatively more significant improvement using (i) originates in the decomposability of the blocks. If a sequence of the blocks is decomposed into the sequence of sub-characters, the morphemes can be highlighted to provide more syntactically/semantically meaningful information to the system, especially the ones that could not have been revealed in the block-preserving environment (iii-v). For example, a Korean word '이상한 (*i-sang-han*, strange)' can be split into '이상하 (*i-sang-ha*, the root of the word)' and '-ㄴ (*-n*, a particle that makes the root an adjective)', making up the circumstances in which the presence and role of the morphemes is pointed out. This property is also reflected in the case of using the feature (ii), although the absolute score is not notable.

[12]It is highly recommended to use the cross-validation if one wants to boost the performance.

Trainable param.s & Training time	*BiLSTM*		*BiLSTM-SA*	
	Param.s	Time / epoch	Param.s	Time / epoch
(i) ***Shin2017***	34,178	13.5m	297,846	18m
(ii) ***Cho2018c***	47,234	16m	310,902	20.5m
(iii) ***Cho2018a*-Sparse**	665,730	33m	772,318	38.5m
(iv) ***Cho2018a*-Dense**	42,626	6.5m	149,214	6m
(v) ***Song2018***	34,178	6m	140,766	6m

Table 3: Computation burden for NSMC models.

3.4.3 Decomposability vs. Local semantics

The point described above is the disadvantage of character-level features (iii-v) in the sense that in such ones, characters cannot be decomposed, even for the sparse multi-hot encoding. The higher performance of (iv-v) compared to the *Jamo*-level features, which is currently displayed, can hence be explained as a result of preserving the cluster of letters. If the computation resource is sufficient so that exploiting deeper networks (e.g., Transformer (Vaswani et al., 2017) or BERT (Devlin et al., 2018)) is available, we infer that (i-ii) may also show compatible or better performance, since the modern self-attention-based mechanisms utilize the positional encodings to grasp the relation between the tokens, advanced from the location-based models we adopted. Nevertheless, it is still quite impressive that (iv) scores the highest even though the utilized dictionary does not incorporate all the available character combinations. It is suspected to be where the distributive semantics on the word pieces are engaged in.

3.4.4 Computation efficiency

In this study, we investigate only on the classification tasks. Notwithstanding they take a short amount of time for training and inference, the measurement on parameter volume and complexity is meaningful (Table 3). It is observed that (v) yields a compatible or better performance with respect to the other schemes, accompanied by less burden of computation. Besides, we argue that the multi-hot encoding (v) has a significant advantage over the rest in terms of multiple usages; it possesses both conciseness of the sub-character (*Jamo*)-level features and local semantics (although not distributional) of the character-level features. Due to these reasons, the derived models are fast in training and also have potential to be effectively used in sentence reconstruction or generation, as shown in Song et al. (2018), where applying large-dimensional one-hot encoding has been considered challenging.

4 Discussion

The primary goal of this paper is to search for a *Jamo*/character-level encoding scheme that best resolves the given task in Korean NLP. Empirically, we found out that the fastText-embedded vectors outperform the other features if provided with the same environment (model architecture). It is highly probable that the distributive semantics plays a significant role in the NLP tasks concerning syntax-semantics, at least in the feature-based approaches (Mikolov et al., 2013; Pennington et al., 2014). However, we experimentally verified that even with traditional feature-based systems, the sparse encoding schemes also perform adequately with the dense one, especially displaying computation efficiency in the multi-hot case.

At this moment, we want to emphasize that the utility of the comparison result is not only restricted to Korean, in that the introduced character encoding schemes are also available in other languages. Although the Korean writing system is unique, the Japanese language incorporates several morae (e.g., small *tsu*) that approximately correspond to the third sound (*cong-seng*) of the Korean characters, which may let the Japanese characters be encoded in a similar manner with the cases of Korean. Also, each character of the Chinese language (and *kanji* in Japanese) can be further decomposed into sub-characters (*bushu* in Japanese) that have meanings as well, as suggested in Nguyen et al. (2017) (e.g., 鯨 "whale" to 魚 "fish" and 京 "capital city").

Besides, many other languages that are used in South Asia (India), such as Telugu, Devanagari, Tamil, and Kannada, have writing system type of Abugida[13] (Daniels and Bright, 1996), the composition of consonant and vowel. The cases are not the same as Korean in view of a writing system since featural decomposition of the Abugida characters is not represented in the way of segmentation of a glyph. However, for example, instead of listing all the CV combinations, one can simplify the representation by segmenting the property of the charac-

[13] https://en.wikipedia.org/wiki/Writing_system

ter into consonant and vowel and making up a two-hot encoded vector. The similar kind of character embedding can be applied to many native Philippine languages such as Ilocano. Moreover, we believe that the argued type of featurization is robust in combating the noisy user-generated texts.

5 Conclusion

In this study, we have reviewed the five different types of (sub-)character-level embedding for a character-rich language. It is remarkable that the dense and multi-hot representation perform best given the classification tasks, and specifically, the latter one has the potential to be utilized beyond the given tasks due to its conciseness and computation efficiency. The utility of the sub-character-level features is also noteworthy in the syntax-semantic tasks that require morphological decomposition. It is expected that the overall performance tendency may provide a useful reference for the text processing of other character-rich languages with conjunct forms in the writing system, including Japanese, Chinese, and the languages of various South and Southeast Asian regions. A brief tutorial on both datasets using embedding methodologies presented in this paper is available online[14].

Acknowledgement

This research was supported by Projects for Research and Development of Police science and Technology under Center for Research and Development of Police science and Technology and Korean National Police Agency funded by the Ministry of Science, ICT and Future Planning (PA-J000001-2017-101). Also, this work was supported by the Technology Innovation Program (10076583, Development of free-running speech recognition technologies for embedded robot system) funded by the Ministry of Trade, Industry & Energy (MOTIE, Korea). The authors appreciate Yong Gyu Park for giving helpful opinions in performing validation and evaluation. After all, the authors want to send great thanks to the three anonymous reviewers for the insightful comments.

[14] https://github.com/warnikchow/kcharemb

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Won Ik Cho, Sung Jun Cheon, Woo Hyun Kang, Ji Won Kim, and Nam Soo Kim. 2018a. Real-time automatic word segmentation for user-generated text. *arXiv preprint arXiv:1810.13113*.

Won Ik Cho, Hyeon Seung Lee, Ji Won Yoon, Seok Min Kim, and Nam Soo Kim. 2018b. Speech intention understanding in a head-final language: A disambiguation utilizing intonation-dependency. *arXiv preprint arXiv:1811.04231*.

Yong Woo Cho, Gyu Su Han, and Hyuk Jun Lee. 2018c. Character level bi-directional lstm-cnn model for movie rating prediction. In *Proceedings of Korea Computer Congress 2018 [in Korean]*, pages 1009–1011.

François Chollet et al. 2015. Keras. `https://github.com/fchollet/keras`.

Peter T Daniels and William Bright. 1996. *The world's writing systems*. Oxford University Press on Demand.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bhuwan Dhingra, Zhong Zhou, Dylan Fitzpatrick, Michael Muehl, and William W Cohen. 2016. Tweet2vec: Character-based distributed representations for social media. *arXiv preprint arXiv:1605.03481*.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Hyuk-Chul Kwon, Ho-Jeong Hwang, Min-Jung Kim, and Seong-Whan Lee. 1995. Contextual postprocessing of a korean ocr system by linguistic constraints. In *icdar*, page 557. IEEE.

Jaeyeon Lee and Kyung-Ah Sohn. 2016. Comparison of decoder performance by representation for korean language in rnn encoder-decoder model. In *Proceedings of the KISS conference [in Korean]*, pages 609–611.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A structured self-attentive sentence embedding. *arXiv preprint arXiv:1703.03130*.

Wang Ling, Isabel Trancoso, Chris Dyer, and Alan W Black. 2015. Character-based neural machine translation. *arXiv preprint arXiv:1511.04586*.

Andrew L Maas, Raymond E Daly, Peter T Pham, Dan Huang, Andrew Y Ng, and Christopher Potts. 2011.

Learning word vectors for sentiment analysis. In *Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*, pages 142–150. Association for Computational Linguistics.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Viet Nguyen, Julian Brooke, and Timothy Baldwin. 2017. Sub-character neural language modelling in japanese. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 148–153.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing*, pages 1532–1543.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *IEEE Transactions on Signal Processing*, 45(11):2673–2681.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Haebin Shin, Min-Gwan Seo, and Hyeongjin Byeon. 2017. Korean alphabet level convolution neural network for text classification. In *Proceedings of Korea Computer Congress 2017 [in Korean]*, pages 587–589.

Chisung Song, Myungsoo Han, Hoon Young Cho, and Kyong-Nim Lee. 2018. Sequence-to-sequence autoencoder based korean text error correction using syllable-level multi-hot vector representation. In *Proceedings of HCLT [in Korean]*, pages 661–664.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research*, 15(1):1929–1958.

Karl Stratos. 2017. A sub-character architecture for korean language processing. *arXiv preprint arXiv:1707.06341*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, pages 5998–6008.

Ziang Xie, Anand Avati, Naveen Arivazhagan, Dan Jurafsky, and Andrew Y Ng. 2016. Neural language correction with character-based attention. *arXiv preprint arXiv:1603.09727*.

Xiang Zhang and Yann LeCun. 2017. Which encoding is the best for text classification in chinese, english, japanese and korean? *arXiv preprint arXiv:1708.02657*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.

Supplementary Material

BiLSTM and *BiLSTM-SA* model specification

Variables

- Sequence length (L) and the number of output classes (N) depend on the task. For NSMC, L = 420 for feature (i-ii) and 140 for (iii-v). For 3i4K, L=240 for feature (i-ii) and 80 for (iii-v). N equals 2 and 7 for NSMC and 3i4K, respectively.
- Character vector dimension (D) depends on the feature. For features (i-v), D equals 67, 118, 2534, 100, and 67, respectively.

BiLSTM

- Input dimension: (L, D)
- RNN Hidden layer width: 64 (=32×2)
- The width of FCN connected to the last hidden layer: 128 (Activation: *ReLU*)
- Output layer width: N (Activation: *softmax*)

BiLSTM-SA

- Input dimension: (L, D)
- The dimension of RNN hidden layer sequence output: (64 (= 32×2), L)
  >> each layer connected to FCN of width: 64 (Activation: *tanh*; equals to d_a in Lin et al. (2017)) [a]
- Auxiliary zero vector size: 64
  >> connected to FCN of width 64 (Activation: *ReLU*, Dropout (Srivastava et al., 2014): 0.3)
  >> connected to FCN of width 64 (Activation: *ReLU*) [b]
- Vector sequence [a] is column-wisely dot-multiplied with [b] to yield the layer of length L
  >> connected to an attention vector of size L (Activation: *softmax*)
  >> column-wisely multiplied to the hidden layer sequence to yield a weighted sum [c] of width 64
  >> [c] is connected to an FCN of width: 256 (Activation: *ReLU*, Dropout: 0.3) × 2 (multi-layer)
- Output layer width: N (Activation: *softmax*)

Settings

- Optimizer: Adam (Kingma and Ba, 2014) (Learning rate: 0.0005)
- Loss function: Categorical cross-entropy
- Batch size: 64 for NSMC, 16 for 3i4K (due to the difference in the corpus volume)
- For 3i4K, class weights were taken into account to compensate the volume imbalance.
- Device: Nvidia Tesla M40 24GB

Incorporating Chains of Reasoning over Knowledge Graph for Distantly Supervised Biomedical Knowledge Acquisition

Qin Dai[1], **Naoya Inoue**[1,2], **Paul Reisert**[2], **Ryo Takahashi**[1] **and Kentaro Inui**[1,2]
[1]Tohoku University, Japan
[2]RIKEN Center for Advanced Intelligence Project, Japan
{daiqin, naoya-i, preisert, ryo.t, inui}@ecei.tohoku.ac.jp

Abstract

The increased demand for structured scientific knowledge has attracted considerable attention on extracting scientific relation from the ever growing scientific publications. Distant supervision is a widely applied approach to automatically generate large amounts of labelled sentences for scientific Relation Extraction (RE). However, the brevity of the labelled sentences would hinder the performance of distantly supervised RE (DS-RE). Specifically, authors always omit the Background Knowledge (BK) that they assume is well known by readers, but would be essential for a machine to identify relationships. To address this issue, in this work, we assume that the reasoning paths between entity pairs over a knowledge graph could be utilized as BK to fill the "gaps" in text and thus facilitate DS-RE. Experimental results prove the effectiveness of the reasoning paths for DS-RE, because the proposed model that incorporates the reasoning paths achieves significant and consistent improvements as compared with a state-of-the-art DS-RE model.

1 Introduction

Scientific Knowledge Graph (KG), such as Unified Medical Language System (UMLS) [1], is extremely crucial for many scientific Natural Language Processing (NLP) tasks such as Question Answering (QA), Information Retrieval (IR) and Relation Extraction (RE). Scientific KG provides large collections of relations between entities, typically stored as (h, r, t) triplets, where h = *head entity*, r = relation and t = *tail entity*, e.g., (*acetaminophen*, *may_treat*, *pain*). However, KGs are often highly incomplete (Min et al., 2013). Scientific KGs, as with general KGs such as Freebase (Bollacker et al., 2008) and DBpedia (Lehmann et al., 2015), are far from complete and this would impede their usefulness in real-world applications. Scientific KGs, on the one hand, face the data sparsity problem. On the other hand, scientific publications have become the largest repository ever for scientific KGs and continue to increase at an unprecedented rate (Munroe, 2013). Therefore, it is an essential and fundamental task to turn the unstructured scientific publications into well organized KG, and it belongs to the task of RE.

One obstacle that is encountered when building a RE system is the generation of training instances. For coping with this difficulty, (Mintz et al., 2009) proposes distant supervision to automatically generate training samples via leveraging the alignment between KGs and texts. They assume that if two entities are connected by a relation in a KG, then all sentences that contain those entity pairs will express the relation. For instance, (*ketorolac_tromethamine*, *may_treat*, *pain*) is a fact triplet in UMLS. Distant supervision will automatically label all sentences, such as Example 1, Example 2 and Example 3, as positive instances for the relation *may_treat*. Although distant supervision could provide a large amount of training data at low cost, it always suffers from wrong labelling problem. For instance, comparing to Example 1, Example 2 and Example 3 should not be seen as the convincing evidences to support the *may_treat* relationship between *ketorolac_tromethamine* and *pain*, but will still be annotated as positive instances by the distant supervision.

[1]https://www.nlm.nih.gov/research/umls/

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 19-28, Hakodate, Japan, September 13-15, 2019

(1) *The analgesic effectiveness of **ketorolac_tromethamine** was compared with hydrocodone and acetaminophen for **pain** from an arthroscopically assisted patellar-tendon autograft anterior cruciate ligament reconstruction.*

(2) *This double-blind, split-mouth, and randomized study was aimed to compare the efficacy of dexamethasone and **ketorolac_tromethamine**, through the evaluation of **pain**, edema, and limitation of mouth_opening.*

(3) *A loading dose of parental **ketorolac_tromethamine** was administered and subjects were later given two staged doses of the same "unknown" drug with **pain** evaluations conducted after each dose.*

To automatically alleviate the wrong labelling problem, (Riedel et al., 2010; Hoffmann et al., 2011) apply multi-instance learning. In order to avoid the handcrafted features and errors propagated from NLP tools, (Zeng et al., 2015) proposes a Convolutional Neural Network (CNN), which incorporate mutli-instance learning with neural network model, and achieves significant improvement in distantly supervised RE (DS-RE). Recently, attention mechanism is applied to effectively extract features from all collected sentences, rather than from the most informative one that previous work has focused on. (Lin et al., 2016) proposes a relation vector based attention mechanism for DS-RE. (Han et al., 2018) proposes a novel joint model that leverages a KG-based attention mechanism and achieves significant improvement than (Lin et al., 2016).

Although the KG-based model outperforms several state-of-the-art DS-RE models, the brevity of textual information would inevitably hinder its performance. Specifically, authors always leave out information that they assume is known to their readers. For instance, Example 2 omits the background connection between *ketorolac_tromethamine* and *pain* and implicitly conveys that the former *may_treat* the latter. Human readers could easily make this inference based on their Background Knowledge (BK) about the target entity pair. However, for a machine, it would be extremely difficult to identify the relationship just from the given sentence without the important BK.

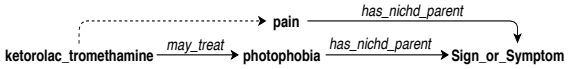


Figure 1: An example of reasoning path.

To address the issue of textual brevity, in this work, we assume that the paths (or reasoning paths) between an entity pair over a KG could be applied as the BK to fill the "gaps" and thereby improve the performance of DS-RE. For instance, one reasoning path between *ketorolac_tromethamine* and *pain* over UMLS is shown in Figure 1. By observing the path, we may infer with some likelihood that $(ketorolac\_tromethamine, may\_treat, pain)$, because *ketorolac_tromethamine* could be prescribed to treat some *Sign_or_Symptom* such as *photophobia*, and *pain* is a *Sign_or_Symptom*, therefore *ketorolac_tromethamine* might be used to treat *pain*. By comprehensively considering the path in Figure 1 and the sentence in Example 2, we could further prove the inference. To this end, we propose the DS-RE model that not only encodes the sentences containing target entity pairs, but also the reasoning paths between them over a KG.

We conduct evaluation on biomedical datasets in which KG is collected from UMLS and textual data is extracted from Medline corpus. The experimental results prove the effectiveness of the incorporation of reasoning paths for improving DS-RE from biomedical datasets.

2 Related Work

RE is a fundamental task in the NLP community. In recent years, Neural Network (NN)-based models have been the dominant approaches for non-scientific RE, which include Convolutional Neural Network (CNN)-based frameworks (Zeng et al., 2014; Xu et al., 2015; Santos et al., 2015) Recurrent Neural Network (RNN)-based frameworks (Zhang and Wang, 2015; Miwa and Bansal, 2016; Zhou et al., 2016). NN-based approaches are also used in scientific RE. For instance, (Gu et al., 2017) utilizes a CNN-based model for identifying *chemical-disease* relations from Medline corpus. (Hahn-

Powell et al., 2016) proposes an LSTM-based model for identifying *causal precedence* relationship between two event mentions in biomedical papers. (Ammar et al., 2017) applies (Miwa and Bansal, 2016)'s model for scientific RE.

Although remarkably good performances are achieved by the models mentioned above, they still train and extract relations on sentence-level and thus need a large amount of annotation data, which is expensive and time-consuming. To address this issue, distant supervision is proposed by (Mintz et al., 2009). To alleviate the noisy data from the distant supervision, many studies model DS-RE as a Multiple Instance Learning (MIL) problem (Riedel et al., 2010; Hoffmann et al., 2011; Zeng et al., 2015), in which all sentences containing a target entity pair (e.g., *ketorolac_tromethamine* and *pain*) are seen as a bag to be classified. To make full use of all the sentences in the bag, rather than just the most informative one in the bag, researchers apply attention mechanism in deep NN-based models for DS-RE. (Lin et al., 2016) proposes a relation vector based attention mechanism to extract feature from the entire bag and outperforms the prior approaches. (Du et al., 2018) proposes multi-level structured self-attention mechanism. (Han et al., 2018) proposes a joint model that adopts a KG-based attention mechanism and achieves significant improvement than (Lin et al., 2016) on DS-RE.

The attention mechanism in deep NN-based models has achieved significant progress on DS-RE. However, the brevity of input sentences could still negatively affect the performance. To address this issue, we assume that the reasoning paths between target entity pairs over a KG could be applied as BK to fill the "gaps" of input sentences and thus promote the efficiency of DS-RE. (Roller et al., 2015) uses some inference pattern learned from UMLS for eliminating potentially related entity pairs from negative training data for DS-RE. (Ji et al., 2017) applies entity descriptions generated form Freebase and Wikipedia as BK, (Lin et al., 2017) utilizes multilingual text as BK and (Vashishth et al., 2018) uses relation alias information (e.g., $founded$ and $co\text{-}founded$ are aliases for the relation $founderOfCompany$) as BK for DS-RE. However, none of these existing approaches mentioned above comprehensively consider multiple sentences containing entity pairs and multiple reasoning paths between them for DS-RE.

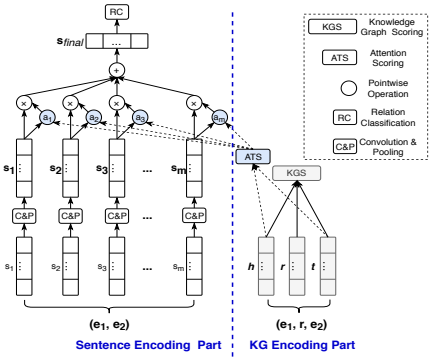


Figure 2: Overview of the base model.

3 Base Model

The success of the joint model proposed by (Han et al., 2018) inspires us to choose the strong model as our base model for biomedical DS-RE. The architecture of the base model is illustrated in Figure 2. In this section, we will introduce the base model proposed by (Han et al., 2018) in two main parts: KG Encoding part and Sentence Encoding part.

3.1 KG Encoding Part

Suppose we have a KG containing a set of fact triplets $\mathcal{O} = \{(e_1, r, e_2)\}$, where each fact triplet consists of two entities $e_1, e_2 \in \mathcal{E}$ and their relation $r \in \mathcal{R}$. Here $\mathcal{E}$ and $\mathcal{R}$ stand for the set of entities and relations respectively. KG Encoding Part then encodes $e_1, e_2 \in \mathcal{E}$ and their relation $r \in \mathcal{R}$ into low-dimensional vectors $\mathbf{h}, \mathbf{t} \in R^d$ and $\mathbf{r} \in R^d$ respectively, where d is the dimensionality of the embedding space. The base model adopts two Knowledge Graph Completion (KGC) models Prob-TransE and Prob-TransD, which are based on TransE (Bordes et al., 2013) and TransD (Ji et al., 2015) respectively, to score a given fact triplet. Specifically, given an entity pair (e_1, e_2), Prob-TransE defines its latent relation embedding $\mathbf{r}_{ht}$ via the Equation 1.

$$\mathbf{r}_{ht} = \mathbf{t} - \mathbf{h} \tag{1}$$

Prob-TransD is an extension of Prob-TransE and introduces additional mapping vectors $\mathbf{h}_p, \mathbf{t}_p \in R^d$ and $\mathbf{r}_p \in R^d$ for e_1, e_2 and r respectively. Prob-TransD encodes the latent relation embedding via

the Equation 2, where $\mathbf{M}_{rh}$ and $\mathbf{M}_{rt}$ are projection matrices for mapping entity embeddings into relation spaces.

$$\begin{aligned} \mathbf{r}_{ht} &= \mathbf{t}_r - \mathbf{h}_r, \\ \mathbf{h}_r &= \mathbf{M}_{rh}\mathbf{h}, \\ \mathbf{t}_r &= \mathbf{M}_{rt}\mathbf{t}, \\ \mathbf{M}_{rh} &= \mathbf{r}_p\mathbf{h}_p^\top + \mathbf{I}^{d\times d}, \\ \mathbf{M}_{rt} &= \mathbf{r}_p\mathbf{t}_p^\top + \mathbf{I}^{d\times d} \end{aligned} \tag{2}$$

The conditional probability can be formalized over all fact triplets $\mathcal{O}$ via the Equations 3 and 4, where $f_r(e_1, e_2)$ is the KG scoring function, which is used to evaluate the plausibility of a given fact triplet. For instance, the score for (*aspirin*, *may_treat*, *pain*) would be higher than that for (*aspirin*, *has_ingredient*, *pain*), because the former is more plausible than the latter. $\theta_\mathcal{E}$ and $\theta_\mathcal{R}$ are parameters for entities and relations respectively, b is a bias constant.

$$P(r|(e_1, e_2), \theta_\mathcal{E}, \theta_\mathcal{R}) = \frac{\exp(f_r(e_1, e_2))}{\sum_{r'\in\mathcal{R}} \exp(f_{r'}(e_1, e_2))} \tag{3}$$

$$f_r(e_1, e_2) = b - \|\mathbf{r}_{ht} - \mathbf{r}\| \tag{4}$$

3.2 Sentence Encoding Part

Sentence Representation Learning. Given a sentence s with n words $s = \{w_1, ..., w_n\}$ including a target entity pair (e_1, e_2), CNN is used to generate a distributed representation $\mathbf{s}$ for the sentence. Specifically, vector representation $\mathbf{v}_t$ for each word w_t is calculated via Equation 5, where $\mathbf{W}^w_{emb}$ is a word embedding projection matrix (Mikolov et al., 2013), $\mathbf{W}^{wp}_{emb}$ is a word position embedding projection matrix, $\mathbf{x}^w_t$ is a one-hot word representation and $\mathbf{x}^{wp}_t$ is a one-hot word position representation. The word position describes the relative distance between the current word and the target entity pair (Zeng et al., 2014). For instance, in the sentence *"Patients recorded pain$_{e_2}$ and aspirin$_{e_1}$ consumption in a daily diary"*, the relative distance of the word *"and"* is [1, -1].

$$\begin{aligned} \mathbf{v}_t &= [\mathbf{v}^w_t; \mathbf{v}^{wp1}_t; \mathbf{v}^{wp2}_t], \\ \mathbf{v}^w_t &= \mathbf{W}^w_{emb}\mathbf{x}^w_t, \\ \mathbf{v}^{wp1}_t &= \mathbf{W}^{wp}_{emb}\mathbf{x}^{wp1}_t, \\ \mathbf{v}^{wp2}_t &= \mathbf{W}^{wp}_{emb}\mathbf{x}^{wp2}_t \end{aligned} \tag{5}$$

The distributed representation $\mathbf{s}$ is formulated via the Equation 6, where, $[\mathbf{s}]_i$ and $[\mathbf{h}_t]_i$ are the i-th value of $\mathbf{s}$ and $\mathbf{h}_t$, M is the dimensionality of $\mathbf{s}$, $\mathbf{W}$ is the convolution kernal, $\mathbf{b}$ is a bias vector, and k is the convolutional window size.

$$[\mathbf{s}]_i = \max_t\{[\mathbf{h}_t]_i\},\ \forall i = 1, ..., M \tag{6}$$

$$\begin{aligned} \mathbf{h}_t &= \tanh(\mathbf{W}\mathbf{z}_t + \mathbf{b}), \\ \mathbf{z}_t &= [\mathbf{v}_{t-(k-1)/2}; ...; \mathbf{v}_{t+(k-1)/2}] \end{aligned}$$

KG-based Attention. Suppose for each fact triplet (e_1, r, e_2), there might be multiple sentences $S_r = \{s_1, ..., s_m\}$ in which each sentence contains the entity pair (e_1, e_2) and is assumed to imply the relation r, m is the size of S_r. As discussed before, the distant supervision inevitably collect noisy sentences, the base model adopts a KG-based attention mechanism to discriminate the informative sentences from the noisy ones. Specifically, the base model uses the latent relation embedding $\mathbf{r}_{ht}$ from Equation 1 (or Equation 2) as the attention over S_r to generate its final representation $\mathbf{s}_{final}$. $\mathbf{s}_{final}$ is calculated via Equation 7, where $\mathbf{W}_s$ is the weight matrix, $\mathbf{b}_s$ is the bias vector, a_i is the weight for $\mathbf{s}_i$, which is the distributed representation for the i-th sentence in S_r.

$$\mathbf{s}_{final} = \sum_{i=1}^{m} a_i\mathbf{s}_i, \tag{7}$$

$$\begin{aligned} a_i &= \frac{\exp(\langle\mathbf{r}_{ht}, \mathbf{x}_i\rangle)}{\sum_{k=1}^{m}\exp(\langle\mathbf{r}_{ht}, \mathbf{x}_k\rangle)}, \\ \mathbf{x}_i &= \tanh(\mathbf{W}_s\mathbf{s}_i + \mathbf{b}_s) \end{aligned}$$

Finally, the conditional probability $P(r|S_r, \theta_s)$ is formulated via Equation 8 and Equation 9, where, θ_S is the parameters in Sentence Encoding Part, $\mathbf{M}$ is the representation matrix of relations, $\mathbf{d}$ is a bias vector, $\mathbf{o}$ is the output vector containing the prediction probabilities of all target relations for the input sentences set S_r, and n_r is the total number of relations.

$$P(r|S_r, \theta) = \frac{\exp(\mathbf{o}_r)}{\sum_{c=1}^{n_r}\exp(\mathbf{o}_c)} \tag{8}$$

$$\mathbf{o} = \mathbf{M}\mathbf{s}_{final} + \mathbf{d} \tag{9}$$

3.3 Optimization

The base model defines the optimization function as the log-likelihood of the objective function in Equation 10.

$$P(G, D|\theta) = P(G|\theta_{\mathcal{E}}, \theta_{\mathcal{R}}) + P(D|\theta_S) \quad (10)$$

where, G and D are KG and textual data respectively. The base model applies Stochastic Gradient Descent (SGD) and L_2 regularization. In practice, the base model optimizes the KG Encoding Part and Sentence Encoding Part in parallel.

4 Proposed Model

As discussed before, the sentences containing the entity pairs of interest tend to omit the BK that the authors assume is known to the readers. However, the omitted BK would be extremely important for a machine to identify the relation between the entity pairs. To fill the "gaps" and improve the efficacy of DS-RE, we assume that the reasoning paths between the entity pairs over a KG could be utilized as BK to compensate for the brevity of the sentences. Motivated by this issue, we propose the DS-RE model that integrates both reasoning paths and sentences.

4.1 Architecture

The proposed model consists of three parts: KG Encoding Part, Sentence Encoding Part and Path Encoding Part, as shown in Figure 3. The KG Encoding Part and Sentence Encoding Part are identical to the base model, except that the final input to the relation classification layer. The Path Encoding Part takes as input a set of reasoning paths, $P_r = \{p_1, ..., p_m\}$, between two entities of interest (e_1, e_2), and encodes them into the final representation of paths, $\mathbf{p}_{final}$. Specifically, let $p = \{e_1, r_1, e_{r_1}, r_2, e_{r_2}, ..., r_i, e_{r_i}..., e_2\}$ denote a path between (e_1, e_2). To express the semantic meaning of a relation in a path, we represent r_i by its component words, rather than treat it as an unit. Therefore, a path will be represented as $p = \{e_1, w_1^{r_1}, w_2^{r_1}, ..., e_{r_1}, , w_1^{r_2}, w_2^{r_2}, ..., e_{r_2}, ..., e_2\}$, where $w_2^{r_1}$ denotes the second word of r_1 (e.g., $treat$ in may_treat relation).

Since a path is represented as a sequence of words, or a special sentence, we apply the similar CNN model used in the Sentence Encoding Part

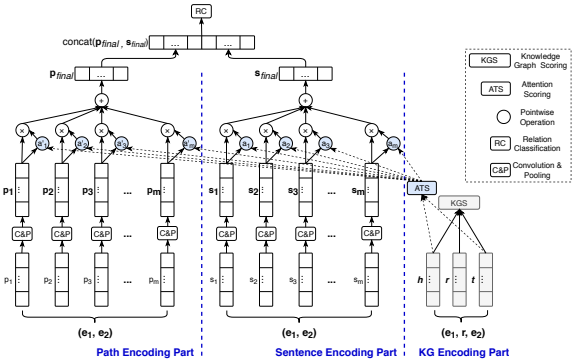


Figure 3: Overview of the proposed model.

to encode the path into vector representation $\mathbf{p}_i$. The Path Encoding Part and Sentence Encoding Part share the word embedding projection matrix $\mathbf{W}_{emb}^{w}$, and word position projection matrix $\mathbf{W}_{emb}^{wp}$ in Equation 5 except the convolutional kernal $\mathbf{W}$ and its corresponding bias vector $\mathbf{b}$ in Equation 6. To utilize evidence from all the paths between target entity pair, we also adopt the KG-based attention mechanism applied in Sentence Encoding Part to calculate the final representation of paths $\mathbf{p}_{final}$. We calculate $\mathbf{p}_{final}$ via Equation 11, where $\mathbf{W}_s$ is the weight matrix, $\mathbf{b}_s$ is the bias vector, a'_i is the weight for $\mathbf{p}_i$, which is the distributed representation for the i-th path in P_r.

$$\mathbf{p}_{final} = \sum_{i=1}^{m} a'_i \mathbf{p}_i, \quad (11)$$

$$a'_i = \frac{\exp(\langle \mathbf{r}_{ht}, \mathbf{x'}_i \rangle)}{\sum_{k=1}^{m} \exp(\langle \mathbf{r}_{ht}, \mathbf{x'}_k \rangle)},$$

$$\mathbf{x'}_i = \tanh(\mathbf{W}_s \mathbf{p}_i + \mathbf{b}_s)$$

Finally, we concatenate the resulting representation $\mathbf{s}_{final}$ and $\mathbf{p}_{final}$ for S_r (the set of input sentences) and P_r (the set of reasoning paths) respectively as the input to the relation classification layer. The conditional probability $P(r|S_r, P_r, \theta_S, \theta_P)$ is formulated via Equation 12 and Equation 13, where, θ_P is the parameters in Path Encoding Part, $\mathbf{M}$ is the representation matrix of relations, $\mathbf{d}$ is a bias vector, $\mathbf{o}$ is the output vector containing the prediction probabilities of all target relations for both input sentences set S_r and input paths set P_r. n_r is the total number of relations.

$$P(r|S_r, P_r, \theta_S, \theta_P) = \frac{\exp(\mathbf{o}_r)}{\sum_{c=1}^{n_r} \exp(\mathbf{o}_c)} \quad (12)$$

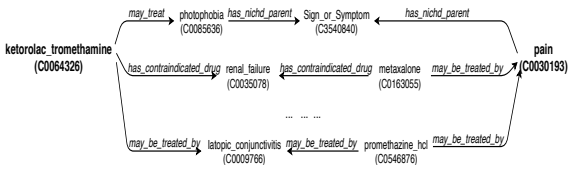


Figure 4: Multiple reasoning paths between *ketorolac_tromethamine* and *pain*.

$$\mathbf{o} = \mathbf{M}[\mathbf{s}_{final}; \mathbf{p}_{final}] + \mathbf{d} \tag{13}$$

Similar to the base model, we define the optimization function as the log-likelihood of the objective function in Equation 14.

$$P(G, D|\theta) = P(G|\theta_{\mathcal{E}}, \theta_{\mathcal{R}}) + P(D|\theta_S, \theta_P) \tag{14}$$

4.2 Reasoning Paths Generation

Let (e_1, e_2) be an entity pair of interest. The set of reasoning paths P_r is obtained by computing all shortest paths in a KG starting from e_1 till e_2. For simulating the situation where the direct relation between a target entity pair is unavailable in a sparse KG, we remove the triplet that directly connect the target entity pair of interest from the KG. Each reasoning path, thus, is at least a two-hop path, namely $p = \{e_1, r_1, e_{r_1}, r_2, e_2\}$. However, if the shortest path is not found due to the sparsity of KG, we will use a padding path to represent the missing path $p = \{r_{padding}\}$. Figure 4 shows the generated paths between *ketorolac_tromethamine* and *pain*.

5 Experiments

Our experiments aim to demonstrate the effectiveness of the proposed model, which is discussed in Section 4, for DS-RE from biomedical datasets.

5.1 Data

The biomedical datasets used for evaluation consist of knowledge graph, textual data and reasoning path, which will be detailed as follows.

Knowledge Graph. We choose the UMLS as the KG. UMLS is a large biomedical knowledge base developed at the U.S. National Library of Medicine. UMLS contains millions of biomedical concepts and relations between them. We follow (Wang et al., 2014), and only collect the fact triplet with RO relation category (RO stands for "has Relationship Other than synonymous, narrower, or broader"), which covers the interesting relations such as *may_treat* and *my_prevent*. From the UMLS 2018 release, we extract about 50 thousand such RO fact triplets (i.e., (e_1, r, e_2)) under the restriction that their entity pairs (i.e., e_1 and e_2) should coexist within a sentence in Medline corpus. They are then randomly divided into training and testing sets for KGC. Following (Weston et al., 2013), we keep high entity overlap between training and testing set, but zero fact triplet overlap. The statistics of the extracted KG is shown in Table 1.

#Entity	#Relation	#Train (triplet)	#Test (triplet)
16,049	295	34,378	12,502

Table 1: Statistics of KG in this work.

Textual Data. Medline corpus is a collection of bimedical abstracts maintained by the National Library of Medicine. From the Medline corpus, by applying the UMLS entity recognizer, QuickUMLS (Soldaini and Goharian, 2016), we extract $682,093$ sentences that contain UMLS entity pairs as our textual data, in which $485,498$ sentences are for training and $196,595$ sentences for testing. For identifying the NA relation, besides the "related" sentences, we also extract the "unrelated" sentences based on a closed world assumption: pairs of entities not listed in the KG are regarded to have NA relation and sentences containing them considered to be the "unrelated" sentences. By this way, we extract $1,394,025$ "unrelated" sentences for the training data, and $598,154$ "unrelated" sentences for the testing data. Table 2 presents some sample sentences in the training data.

Reasoning Path. Following the Section 4.2, we extract $197,396$ paths for not NA triplets ($139,224$ / $58,172$ for training / testing) and $679,408$ for NA triplets ($474,263$ / $205,145$ for training / testing), under the restriction that each entity in a path should be observed in Medline corpus.

5.2 Parameter Settings

We base our work on (Han et al., 2018) and its implementation available at `https://github.com/thunlp/JointNRE`, and thus adopt identical optimization process. We use the default settings

Fact Triplet	Textual Data
(insulin, gene_product_plays_role_in_biological_process, energy_expenditure)	s_1 : *These results indicate that hyperglucagonemia during* $\underline{insulin}_{e_1}$ *deficiency results in an increase in* $\underline{energy\_expenditure}_{e_2}$*, which may contribute to the catabolic_state in many conditions.* s_2 : *It was hypothesized that the waxy maize treatment would result in a blunted and more sustained glucose and* $\underline{insulin}_{e_1}$ *response, as well as* $\underline{energy\_expenditure}_{e_2}$ *and appetitive responses.* s_3 : ...
(IRI, NA, insulin)	s_1 : *Plasma insulin immunoreactivity (*$\underline{IRI}_{e_1}$*) results from high molecular weight substances with insulin immunoreactivity (HWIRI), proinsulin (PI) and* $\underline{insulin}_{e_2}$ *(I).* s_2 : *The beads method demonstrated high* $\underline{IRI}_{e_1}$ *values in both* $\underline{insulin}_{e_2}$ *fractions and the fractions containing serum_proteins bigger than 40,000 molecular weight.* s_3 : ...

Table 2: Examples of textual data extracted from Medline corpus.

of parameters [2] provided by the base model. Since we address the DS-RE in biomedical domain, we use the Medline corpus to train the domain specific word embedding projection matrix $\mathbf{W}^{w}_{emb}$ in Equation 5.

5.3 Result and Discussion

We investigate the effectiveness of our proposed model with respect to enhancing the DS-RE from biomedical datasets. We follow (Mintz et al., 2009; Weston et al., 2013; Lin et al., 2016; Han et al., 2018) and conduct the held-out evaluation, in which the model for DS-RE is evaluated by comparing the fact triplets identified from textual data (i.e., the set of sentences containing the target entity pairs) with those in KG. Following the evaluation of previous works, we draw Precision-Recall curves and report the micro average precision (AP) score, which is a measure of the area under the Precision-Recall curve (higher is better), as well as Precision@N (P@N) metrics, which gives the percentage of correct triplets among top N ranked candidates.

Precision-Recall Curves. The Precision-Recall (PR) curves are shown in Figure 5, where "CNN+AVE" represents that the DS-RE model uses the average vector of sentences as $\mathbf{s}_{final}$ in Equation 7. "JointE+KATT" (or "JointD+KATT") represents that the DS-RE model applies Prob-TransE (or Prob-TransD) as its KG Encoding Part for attention calculation. "(TEXT)" indicates that the model only takes the textual data as input (i.e., the set of sentences containing target entity pairs). "(PATH)" indicates the DS-RE model only takes the reasoning paths between entity pairs as its input. "(TEXT+PATH)" indicates the DS-RE model takes both the textual data and reasoning paths as its input.

[2] As a preliminary study, we only adopt the default hyperparameters, but we will tune them for our task in the furture.

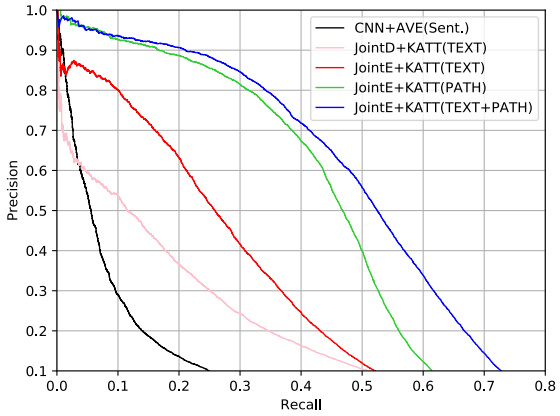


Figure 5: Precision-Recall curves.

The results show that:

(1) The proposed model (i.e., "JointE+KATT(PATH+TEXT)") significantly outperform the base model (i.e., "JointE+KATT(TEXT)"), proving that reasoning paths are useful BK for biomedical DS-RE. This result inspires us to explore other reasoning strategy such as by reasoning across multiple documents. (2) "JointE+KATT(PATH+TEXT)" achieves better overall performance than "JointE+KATT(PATH)", demonstrating the mutual complementary relationship between the sentences containing entity pairs and the reasoning paths between them. Specifically, on the one hand, as discussed in Section 1, reasoning paths could provide BK for interpreting the implicitly expressed relation in sentences. On the other hand, due to the sparsity of KG, it is by no means certain that all entity pairs are fully connected by plausible reasoning paths in the KG. In that case, the sentences could provide the informative evidence to identify the relation between them.

AP and P@N Evaluation. The results in terms of P@1k, P@2k, P@3k, P@4k, P@5k, the mean of them and AP are shown in Table 3. From the table, we have similar observation to the PR curves: (1) The proposed model (i.e., "JointE+KATT(TEXT+PATH)") significantly outperforms the base model for all measures. (2) "JointE+KATT(TEXT+PATH)" outperforms "JointE+KATT(PATH)" in most of the metrics.

Model	P@1k	P@2k	P@3k	P@4k	P@5k	Mean	AP
CNN+AVE	0.852	0.751	0.685	0.640	0.602	0.706	0.098
JointD+KATT(TEXT)	0.628	0.614	0.552	0.495	0.446	0.547	0.186
JointE+KATT(TEXT)	0.835	0.759	0.692	0.629	0.564	0.696	0.272
JointE+KATT(PATH)	**0.945**	0.911	0.881	0.842	0.796	0.875	0.432
JointE+KATT(TEXT+PATH)	0.941	**0.922**	**0.897**	**0.865**	**0.818**	**0.889**	**0.496**

Table 3: P@N and AP for different RE models, where k=1000.

Case Study. Table 4 shows the comparison of the attention distribution between "JointE+KATT(TEXT)" (Base) and "JointE+KATT(TEXT+PATH)" (Proposed). The first and second columns represent the attention distribution (the highest and the lowest) over input sentences. From the Table 4, we can see that the proposed model that incorporates reasoning paths is more capable of selecting informative sentences than the base model, because it "focuses" on the second sentence that explicitly describes the may_treat relation via the word "reduction", in contrast, the base model "ignores" such informative sentence. Table 5 shows the attention allocated by our proposed model for given reasoning paths. The first path generally means if two chemicals should not be used in the case of (or contraindicated with) drug_allergy, they will treat lung_tumor. In contrast, the second path generally means if two chemicals treat Histiocytoses (an excessive number of cells), they will also treat lung_tumor. Apparently the second one that our proposed model focused on is more plausible. This indicates that our proposed model has the capacity of identifying the plausible reasoning path.

6 Conclusion and Future Work

In this work, we tackle the task of DS-RE from biomedical datasets. However, the biomedical DS-RE could be negatively affected by the brevity of text. Specifically, authors always omit the BK that would be important for a machine to identify relationships between entities. To address this issue, in this work, we assume that the reasoning paths over a KG could be utilized as the BK to fill the "gaps" in text and thus facilitate DS-RE. Experimental results prove the effectiveness of the combination, because our proposed model achieves significant and consistent improvements as compared with a state-of-the-art DS-RE model. Although the reasoning paths over KG are useful for DS-RE, the sparsity of KG would hinder their effectiveness. Therefore, in the future, beside the reasoning paths over KG, we will also utilize the reasoning paths across multiple documents for our task. For instance, reasoning across Document1 and Document2, shown below, would facilitate the relation identification between "Aspirin" and "inflammation".

Document1: "***Aspirin*** *and other* ***nonsteroidal anti-inflammatory drugs*** *(NSAID) show ...*"

Document2: "***Nonsteroidal anti-inflammatory drugs*** *reduce* ***inflammation*** *by ...*"

Base	Proposed	Sentences for (**Mitomycin_C (MCC)**, may_treat, **stomach/gastric_tumor**)
High	Low	The additive effect in the combination of TNF and Mitomycin_C was observed against two **Mitomycin_C** resistant **gastric_tumors**.
Low	High	One-quarter or one-half maximum tolerated doses (MTDs) of 5-FU or **MMC** resulted in a significant reduction of **stomach_tumor** growth, ...

Table 4: Comparison of attention between base model and proposed model, where High (or Low) represents the highest (or lowest) attention.

Attention	Paths for (**etoposide**, may_treat, **lung_tumor**)
Low	**etoposide** $\overleftarrow{has\_contraindicated\_drug}$ drug_allergy $\overrightarrow{has\_contraindicated\_drug}$ S-Liposomal Doxorubicin $\overrightarrow{may\_treat}$ **lung_tumor**
High	**etoposide** $\overrightarrow{may\_be\_treated\_by}$ Histiocytoses $\overleftarrow{may\_be\_treated\_by}$ Vinblastine $\overrightarrow{may\_treat}$ **lung_tumor**

Table 5: Some examples of attention distribution over reasoning paths from "JointE+KATT(TEXT+PATH)".

Acknowledgement

This work was supported by JST CREST Grant Number JPMJCR1513, Japan and KAKENHI Grant Number 16H06614.

References

Waleed Ammar, Matthew Peters, Chandra Bhagavatula, and Russell Power. 2017. The ai2 system at semeval-2017 task 10 (scienceie): semi-supervised end-to-end entity and relation extraction. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 592–596.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. AcM.

Antoine Bordes, Nicolas Usunier, Alberto Garcia-Duran, Jason Weston, and Oksana Yakhnenko. 2013. Translating embeddings for modeling multi-relational data. In *Advances in neural information processing systems*, pages 2787–2795.

Jinhua Du, Jingguang Han, Andy Way, and Dadong Wan. 2018. Multi-level structured self-attentions for distantly supervised relation extraction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2216–2225.

Jinghang Gu, Fuqing Sun, Longhua Qian, and Guodong Zhou. 2017. Chemical-induced disease relation extraction via convolutional neural network. *Database*, 2017.

Gus Hahn-Powell, Dane Bell, Marco A Valenzuela-Escárcega, and Mihai Surdeanu. 2016. This before that: Causal precedence in the biomedical domain. *arXiv preprint arXiv:1606.08089*.

Xu Han, Zhiyuan Liu, and Maosong Sun. 2018. Neural knowledge acquisition via mutual attention between knowledge graph and text. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*, pages 541–550. Association for Computational Linguistics.

Guoliang Ji, Shizhu He, Liheng Xu, Kang Liu, and Jun Zhao. 2015. Knowledge graph embedding via dynamic mapping matrix. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, volume 1, pages 687–696.

Guoliang Ji, Kang Liu, Shizhu He, and Jun Zhao. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes, Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. 2015. Dbpedia–a large-scale, multilingual knowledge base extracted from wikipedia. *Semantic Web*, 6(2):167–195.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 2124–2133.

Yankai Lin, Zhiyuan Liu, and Maosong Sun. 2017. Neural relation extraction with multi-lingual attention. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 34–43.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Bonan Min, Ralph Grishman, Li Wan, Chang Wang, and David Gondek. 2013. Distant supervision for relation extraction with an incomplete knowledge base. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 777–782.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. Association for Computational Linguistics.

Makoto Miwa and Mohit Bansal. 2016. End-to-end relation extraction using lstms on sequences and tree structures. *arXiv preprint arXiv:1601.00770*.

Randall Munroe. 2013. The rise of open access. *Science*, 342(6154):58–59.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163. Springer.

RA Roller, E Agirre, A Sorora, and M Stevenson. 2015. Improving distant supervision using inference learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th*

International Joint Conference on Natural Language Processing. Association for Computational Linguistics.

Cicero Nogueira dos Santos, Bing Xiang, and Bowen Zhou. 2015. Classifying relations by ranking with convolutional neural networks. *arXiv preprint arXiv:1504.06580*.

Luca Soldaini and Nazli Goharian. 2016. Quickumls: a fast, unsupervised approach for medical concept extraction. In *MedIR workshop, sigir*.

Shikhar Vashishth, Rishabh Joshi, Sai Suman Prayaga, Chiranjib Bhattacharyya, and Partha Talukdar. 2018. Reside: Improving distantly-supervised neural relation extraction using side information. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1257–1266.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph embedding by translating on hyperplanes. In *AAAI*, volume 14, pages 1112–1119.

Jason Weston, Antoine Bordes, Oksana Yakhnenko, and Nicolas Usunier. 2013. Connecting language and knowledge bases with embedding models for relation extraction. *arXiv preprint arXiv:1307.7973*.

Kun Xu, Yansong Feng, Songfang Huang, and Dongyan Zhao. 2015. Semantic relation classification via convolutional neural networks with simple negative sampling. *arXiv preprint arXiv:1506.07650*.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, Jun Zhao, et al. 2014. Relation classification via convolutional deep neural network. In *COLING*, pages 2335–2344.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762.

Dongxu Zhang and Dong Wang. 2015. Relation classification via recurrent neural network. *arXiv preprint arXiv:1508.01006*.

Peng Zhou, Wei Shi, Jun Tian, Zhenyu Qi, Bingchen Li, Hongwei Hao, and Bo Xu. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 207–212.

Epistemic Marker, Event Type and Factivity in Emotion Expressions

Xuefeng Gao **Chu-Ren Huang** **Sophia Yat-Mei Lee**
Department of Chinese and Bilingual Studies
The Hong Kong Polytechnic University
Hong Kong
xue-feng.gao@connect.polyu.hk, churen.huang@polyu.edu.hk,
ym.lee@polyu.edu.hk

Abstract

Epistemic markers are shown to be an effective linguistic device to introduce cause events of emotions. Linguistically, epistemicity is highly related to factivity. Yet the possible interaction between emotion-causing event types and factivity has not been explored before. This paper reports a corpus-based study on factivity related issues, focusing on the construction "subject + emotion word + epistemic marker + cause event". The interaction between the epistemic marker and event type in sentences with HAPPINESS emotion, and the factivity of negative epistemic marker are analyzed to present a clear picture of the relationship between epistemic marker and emotion. Our study shows that MOVEMENT and LIFE are the two most frequent event types which are introduced by epistemic markers. Moreover, negative epistemic markers do not have any effect of the factivity of the proposition in complement clause and the polarity of emotions. The emotion of the whole sentence depends on the reversal of the event and the negative epistemic marker.

1 Introduction

It is common that the utterance that we make marks our stance. Specifically, people often use stance markers explicitly or implicitly to express their perspectives, evaluations and attitudes. The stance markers include epistemic marker, evidential marker and attitudinal marker. It is found that epistemic markers are often employed to introduce cause event in the construction "subject + emotion word + epistemic marker + cause event" (Lee 2010, 2019; Lee and Huang, 2018). Although some researchers have directed their attention to Chinese epistemic markers, they focus more on the meaning and grammaticalization of these epistemic markers (Yap and Chor 2014, 2019). Other elements in the utterance and the interaction with emotion are often neglected.

The current study aims to examine the construction "subject + emotion word + epistemic marker + cause event" in terms of event type and factivity in emotion expressions. Since it is shown that epistemic markers are most frequently used with HAPPINESS emotion (Lee et al. 2013, Lee 2019) and far exceeding any other emotion, we will also focus on the emotion of happiness in this study. The two research questions are as follows:

1. Which type of cause event is most frequently introduced by epistemic markers in HAPPINESS emotion, especially in terms of factivity?
2. Will the absence of epistemic markers, hence under-specification of factivity, influence the polarity of the emotion?

This paper is organized as follows. In Section 2, we will give an overview of the previous studies in relation to the epistemic marker and their interaction with emotion. Then cause event types introduced by epistemic markers in HAPPINESS emotion will be presented in Section 3. Section 4 will discuss the factivity in setences with negative epistemic markers. Finally, we conclude the paper in Section 5.

2 Related Work

Emotion as an essential facet of cognition has been studied in many different disciplines, such as

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 29-36, Hakodate, Japan, September 13-15, 2019

linguistics, psychology, sociology, neuroscience and computer science. It is found that transitivity and epistemicity are two factors related to emotion expressions linguistically (Lee 2010, 2019; Lee and Huang, 2018). The definition of epistemicity refers to 'pertaining to how a person views the facts of the world, or how they view another person's view of such facts' (Givón, 2009: 315). Although the concept of epistemicity is still debatable, there is a general agreement that epistemicity indicates the certainty of a proposition which shows the attitude of the speaker. There are three types of words that always reveal the epistemicity: (1) epistemic verbs (e.g. think, see); (2) epistemic adverbs (e.g. perhaps, supposedly); (3) modal auxiliaries (e.g. might, may). Epistemic verbs are the verbs which 'perceive or cognize a state or an event, or utters a proposition concerning a state or event' (Givón 1993, I:133), and they are often perception-cognition-utterance (PCU) verbs. The definition of PCU verbs is '(1) the main clause codes mental or verbal activity, with a verb (or adjective) of perception, cognition, mental attitude or verbal utterance; (2) the state or event coded in the complement clause is the object of the mental or verbal activity coded by the main verb; (3) no coreference restrictions hold between arguments of the main clause and complement clause' (Givón 1993, II:4).

There are three types of PCU verbs which are divided by semantic criteria: PCU verbs with preference/aversion, PUC verbs with epistemic attitude, and utterance verbs. For epistemic verbs (PCU verbs with epistemic attitude), they code the relative certainty which shows the reality of the state or event in the complement clause. There are four types of epistemic verbs which code the different degree of the certainty: (1) high epistemic certainty; (2) weak epistemic certainty; (3) low epistemic certainty; and (4) negative epistemic certainty. Givón (1993) argues that epistemic verbs with high epistemic certainty are marked as factive or presuppositional which means that the proposition in the complement clause is believed to be true in spite of the main-clause proposition's truth value.

Factivity is a significant feature for epistemic verbs. Kiparsky and Kiparsky (1971: 348) defined factivity as follows:

> The speaker presupposes that the embedded clause expresses a true proposition, and makes some assertion about that proposition. All predicates which behave syntactically as factives have this semantic property.

Factivity includes four categories: (1) factive; (2) semi-factive; (3) non-factive; (4) negative factive. The different epistemic markers are corresponding with different degree of factivity. Table 1 shows the epistemic certainty continuum (Givón, 1993; Lee, 2019: 73).

Epistemicity	Factivity	Epistemic verbs
Strong epistemic certainty	Factive	*know, remember, forget, see*
Weak epistemic certainty	Semi-factive	*think, assume, guess, suspect*
Epistemic uncertainty	Non-factive	*hope, wonder, doubt*
Negative epistemic certainty	Negative factive	*pretend, lie*

Table1: Epistemic certainty continuum (Givón, 1993; Lee, 2019: 73)

Lee (2010, 2019) and Lee and Huang (2018) found that there was a close relationship between epistemic markers and emotion causes. They argued that epistemic markers were often employed to introduce a cause event of the emotion, and they tended to collocate with change-of-state emotion verbs rather than homogeneous ones. In order to verify the hypothesis, a corpus-based approach was applied in their study. The findings indicated that there were five categories of epistemic verbs: SEEING, HEARING, KNOWING, DISCOVERY and EXISTENCE. SEEING epistemic verbs were most frequently used in the corpus, followed by EXISTENCE epistemic verbs, DISCOVERY epistemic verbs, KNOWING epistemic verbs and HEARING epistemic verbs. Five basic emotions were analyzed and it was found that HAPPINESS most frequently occurred with epistemic markers marking causes, followed by SURPRISE and FEAR. ANGER emotion and SADNESS emotion had limited ability to attach epistemic markers. The epistemic markers have eliminated their original meaning and tend to express the cognitive awareness of the experiencer. The explicit usage of epistemic markers indicates the high motivation for experiencers to claim the certainty of the emotion, e.g., HAPPINESS and SURPRISE, while epistemic markers do not tend to associate with emotions which are triggered by

obscure and unknown events, e.g., SADNESS. The frequency of epistemic markers co-occurred with emotions is HAPPINESS > SURPRISE > FEAR > ANGER >SADNESS.

3 Cause Event Types in HAPPINESS Emotion

Lee (2010, 2019) and Lee and Huang (2018) stated that cause events were always followed by different epistemic markers and epistemic markers were most frequently used explicitly in HAPPINESS emotion to mark causes. The epistemic markers were always followed by the emotion words and finally formed a construction "(subject) + emotion word + epistemic marker + cause event". However, which type(s) of cause event tend to co-occur with epistemic markers and HAPPINESS emotions?

In order to analyze this issue, the data in our study were extracted from BLCU Corpus Center (BCC, Xun, 2016), which contains 15 billion characters. Data from BCC is mainly retrieved from news, literature, *weibo*, etc. The reason why we use this corpus is Chang et al (2000) and Lee (2010, 2019) have been focused on two influential corpora Sinica Corpus and Chinese Gigaword Corpus respectively. Another reason is that constructions can be easily searched in BCC. Then six epistemic markers were chosen in five subcategories as the keywords combined with emotion word 高兴 *gāoxìng* 'glad' to search in the corpus. The six epistemic markers were 看到 *kàndào* 'to see' (SEEING), 听到 *tīngdào* 'to hear' (HEARING), 知道 *zhīdào* 'to know' (KNOWING), 得知 *dézhī* 'to know' (KNOWING), 发现 *fāxiàn* 'to find' (DISCOVERY), 有 *yǒu* 'to have' (EXISTENCE). Two KNOWING epistemic markers are chosen due to the limited data in the corpus. The construction "高兴+看到 *gāoxìng* + *kàndào* 'glad to see'" will be treated as keyword to search in the corpus. In order to balance the genres of data, both *Weibo* data and 多领域 (various fields) data are considered. Table 2 shows the distribution of different constructions.

	看到 *kàndào* 'to see'	听到 *tīngdào* 'to hear'	知道/得知 *zhīdào* /*dézhī* 'to know'	发现 *fāxiàn* 'to find'	有 *yǒu* 'to have'	Total
Weibo	138	12	7	1	281	439
Various fields	526	204	80	16	653	1479
Total	664	216	87	17	934	1918

Table 2: Distribution of "高兴 *gāoxìng* 'glad' + epistemic markers" in the corpus

But some epistemic markers in these entries do not function as epistemic markers. For example, 看到你的博文 *kàndào nǐde bówén* 'to see your posts' is a whole embedded clause, and 看到 *kàndào* 'to see' would be better to be a part of the embedded clause rather than the epistemic marker in (1). While 看到 *kàndào* 'to see' in (2) is regarded as an epistemic marker because 看到 *kàndào* 'to see' is the cognitive perception rather than the real seeing and the embedded clause 一切进行得非常顺利 *yíqiē jìnháng dé fēicháng shùnlì* 'all things go very well' is the cause event of HAPPINESS emotion.

(1) 很高兴看到你的博文，希望保持联系。
hěn gāoxìng kàndào nǐde bówén
very glad see your post
xīwàng bǎochíliánxì
hope keep in touch
'(I am) glad to see your posts and hope we can keep in touch.'

(2) 可娜很高兴**看到**一切进行得非常顺利。
Kěnà hěn gāoxìng kàndào yíqiè
Kena very glad to see all things
jìnxíng de fēicháng shùnlì
go DE very well
'Kena was very glad to see that all things went very well.'

After filtering the data without the function of epistemic marker, data would be annotated based on the annotation scheme. The event classification which is employed in this study is Automatic Content Extraction (ACE). There are eight types of event: LIFE, MOVEMENT, TRANSACTION, BUSINESS, CONFLICT, CONTACT, PERSONELL, and JUSTICE. The annotation scheme is divided into two parts. Firstly, read the whole sentence and identify if the sentence involves negation (as in (3)) or interrogative (as in (4)) because sentences with this two features may result in neutrality and lose the ability to express HAPPINESS emotion. The emotions in both (3) and (4) are neutral, so we do not include these sentences. The next step is to identify the event type if the sentence contains both epistemic marker and cause event. For example, 知道 *zhīdào* 'to know' is the KNOWING epistemic marker in (5) followed by the cause event of the HAPPINESS emotion 她急着想离婚、想嫁给那个宝贝银行家 *tā jízhe xiǎng líhūn xiǎng jiàgěi nàge bǎobèi yínhángjiā* 'she wants to divorce quickly and marries that banker'. The cause event is about divorce and remarriage, so this cause event is classified as LIFE.

(3) 城妈妈好像并不高兴**看到**捣蛋的孩子回来。
Chéng māma hǎoxiàng bìng bù
Cheng mother seem at all NEG
gāoxìng kàndào dǎodàn de háizi
happy to see naughty DE children
huílai
come back
'Mother Cheng seems not happy to see that naughty children came back.'

(4) 你是不是很高兴**知道**别人都怕你?
nǐ shì bù shì hěn gāoxìng zhīdào
you be NEG be very glad to know
biérén dōu pà nǐ
others all be afraid of you
'Are you very glad to know that other people are afraid of you?'

(5) 他很高兴**知道**她急着想离婚、想嫁给那个宝贝银行家。
tā hěn gāoxìng zhīdào tā jízhe
he very glad to know she eager
xiǎng líhūn xiǎng jiàgěi nàge
want divorce want marry to that
bǎobèi yínhángjiā
darling banker
'He was very glad to know that she was eager to divorce quickly and married that banker.'

Table 3 shows the distribution of event types which is introduced by different epistemic markers in our dataset. As shown in Table 3, the most frequent epistemic marker in HAPPINESS emotion expressions is SEEING 看到 *kàndào* 'to see', followed by EXISTENCE 有 *yǒu* 'to have', HEARING 听到 *tīngdào* 'to hear', KNOWING 知道 *zhīdào*/得知 *dézhī* 'to know', DISCOVERY 发现 *fāxiàn* 'to find', which is mostly the same as the work done by Lee (2019). The only difference is the frequency of HEARING and KNOWING epistemic markers which may due to two knowing epistemic markers are included in our study, but it will not affect the result of this study. It is also found that 34% of these six words are used as epistemic markers co-occurring with emotion words in the construction "emotion word + epistemic marker + cause event". As for event types, MOVEMENT is most frequently employed as cause event in HAPPINESS emotion, followed by BUSINESS, LIFE, CONTACT, PERSONELL, CONFLICT, JUSTICE and TRANSACTION.

Epistemic marker	看到 *kàndào* 'to see'		听到 *tīngdào* 'to hear'		知道/得知 *zhīdào* /*dézhī* 'to know'		发现 *fāxiàn* 'to find'		有 *yǒu* 'to have'		Total	
	Token	%	Token	%	Token	%	Token	%	Token	%	Token	%
BUSINESS	95	28	10	14	8	13	0	0	9	5	122	18.8
CONFLICT	17	5	4	6	3	5	2	15	16	10	42	6.5
CONTACT	14	4	10	14	8	13	4	31	80	48	116	17.9

JUSTICE	10	3	2	3	1	2	0	0	0	0	13	2.0
LIFE	61	18	19	28	15	25	3	23	23	14	121	18.7
MOVEMENT	111	33	17	25	21	35	3	23	24	14	176	27.2
PERSONELL	31	9	7	10	3	5	1	8	14	8	56	8.6
TRANSACTION	0	0	0	0	1	2	0	0	1	1	2	0.3
TOTAL	339	100	69	100	60	100	13	100	167	100	648	100

Table 3: Distribution of event types introduced by epistemic markers

Compared with the distribution of event types which is introduced by epistemic markers in *Weibo* posts, the result seems a little bit different. As shown in Table 4, the most frequent cause event type which triggers HAPPINESS emotion and is introduced by epistemic markers is LIFE, followed by CONTACT, MOVEMENT and BUSINESS. It shows that people prefer to show their own feelings through their own life experience or events occurred that have around them in social media. For example, 你们结婚晒幸福 *nǐmen jiéhūn shài xìngfú* 'you marry and show off your happiness' is the embedded clause of the 看到 *kàndào* 'to see', and it is also the cause event of HAPPINESS emotion, which is identified as LIFE because the main idea is about marriage.

(6) 姐妹们，很高兴**看到**你们结婚晒幸福。
jiěmèi men hěn gāoxìng kàndào nǐmen
bestie PL very glad to see your
jiéhūn shài xìngfú
marry show off happiness
'Besties, (I was) so glad to see that you marry and show off your happiness.'

Epistemic marker	看到 *kàndào* 'to see'	听到 *tīngdào* 'to hear'	知道[1] *zhīdào* 'to know'	发现 *fāxiàn* 'to find'	有 *yǒu* 'to have'	Total
BUSINESS	17	0	0	0	3	20
CONFLICT	0	0	0	0	8	8
CONTACT	3	0	0	0	27	30
JUSTICE	3	4	0	0	0	7
LIFE	28	1	3	1	11	44
MOVEMENT	11	1	1	0	15	28
PERSONELL	10	0	1	0	6	17
TRANSACTION	0	0	0	0	0	0
TOTAL	72	6	5	1	70	154

Table 4: Distribution of event types introduced by epistemic markers in *Weibo* posts

HAPPINESS emotion is most frequently used with epistemic markers and SEEING epistemic marker is the most frequent one. It is found that sensory verbs no longer indicate the sensory perception of the cause event, but reflect the cognitive mental process of the cause event. HAPPINESS emotion tends to use epistemic markers because

[1] There is not "高兴得知" *Weibo* data in the corpus.

some motivations need to stimulate experiencers to cause HAPPINESS emotion, compared with other negative emotions (Lee and Huang 2018). As for the cause events, HAPPINESS emotion tends to be stimulated by events which are aspiring (MOVEMENT) and life-oriented (LIFE). The two most frequent event types show that the cause event in HAPPINESS emotion should be factive and then evoke the emotion of happiness, but it is rare in negative emotions. The result is also consistent with Huang and Chang (1996) which indicates that aspectual *–qilai* can only co-occur with a UP metaphor and emotion, and the construction has semantic and syntactic restrictions on the collocation with other elements. These cause events which are introduced by epistemic markers imply that experiencers express HAPPINESS emotion by means of sensory organs and the cause events are factive.

4 Factivity in Sentences with Negative Epistemic Markers

Lee (2010, 2019) and Lee and Huang (2018) found that epistemic markers often introduced the cause event. As in example (7), epistemic marker 看到 *kàndào* 'to see' is between the emotion words 高兴 *gāoxìng* 'glad' and cause event 他能一直坚持这项事业 *tā néng yizhí jiānchí zhè xiàng shìyè* 'he can insist on this business all the time' and it introduces the cause event of emotion HAPPINESS. The presupposition that the embedded clause 他能一直坚持这项事业 *tā néng yizhí jiānchí zhè xiàng shìyè* 'he can insist on this business all the time' is a true proposition regardless of the truth value of the main-clause, so the usage of the strong epistemic marker 看到 *kàndào* 'to see' presuppose the truth of the propositions which means that it is factive. Therefore, the embedded clause in (7) not only codes the high certainty of the event, but also causes the HAPPINESS emotion.

(7) 我们很**高兴看到**他能一直坚持这项事业。
wǒmen hěn gāoxìng kàndào tā néng
we very happy to see he can
yizhí jiānchí zhè xiàng shìyè
all the time insist this CL business
'We were very glad to see that he can insist on this business all the time.'

Two types of presupposition tests are always used to test the presupposition whether remains present when other elements modify in certain aspects. The two classical presupposition tests are constancy under negation and constancy under yes/no question. If the epistemic verbs are factive predicates, they can bear the test. If they are non-factive verbs, the presupposition tests do not work. As in (8), the propositions of the embedded clause in (8a, 8b and 8c) are all true due to the factive epistemic verbs 知道 *zhīdào* 'to know' regardless of the negation or yes/no question test of the sentence.

(8) a. 他**知道**小明已经到了。
>> Xiaoming has arrived
tā zhīdào Xiǎomíng yǐjīng dào le
he know Xiaoming already arrive LE
'He knew that Xiaoming had arrived.'

b.他不知道小明已经到了。
>> Xiaoming has arrived
tā bù zhīdào Xiǎomíng yǐjīng dào le
he NEG know Xiaoming already arrive LE
He didn't know that Xiaoming had arrived.

c. 他知道小明已经到了吗？
>> Xiaoming has arrived
tā zhīdào Xiǎomíng yǐjīng dào le ma
he know Xiaoming already arrive LE SFP
'Did he know that Xiaoming had arrived?'

But when the negation test is applied in (9a), different findings are observed. There are two interpretations when negator 没 *méi* is used in (9b) due to the scope of the negator. The first interpretation indicates that the proposition in the complement clause 他获奖 *tā huò jiǎng* 'he receives an award' is not true which can be regarded as negative factive as defined by Givón (1993) when the scope of the negator only includes the epistemic marker 听到 *tīngdào* 'to hear'. Another interpretation states the proposition in the complement clause can be either true or false because the scope of the negator is the whole parts followed by it. We cannot identify if he receives an award or not from the second interpretation. But in this paper, we will

focus on the first interpretation and discuss the factivity in this construction. This complex construction involves four elements: the adjectival emotion words, negator, epistemic marker, and complement clause. These four elements will contribute to the whole sentence in terms of syntax and semantics.

Apter (2007) pointed out that there were three types of reversal: contingent reversal, frustration reversal and satiation reversal. Contingent reversal is the first level which is represented by lexical opposite pairs triggered by external context. The second level can be linked to frustration reversal which caused by rejection of the original state and the negator signals this rejection. The last level is the satiation reversal which means that the reversal is implicitly marked which will present through the contrary of the meaning. As in (9b), the proposition in complement clause shows the rejection of the original state 他获奖 *tā huò jiǎng* 'he receives an award' but it reverses to failure because of the frustration reversal. The negator 没 *méi* marked the whole sentence and the proposition of the complement clause is still true and it is factive rather than negative factive. Although the complement clause indicates his failure, the emotion of the sentence is still HAPPINESS. Therefore, the emotion of the whole sentence depends on the reversal of the event and the negative epistemic marker. The negative epistemic markers will not influence the polarity of emotion.

(9) a. 我很**高兴**听到他获奖了。
>> he receives an award
wǒ hěn gāoxìng tīngdào tā huò
I very glad to hear he receive
jiǎng le
award LE
'I was very glad to hear that he received an award.'

b. 我很**高兴**没听到他获奖。
>> reversal of 'he receives an award'
wǒ hěn gāoxìng méi tīngdào tā
I very glad NEG to hear he
huò jiǎng
receive award
'I was very glad not to hear that he received an award.'

5 Conclusion

This paper explores how epistemic marker interacts with emotions in terms of event type and factivity. We show that the epistemic marker is frequently used in the construction "subject + emotion word + epistemic marker + cause event" and then analyze the cause event types in HAPPINESS emotion sentences that epistemic markers are most frequently used. MOVEMENT and LIFE are two event types which have been most frequently found to be introduced by epistemic markers. These cause events which are introduced by epistemic markers suggest that experiencers express HAPPINESS emotions with the help of sensory organs. Moreover, we found that negative epistemic marker will not affect the factivity of the proposition in complement clauses and the polarity of the emotion. The emotion of the whole sentence relies on the reversal of the event and the negative epistemic marker.

Acknowledgments

This research work is supported by a General Research Fund project sponsored by the Research Grants Council, Hong Kong (Project No. 15609715) and a Faculty Research Grant by the The Hong Kong Polytechnic University (Project No. 1-ZVMF).

References

Apter, Michael. 2007. Reversal Theory: The Dynamics of Motivation, Emotion, and Personality. Oxford: Oneworld.

Chang, Li-li, Keh-jiann Chen, and Chu-Ren Huang. 2000. Alternation across Semantic Field: A Study of Mandarin Verbs of Emotion. In Yung-O Biq (Ed), Special Issue on Chinese Verbal Semantics. Computational Linguistics and Chinese Language Processing, 5(1): 61–80.

Givón, Talmy. 1993. English Grammar: A Function-Based Introduction, 2 Vols. Amsterdam: John Benjamins.

Givón, Talmy. 2009. The ontogeny of complex verb phrases: How children learn a negotiate fact and desire. In Talmy Givon, and Masayoshi Shibatani (Eds), Syntactic Complexity: Diachrony, Acquisition, Neuro-cognition, Evolution, pp.311-388. Amsterdam: John Benjamins.

Huang, Chu-Ren, and Shen-ming Chang. 1996. Metaphor, metaphorical extension, and grammaticalization: A study of Mandarin Chinese. In Adele E Goldberg (Ed), Conceptual Structure, Discourse and Language, pp. 201-216. Stanford, California: Center for the Study of Language and Information.

Kiparsky, Paul, and Carol Kiparsky. 1971. Fact. In Danny Steinberg, and Leon Jacobovits (Eds.), Semantics, pp.345-369. Cambridge, MA: University Press.

Lee, Sophia Yat-Mei. 2010. A Linguistic Approach to Emotion Detection and Classification. Ph. D. dissertation. Hong Kong: The Hong Kong Polytechnic University.

Lee, Sophia Yat-Mei. 2019. Emotion and Cause: Linguistic Theory and Computational Implementation. Berlin: Springer.

Lee, Sophia Yat-Mei, Ying Chen, Chu-Ren Huang, and Shoushan Li. 2013. Detecting emotion causes with a linguistic rule-based approach. Computational Intelligence, Special Issues on Computational Approaches to Analysis of Emotion in Text, 29 (3): 390-416.

Lee, Sophia Yat-Mei, and Chu-Ren Huang. 2018. A linguistic analysis of emotion and cause. Contemporary Linguistics, 20(3): 357-373.

Xun, Endong, Gaoqi Rao, Xiaoyue Xiao, and Jiaojiao Zang. 2016. The construction of the BCC Corpus in the age of Big Data. Corpus Linguistics, 3(1): 93-109.

Yap, Foong Ha, and Winnie Oi-Wan Chor. 2014. Epistemic, evidential and attitudinal markers in clause-medial position in Cantonese. In Elisabeth Leiss, and Werner Abraham (Eds.), Modes of Modality: Modality, Typology and Universal Grammar, pp.219-260. Amsterdam: John Benjamins.

Yap, Foong Ha, and Winnie Oi-Wan Chor. 2019 The grammaticalization of stance markers in Chinese. In Chris Shei (Ed), The Routledge Handbook of Chinese Discourse Analysis, pp. 230-243. Routledge.

AMR Normalization for Fairer Evaluation

Michael Wayne Goodman
Nanyang Technological University
Singapore
goodmami@uw.edu

Abstract

Abstract Meaning Representation (AMR; Banarescu et al., 2013) encodes the meaning of sentences as a directed graph and Smatch (Cai and Knight, 2013) is the primary metric for evaluating AMR graphs. Smatch, however, is unaware of some meaning-equivalent variations in graph structure allowed by the AMR Specification and gives different scores for AMRs exhibiting these variations. In this paper I propose four normalization methods for helping to ensure that conceptually equivalent AMRs are evaluated as equivalent. Equivalent AMRs with and without normalization can look quite different—comparing a gold corpus to itself with relation reification alone yields a difference of 25 Smatch points, suggesting that the outputs of two systems may not be directly comparable without normalization. The algorithms described in this paper are implemented on top of an existing open-source Python toolkit for AMR and will be released under the same license.

1 Introduction

Abstract Meaning Representation (AMR; Banarescu et al., 2013) encodes the meaning of sentences in a rooted, directed acyclic graph of concepts (labeled nodes) and relations (labeled edges). It was introduced as being to semantics what the Penn Treebank (Marcus et al., 1994) was to syntax—a simple pairing of sentences and hand-authored annotations—and aimed to coalesce multiple aspects of semantic annotation that had previously been done separately, such as named entity recognition, role labeling, and coreference resolution, into one form.

Research efforts targeting AMR often use the Smatch metric (Cai and Knight, 2013) for evaluation. Smatch views AMR graphs as bags of triples and attempts to find a mapping of nodes between two AMRs that results in the highest F-score in terms of matching triples. The result is a single score for a list of AMR pairs. As AMR encodes many aspects of meaning in one graph, some have found it useful to divide up the parts of the graph that Smatch evaluates so as to inspect a parser's aptitude in each task (Damonte et al., 2017). Nevertheless, Smatch remains the primary underlying method for comparing AMRs and thus ensuring that it is a fair metric is important for the task of semantic parsing.

The AMR Specification[1] describes some features of the representation that expand its expressiveness and improve its legibility, such as reifying graph edges to nodes so that the meaning of the edge can be used by other parts of the graph, and rules for inverting edges so the graph can be linearized into the PENMAN format (Matthiessen and Bateman, 1991). The specification says that these alternations express the same meaning, but they result in different triples used by Smatch for comparison.

In this paper, I investigate the effects these differences have on comparison and propose normalization methods to aid in resolving them. Normalization is intended as a preprocessing step to evaluation and is done to both the gold and test corpus. The purpose is not to yield higher Smatch scores or to change system outputs, but to ensure that conceptually equivalent AMRs evaluate as equivalent and that no system is unfairly penalized or rewarded.

[1] https://github.com/amrisi/amr-guidelines

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 37-46, Hakodate, Japan, September 13-15, 2019

2 Background

While AMR and its PENMAN notation are often considered one and the same, I find that distinguishing them aids the discussion of the Smatch metric, so in this section I explain all three in turn.

2.1 PENMAN Graph Notation

PENMAN notation for AMR is a variation of Sentence Plan Language (Kasper and Whitney, 1989) for the PENMAN project (Matthiessen and Bateman, 1991). The notation is applicable to graphs that are: (1) directed and acyclic (DAGs), (2) connected, (3) with a distinguished root called the *top*, and (4) with labeled nodes and edges.[2] The basic syntax for nodes and edges is as follows:

⟨*node*⟩ ::= '(' ⟨*id*⟩ '/' ⟨*node-label*⟩ ⟨*edge*⟩* ')'

⟨*edge*⟩ ::= ':'⟨*edge-label*⟩ (⟨*const*⟩|⟨*id*⟩|⟨*node*⟩)

The recursion of nodes as targets of edges can only capture projective structures such as trees. In order to encode multiple roots (besides the top node), edges are inverted so the source becomes the target by appending `-of` to the edge label. For reentrancies, node identifiers, also and hereafter called *variables*, are reused.[3] Figure 1 shows an example PENMAN serialization, with all the above features, along with the graph it describes.

```
(n1 / A
   :attr "value"
   :edge1 (n2 / B)
   :edge2-of (n3 / C
      :edge3 n2))
```

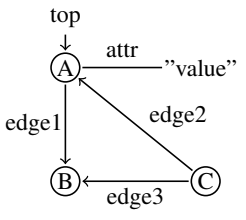


Figure 1: PENMAN notation and the equivalent graph

This paper uses the relative terms *parent* and *child* for the nodes of an edge in the tree structure and *source* and *target* for nodes in the directed graph edges (i.e., such that parent=source in regular edges and parent=target in inverted edges). Edges whose target is a constant are *attributes*. The place where a node specifies its label is the *node definition*.

AMR, described in the next section, uses PENMAN notation to serialize its graph structure. While AMR and PENMAN share a history, the graph notation is not restricted to AMR and could in principle be used for any graphs that meet its criteria. For example it has also been used to encode Dependency Minimal Recursion Semantics (DMRS; Copestake, 2009) for neural text generation (Hajdik et al., 2019) and machine translation (Goodman, 2018).

2.2 Abstract Meaning Representation

Where PENMAN notation is the serialization format, Abstract Meaning Representation (Banarescu et al., 2013) is the semantic framework. As AMR graphs encode semantic information, it refers to node labels as *concepts*, to edges as *relations*, and to edge labels as *roles*. AMR defines in the specification and annotation documentation[4] the inventories of valid concepts and roles and their usage. An AMR graph serialized in PENMAN notation, as in Fig. 2, is simply called an *AMR*, but it can also be represented as a sequence of triples, as in Fig. 3. Node labels are represented by `instance` triples[5] and the top node is indicated with the `:TOP` triple.

```
(d / drive-01
   :ARG0 (h / he)
   :manner (c / care-04
      :polarity -))
```

Figure 2: AMR for *He drives carelessly.*

```
instance(d, drive-01) ^
instance(h, he) ^
instance(c, care-04) ^
TOP(top, d) ^
ARG0(d, h) ^
manner(d, c) ^
polarity(c, -)
```

Figure 3: Triples for *He drives carelessly.*

Several PENMAN graphs may correspond to the same set of triples. A tree-structured graph as in

[2] No technical reason precludes cyclic and unlabeled graphs in PENMAN but I will consider these errors for this paper.

[3] Kasper and Whitney (1989) allowed node attributes and edges to be distributed across multiple references to the node but I will not consider this feature in this paper.

[4] `https://www.isi.edu/~ulf/amr/lib/roles.html`

[5] Some prefer `instance-of` but the choice is arbitrary; I use `instance` to avoid the ramifications of inverted edges.

Fig. 2 has limited options—the branches for `:ARG0` and `:manner` can swap positions, but that's it—but graphs with reentrancies can "rotate" on the reentrant nodes. For example, the graph in Fig. 1 could also be represented as in Fig. 4 or 26 other ways.[6] These alternative serializations do not affect the meaning as determined by the triples (used in evaluation as discussed below), but they can cause issues for systems that learn the serialized character sequences (e.g., Konstas et al., 2017; van Noord and Bos, 2017). Konstas et al. (2017) found that human annotators preferred to insert non-core and inverted relations in the same order as in the original sentence, which leaked ordering information.

```
(n1 / A
   :edge1 (n2 / B
      :edge3-of (n3 / C
         :edge2 n1))
   :attr "value")
```

Figure 4: Alternative serialization of the graph in Fig. 1

While AMR lacks a notion of scope and has no direct model theoretic interpretation,[7] it can encode partial scope information implicitly. For example, the AMRs for *the fast car is red* and *the red car is fast* would differ only by which concept, `fast-02` or `red-02`, is the top of the graph (AMR calls this "focus"). If the examples were, instead, *the fast car that is red* and *the red car that is fast*, then `car` would be the top of both and the triples would be the same, but the PENMAN serializations could differ. Furthermore, reentrancies in AMR present a choice of which occurrence of a variable gets the node definition. It would not be surprising, therefore, for annotators to prefer different PENMAN arrangements for sentences with the same triples, as in Figs. 5 and 6. Put another way, the PENMAN serialization can encode information not present in the triples.

The AMR Specification also describes equivalent[8] variants where the triples do in fact differ. One case is the roles `:domain` and `:mod`, which are considered equivalent in the inverse (i.e., `:domain-of` is equivalent to `:mod`, etc.). The other case is reified relations, where a relation between two nodes becomes a binary node, which is useful when the relation itself interacts with other parts of the graph. These are explained further in Sections 3.1 and 3.2.

```
(b / bite-01
   :ARG0 (d / dog
      :ARG0-of (c / chase-01
         :ARG1 (b / boy)))
   :ARG1 b)
```

Figure 5: AMR for *The dog chasing the boy bit him.*

```
(b / bite-01
   :ARG0 d
   :ARG1 (b / boy
      :ARG1-of (c / chase-01
         :ARG0 (d / dog))))
```

Figure 6: AMR for *The boy chased by the dog was bit by it.*

2.3 Smatch

Smatch (Cai and Knight, 2013) is the primary metric used for AMR evaluation. It estimates the "overlap" between two AMRs by finding a mapping of variables that optimizes the number of matching triples. Precision is defined as $\frac{M}{T}$ and recall as $\frac{M}{G}$ where M is the number of matching triples, T is the number of test triples, and G is the number of gold triples,[9] and the final Smatch score is the F-score of these two. Finding an ideal mapping is an NP-complete task, so Smatch approximates it using greedy search with random restarts to avoid local optima. As regular and inverted relations in AMR are the same when presented as triples, any rearrangement of the PENMAN form for the same triples (as discussed in Section 2.2) will yield the same results as long as the top node does not change, exempting search errors.

Smatch is naïve with respect to AMR-specific interpretations of PENMAN graphs—it only considers the most direct translation of PENMAN graphs to

[6] There are 6 rotations and each rotation has 2 or 6 arrangements by swapping branch positions; more are possible when the top node is not fixed.

[7] Bos (2016) proposed a transformation to first-order logic and also found that a minor change to AMR could allow negation scope to be accurately encoded. Stabler (2017) extended this work and included tense and number features.

[8] Equivalent only by the AMR Specification, not necessarily logical equivalence by a mapping of AMR to logical forms.

[9] The Smatch utility I use (see Section 4) does not specify gold and test, only the first and second arguments. Swapping these arguments swaps precision and recall. I set the gold corpus to the second argument.

triples. It does not consider equivalent alternations where the triples do change (such as `:domain` vs `:mod` alternations and relation reifications) as equivalent, and these alternations will lead to score differences. Smatch is also not robust to subtly invalid graphs, such as inverted edges whose source (i.e., child in the tree structure) is a constant.[10] In this case, the triple will be ignored completely, leading to an inflated score.

Moreover, Smatch gives no credit for a correct role or value unless both are correct. For example, the first line in the Little Prince corpus is *Chapter 7* with the AMR `(c / chapter :mod 7)`, but all three parsers I tested failed to output the correct relation (one gave `:quant 7`, another `:li 7`, and another `:op1 7`). They are therefore all penalized in recall for missing the `:mod 7` relation and again in precision for their incorrect attempt, and none get credit for the correct value of `7`. Omitting the relation entirely (e.g., `(c / chapter)`) yields a higher score, but that's hardly ideal.

The AMR normalizations described in this paper ensure equivalent AMRs have the same triples and thus the same score. In addition, two of the normalizations involve reification which replaces a single triple with several, and this presents a tradeoff: it can allow "partial credit" for getting the role or the value correct, but getting both wrong hurts the score worse than getting a single relation wrong.

3 AMR Normalization

This section describes two meaning-preserving AMR normalizations and two meaning-augmenting normalizations. The first two include canonical role inversions and relation reification, while the latter two include attribute reification and PENMAN structure preservation.

3.1 Canonical Role Inversions

The roles of inverted relations are marked with an `-of` suffix, and generally they are deinverted by removing the suffix. AMR, however, specifies several roles whose canonical form contains the suffix `-of`, namely `:consist-of`, `:prep-on-behalf-of`, and `:prep-out-of`, and the inverse form of these therefore requires an additional suffix (e.g., `:prep-out-of-of`). In addition there is `:mod` which is equivalent to the inverse of `:domain`, and vice-versa.[11] If a gold corpus contained `:mod` while the test corpus used `:domain-of`, Smatch would not see these as equivalent and the score would drop.

By normalizing inverted roles to their canonical forms, such as `:domain-of` → `:mod`, `:consist` → `:consist-of-of`, the Smatch score will not differ for such alternations. Some may argue that normalizing invalid roles such as `:consist` in this way is meaning-altering, but as the naïve inversions of these roles are not separately defined roles in AMR there is no chance of conflation, and in this case I take the position that practicality beats purity.

3.2 Relation Reifications

Some specific relations in AMR can be reified into concepts with separate relations for the original relation's source and target. For example, Fig. 7 is equivalent to Fig. 2 with `:manner` reified to `have-manner-91`. While its possible to reify every eligible relation, in practice all are collapsed unless it is necessary to have the node, so Fig. 2 would generally be preferred over Fig. 7.

```
(d / drive-01
   :ARG0 (h / he)
   :ARG1-of (h2 / have-manner-91
      :ARG2 (c / care-04
         :polarity -))
```

Figure 7: AMR for *He drives carelessly* with `:manner` reified to `have-manner-91`

```
(d / drive-01
   :ARG0 (h / he)
   :ARG1-of (h2 / have-manner-91
      :ARG2 (c / care-04)
      :polarity -))
```

Figure 8: AMR for *He doesn't drive carefully.*

There are three situations where reification is useful: (1) when the meaning of the relation itself is the focus or the argument of another concept instance; (2) when it breaks a cycle in the graph; and (3) when

[10] Only nodes, not constants, may specify relations. These invalid graphs occur occasionally in the output of some parsers.

[11] The specification suggests that `:mod` *is* the inverse of `:domain`, but that could not be true as `:mod` appears in attribute relations and a relation's source cannot be a constant.

Role	Concept	Source	Target
:degree	have-degree-92	:ARG1	:ARG2
:manner	have-manner-91	:ARG1	:ARG2
:purpose	have-purpose-91	:ARG1	:ARG2

Table 1: Sample of reification definitions

an annotator uses a "shortcut" role in a relation. Situation (1) is the only case that is strictly necessary. For example, Fig. 8 is used to express *He doesn't drive carefully*, where the have-manner property is negated rather than the manner itself. The breaking of cycles in situation (2) is possible because reification replaces an edge with a node and two outgoing edges, thus becoming a new root (but not necessarily the graph's top). These kinds of reifications ensure that the graph remains a DAG—a property that may be useful for some applications. The "shortcut" roles of situation (3) are a feature of the AMR Editor (Hermjakob, 2013) provided as a convenience to annotators. They are always reified automatically by the editor and therefore might be considered not part of the official role inventory in the AMR framework. Annotators not using the editor, however, might use them as they are listed in the specification, so it is still useful to reify these in normalization.

In implementation, reification is not complicated. The process uses a defined mapping of roles to AMR fragments containing the reified concept and the roles that capture the original relation's source and target. A sample of these definitions is shown in Table 1; the full list is given in Appendix A. Reification uses this mapping to replace some relation `(a :<role> b)` with `(a :<source>-of (c / <concept> :<target> b))` for regular relations and `(a :<target>-of (c / <concept> :<source> b))` for inverted relations. Reification used in normalization will always have one inverted edge as the original AMR would not have had any way to focus the pre-reified relation.

Collapsing, or dereifying, nodes to edges is slightly more complicated because there are more restrictions on when it can be applied. A node can only be collapsed if it does not participate in relations (including the `:TOP` relation) other than those resulting from reification.[12] For example, `have-manner-91` in Fig. 7 can be collapsed but it cannot be in Fig. 8 because in the latter it is involved in the `:polarity` relation. The change to the graph itself is just the opposite of reification: `(a :<source>-of (c / <concept> :<target> b))` becomes `(a :<role> b)` and `(a :<target>-of (c / <concept> :<source> b)` becomes `(a :<role>-of b)`.

There are additional complexities when the reification mapping is not one-to-one; that is, when it maps multiple relations to the same concept or a single relation to multiple alternative concepts. For the first case, normalization always introduces a new node for each reified relation, even when multiple relations on the same node are mapped to the same concept. This case only occurs with the shortcut roles `:employed-by`/`:role` and `:subset`/`:superset`. For the second case the relations will not be reified because it is undecidable which of the competing concepts should be used, and likewise in dereification information would be lost by collapsing both concepts to the same relation. This case occurs with `:poss` reifying to either `own-01` or `have-03`, and `:beneficiary` reifying to either `benefit-01` and `receive-01`.

The effect of reification on the Smatch score can be large. By reifying one relation to a node with two relations, the net total of triples increases by two. In the gold corpus (see Sections 4 and 5), roughly 15% of triples were reifiable, so a fully-reified corpus would contain roughly 30% more triples. The result is that Smatch will require more time and memory to compute a score, and the search for the variable mapping may become less stable because there are more nodes to search over. This normalization can affect the Smatch score by amplifying certain kinds of errors and giving partial credit for others. Table 2 shows a gold item (the top AMR for *five apples*) and several test AMRs with various differences. The Collapsed column shows the Smatch score between the gold and test AMRs when the relations are left as-is, and the Reified column shows the score when both gold and test are reified. Smatch's preference for missing versus incorrect relations becomes a dispreference unless the test AMR's role differs and is not reifiable (`:unit` in Table 2).

[12] While it is possible to pull out and collapse the information relating to the reified relation and leave in place the node and its additional relations, I do not do so here.

AMR	Collapsed	Reified
`(a / apple :quant 5)`	-	-
`(a / apple)`	0.80	0.57
`(a / apple :quant` **`1`**`)`	0.67	0.80
`(a / apple` **`:mod`** `5)`	0.67	0.80
`(a / apple` **`:mod 1`**`)`	0.67	0.60
`(a / apple` **`:unit`** `5)`	0.67	0.50
`(a / apple` **`:unit 1`**`)`	0.67	0.50

Table 2: Difference in Smatch score with and without reification (top is gold, rest are test, bold are differences)

3.3 Attribute Reification

As mentioned in Section 2.3, Smatch silently drops triples whose source is not a valid variable, leading to inflated scores. While canonical role inversions (such as `:domain-of` to `:mod`) help here, the situation can be completely averted by reifying every constant into a node with a new unique variable and with the constant as the node's concept. For example, `:mod 7` becomes `:mod (_/ 7)`. The result is not meaning-equivalent as the alternation is not provided by the AMR Specification, but it will at least allow each triple to be considered in evaluation. The effect on Smatch is that each attribute triple is replaced with a relation and a concept triple, thus increasing the number of triples by one for each constant. It also allows for partial credit, similar to reification.

3.4 PENMAN Structure Preservation

Section 2.2 described two kinds of variation in PENMAN that correspond to the same triples: the order of serialized relations on a node and which occurrence of a node contains the node definition. As discussed, these differences can be used to encode nuance or hints to the surface form that the AMR annotates. In order to preserve the information encoded by the location of node definitions, additional `:TOP` relations may be used to indicate which node is the top of the node being defined. These parallel the tree structure rather than the DAG, so they do not invert if the child of an inverted relation (i.e., the relation's source) is a node definition.[13] Inserting these relations into an AMR with n nodes results in $n-1$ new triples as one is not inserted for the top node in the graph. The effect on Smatch is a boost in the score of AMRs that define nodes in the same place.

[13]These `:TOP` relations could lead to a cyclic structure so it is not recommended as a general annotation practice.

4 Experiment Setup

For information about roles and their reifications I use the AMR 1.2.6 Specification[14] and the annotator documentation of roles as of May 1, 2019.[15] For reification I use all non-ambiguous mappings, which excludes `:beneficiary` and `:poss`, and for dereification I also exclude mappings of shortcut roles. My experiments use the training portion of the freely-available Little Prince corpus (version 1.6).[16] For reading and writing PENMAN graphs I use the open-source Penman package for Python.[17] I used JAMR (Flanigan et al., 2016),[18] CAMR (Wang et al., 2016),[19] and AMREager (Damonte et al., 2017)[20] for producing system outputs. All systems use their included models trained on the LDC2015E86 (SemEval Task 8) data, which is out-of-domain for the Little Prince corpus but the parsers then all use comparable models. For comparison I use Smatch (Cai and Knight, 2013).[21]

5 Corpus Analysis

I first inspect the corpus to understand the distribution of normalizable AMRs. Table 3 shows the number of nodes and triples in The Little Prince corpus (1,274 AMRs) for both gold annotations and system outputs. These counts are used for calculating the percentages in Tables 4 and 5.

Table 4 shows the percentage of graphs and triples that have the non-canonical `:domain-of` and `:mod-of` relations. They do not appear in the gold annotations, CAMR output, or AMREager output,

[14]https://github.com/amrisi/amr-guidelines

[15]https://www.isi.edu/~ulf/amr/lib/roles.html

[16]https://amr.isi.edu/download.html

[17]https://github.com/goodmami/penman/

[18]Semeval-2016 branch as of March 21, 2019: https://github.com/jflanigan/jamr

[19]Master branch as of February 19, 2018: https://github.com/c-amr/camr

[20]Master branch as of April 11, 2019: https://github.com/mdtux89/amr-eager

[21]https://github.com/snowblink14/smatch/

Corpus	# Nodes	# Triples
Gold	8,189	16,832
JAMR	8,115	15,509
CAMR	7,404	13,922
AMREager	7,461	15,226

Table 3: Corpus sizes

but do in the JAMR output along with a small number of `:consist` relations (not shown), and no corpus used non-canonical inversions of the `:prep-*` relations. This is not unexpected, as the gold corpus does not contain any instances of these roles, so data-driven parsers would have no examples to learn from. A parser that assembles the graph and inverts as necessary to serialize may be susceptible.

	% `:domain-of`		% `:mod-of`	
Corpus	Graphs	Triples	Graphs	Triples
Gold	0	0	0	0
JAMR	5.81	0.52	8.63	0.80
CAMR	0	0	0	0
AMREager	0	0	0	0

Table 4: Non-canonical role inversions

Table 5 shows the percentage of graphs and relations that are reifiable and the percentage of graphs and nodes that are collapsible. All systems have roughly as many reifiable graphs and relations as the gold corpus. CAMR is the only system that outputs reified relations that can be collapsed, although the number is miniscule.

	% Reifiable		% Collapsible	
Corpus	Graphs	Rels	Graphs	Nodes
Gold	78.96	15.23	0	0
JAMR	73.94	13.07	0	0
CAMR	68.68	14.22	0.16	0.03
AMREager	76.92	17.02	0	0

Table 5: Reifiable relations and collapsible nodes

Using Smatch to compare two versions of the gold corpus—one original and one with reified relations—yields an F-Score of 0.75, or a drop of 25 Smatch points. This result is an estimate of the range of score variation when a system perfectly reproduces the gold corpus but makes the opposite decision regarding reification.

6 System Evaluation

Here I test the effect the normalizations have on Smatch when evaluating system outputs to the gold corpus. Table 6 shows the results of the three systems with various normalizations. While JAMR was the only parser that output non-canonical roles, normalizing the roles did not help its score; in fact, the score dropped slightly. Some of JAMR's non-canonical roles were inverted relations to constants, so Smatch was ignoring them. Normalizing them would thus hurt the score unless the normalized relations were correct. Reification (both kinds) generally led to higher scores, meaning that most relations that were reified were fully or partially correct. One result that stands out is structure preservation; for both JAMR and AMREager it led to decreased scores but it helped CAMR, showing that CAMR is more likely to place node definitions where an annotator would. Finally, the normalization helped AMREager close the gap with JAMR, and in some configurations even surpass it.

7 Related Work

Konstas et al. (2017) normalized AMRs is a destructive way in order to reduce data sparsity for their character-based neural parser and generator. My normalization methods can also reduce sparsity but they also generally increase the size and complexity of the graph, so it's not clear if it would aid character-based models. Damonte et al. (2017) found that parsers do well on different sub-tasks, such role labeling and word-sense disambiguation, and ran Smatch on different subsets of the triples in order to highlight a parser's performance in each task. In addition, Damonte et al. also found that Smatch weighted certain error types more than others, although they looked at more application-specific error types, like the representation of proper names. In contrast, I compare using the full graphs as the goal is normalization, not specialization. My normalization methods are mostly compatible with the subtask evaluation of Damonte et al. 2017 but some the evaluation tasks look for certain roles which disappear on reification. Anchiêta et al. (2019) also noticed that Smatch gives more weight to the top node of the graph, but they reached different conclusions. Where I proposed adding `:TOP` re-

System	Normalization I	A	R	S	Score P	R	F
					0.60	0.56	0.58
	✓				0.60	0.55	0.57
		✓			0.61	0.56	0.58
			✓		0.63	0.57	0.60
JAMR				✓	0.59	0.55	0.57
	✓		✓		0.63	0.57	0.60
		✓	✓		0.64	0.57	0.60
	✓	✓	✓		0.64	0.57	0.60
	✓	✓	✓	✓	0.61	0.56	0.59
					0.67	0.56	0.61
	✓				0.67	0.56	0.61
		✓			0.67	0.55	0.60
			✓		**0.70**	0.57	**0.63**
CAMR				✓	0.68	**0.58**	**0.63**
	✓		✓		0.69	0.57	**0.63**
		✓	✓		**0.70**	0.56	0.62
	✓	✓	✓		**0.70**	0.56	0.62
	✓	✓	✓	✓	**0.70**	**0.58**	**0.63**
					0.57	0.52	0.55
	✓				0.57	0.52	0.55
		✓			0.57	0.53	0.55
			✓		0.61	0.57	0.59
AMREager				✓	0.59	0.54	0.56
	✓		✓		0.61	0.57	0.59
		✓	✓		0.60	**0.58**	0.59
	✓	✓	✓		0.60	**0.58**	0.59
	✓	✓	✓	✓	0.61	0.57	0.59

Table 6: Smatch results comparing gold to system outputs with the original graphs, canonical role inversions (I), attribute reification (A), relation reification (R), and structure preservation (S)

lations to all nodes to preserve the PENMAN structure, they discard the `:TOP` node, meaning that the AMRs for *the fast car is red* and *the red car is fast* are evaluated as equivalent. Barzdins and Gosko (2016) presented extensions to Smatch including a visualization of per-sentence error patterns and an ensemble selection from multiple test AMRs per gold AMR. The latter extension could in principle be combined with the normalization procedures I have described, however it would need to be augmented to allow for the normalizations of the gold corpus as well as the test corpus.

8 Conclusion and Future Work

AMR provides flexibility with the way that equivalent graphs are encoded. This flexibility can make life easier for annotators and parsers alike, but it also means that evaluation tools not aware of these allowed alternations can give unfair results. I introduced four normalization methods in this paper. Of these, canonical role inversion, relation reification, and attribute reification are intended to tame the variation that can reasonably appear in parser outputs. The fourth, PENMAN structure preservation, makes evaluation more strictly account for annotation choices which may implicitly encode subtle distinctions in meaning, like scope or nuance.

The evaluation results when comparing a normalized test corpus to the similarly normalized gold corpus are not drastically different. I think this result is a good thing, particularly because comparing a corpus to itself with and without normalization has a very large difference in scores. It suggests that normalization, done to both sides, resolves small differences. While one parser I tested, CAMR, maintained its lead with normalized outputs, the third-place parser AMREager nearly caught up to the second-place JAMR. The relative changes in evaluation scores may important for determining state-of-the-art parsers or for shared task competitions.

The normalizations may be useful not only for evaluation but for preprocessing for data-driven workflows. By removing sources of variation, data sparsity can be reduced which could benefit parser training. The increase in graph size due to the normalization, however, may counteract the benefits. I leave this question open to future research.

The code for this paper is available online at `https://github.com/goodmami/norman`.

References

Rafael Torres Anchiêta, Marco Antonio Sobrevilla Cabezudo, and Thiago Alexandre Salgueiro Pardo. 2019. Sema: an extended semantic evaluation for AMR. In *Proceedings of the 20th Computational Linguistics and Intelligent Text Processing*. Springer International Publishg.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2013. Abstract meaning representation for sembanking. In *Proceedings of the 7th Linguistic Annotation Workshop and Interop-*

erability with Discourse, pages 178–186, Sofia, Bulgaria. Association for Computational Linguistics.

Guntis Barzdins and Didzis Gosko. 2016. RIGA at SemEval-2016 task 8: Impact of Smatch extensions and character-level neural translation on AMR parsing accuracy. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1143–1147, San Diego, California. Association for Computational Linguistics.

Johan Bos. 2016. Expressive power of abstract meaning representations. *Computational Linguistics*, 42(3):527–535.

Shu Cai and Kevin Knight. 2013. Smatch: an evaluation metric for semantic feature structures. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, volume 2, pages 748–752.

Ann Copestake. 2009. **Invited Talk:** slacker semantics: Why superficiality, dependency and avoidance of commitment can be the right way to go. In *Proceedings of the 12th Conference of the European Chapter of the ACL (EACL 2009)*, pages 1–9, Athens, Greece. Association for Computational Linguistics.

Marco Damonte, Shay B. Cohen, and Giorgio Satta. 2017. An incremental parser for abstract meaning representation. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 536–546, Valencia, Spain. Association for Computational Linguistics.

Jeffrey Flanigan, Chris Dyer, Noah A Smith, and Jaime Carbonell. 2016. CMU at SemEval-2016 task 8: Graph-based AMR parsing with infinite ramp loss. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1202–1206.

Michael Wayne Goodman. 2018. *Semantic Operations for Transfer-based Machine Translation*. Ph.D. thesis, University of Washington, Seattle.

Valerie Hajdik, Jan Buys, Michael Wayne Goodman, and Emily M. Bender. 2019. Neural text generation from rich semantic representations. In *Proceedings of the 2019 Conference on the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT))*, Minneapolis, Minnesota.

Ulf Hermjakob. 2013. AMR editor: A tool to build abstract meaning representations. Technical report, ISI.

Robert Kasper and Richard Whitney. 1989. SPL: A sentence plan language for text generation. *University of Southern California*.

Ioannis Konstas, Srinivasan Iyer, Mark Yatskar, Yejin Choi, and Luke Zettlemoyer. 2017. Neural AMR: Sequence-to-sequence models for parsing and generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 146–157, Vancouver, Canada. Association for Computational Linguistics.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The penn treebank: Annotating predicate argument structure. In *Proceedings of the Workshop on Human Language Technology*, HLT '94, pages 114–119, Stroudsburg, PA, USA. Association for Computational Linguistics.

Christian Matthiessen and John A Bateman. 1991. *Text generation and systemic-functional linguistics: experiences from English and Japanese*. Pinter Publishers.

Rik van Noord and Johan Bos. 2017. Neural semantic parsing by character-based translation: Experiments with abstract meaning representations. *Computational Linguistics in the Netherlands Journal*, 7:93–108.

Edward Stabler. 2017. Reforming AMR. In *International Conference on Formal Grammar*, pages 72–87. Springer.

Chuan Wang, Sameer Pradhan, Xiaoman Pan, Heng Ji, and Nianwen Xue. 2016. CAMR at semeval-2016 task 8: An extended transition-based AMR parser. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pages 1173–1178, San Diego, California. Association for Computational Linguistics.

A Relation Reifications

Role	Concept	Source	Target	Reifies	Dereifies	Shortcut
:accompanier	accompany-01	:ARG0	:ARG1	✓	✓	
:age	age-01	:ARG1	:ARG2	✓	✓	
:beneficiary	benefit-01	:ARG0	:ARG1			
:beneficiary	receive-01	:ARG2	:ARG0			
:cause	cause-01	:ARG1	:ARG0	✓		✓
:concession	have-concession-91	:ARG1	:ARG2	✓	✓	
:condition	have-condition-91	:ARG1	:ARG2	✓	✓	
:cost	cost-01	:ARG1	:ARG2	✓		✓
:degree	have-degree-92	:ARG1	:ARG2	✓	✓	
:destination	be-destined-for-91	:ARG1	:ARG2	✓	✓	
:domain	have-mod-91	:ARG2	:ARG1	✓	✓	
:duration	last-01	:ARG1	:ARG2	✓	✓	
:employed-by	have-org-role-91	:ARG0	:ARG1	✓		✓
:example	exemplify-01	:ARG0	:ARG1	✓	✓	
:extent	have-extent-91	:ARG1	:ARG2	✓	✓	
:frequency	have-frequency-91	:ARG1	:ARG2	✓	✓	
:instrument	have-instrument-91	:ARG1	:ARG2	✓	✓	
:li	have-li-91	:ARG1	:ARG2	✓	✓	
:location	be-located-at-91	:ARG1	:ARG2	✓	✓	
:manner	have-manner-91	:ARG1	:ARG2	✓	✓	
:meaning	mean-01	:ARG1	:ARG2	✓		✓
:mod	have-mod-91	:ARG1	:ARG2	✓	✓	
:name	have-name-91	:ARG1	:ARG2	✓	✓	
:ord	have-ord-91	:ARG1	:ARG2	✓	✓	
:part	have-part-91	:ARG1	:ARG2	✓	✓	
:polarity	have-polarity-91	:ARG1	:ARG2	✓	✓	
:poss	own-01	:ARG0	:ARG1			
:poss	have-03	:ARG0	:ARG1			
:purpose	have-purpose-91	:ARG1	:ARG2	✓	✓	
:quant	have-quant-91	:ARG1	:ARG2	✓	✓	
:role	have-org-role-91	:ARG0	:ARG2	✓		✓
:source	be-from-91	:ARG1	:ARG2	✓	✓	
:subevent	have-subevent-91	:ARG1	:ARG2	✓	✓	
:subset	include-91	:ARG2	:ARG1	✓		✓
:superset	include-91	:ARG1	:ARG2	✓		✓
:time	be-temporally-at-91	:ARG1	:ARG2	✓	✓	
:topic	concern-02	:ARG0	:ARG1	✓	✓	
:value	have-value-91	:ARG1	:ARG2	✓	✓	

Table 7: Full mapping of roles and concepts used for reification, dereification, and editor shortcuts

A CCG-based Compositional Semantics and Inference System for Comparatives

Izumi Haruta
Ochanomizu University
haruta.izumi@is.ocha.ac.jp

Koji Mineshima
Ochanomizu University
mineshima.koji@ocha.ac.jp

Daisuke Bekki
Ochanomizu University
bekki@is.ocha.ac.jp

Abstract

Comparative constructions play an important role in natural language inference. However, attempts to study semantic representations and logical inferences for comparatives from the computational perspective are not well developed, due to the complexity of their syntactic structures and inference patterns. In this study, using a framework based on Combinatory Categorial Grammar (CCG), we present a compositional semantics that maps various comparative constructions in English to semantic representations, and introduce an inference system that effectively handles logical inference with comparatives, including those involving numeral adjectives, antonyms, and quantification. We evaluate the performance of our system on the FraCaS test suite and show that the system can handle a variety of complex logical inferences with comparatives.

1 Introduction

Gradability is a pervasive phenomenon in natural language and plays an important role in natural language understanding. Gradable expressions can be characterized in terms of the notion of *degree*. Consider the following examples:

(1) a. My car is more *expensive* than yours.
b. My car is *expensive*.

The sentence (1a), in which the comparative form of the gradable adjective *expensive* is used, compares the price of two cars, making it a comparison between degrees. The sentence (1b), which contains the positive form of the adjective, can be regarded as a construction that compares the price of the car to some implicitly given degree (i.e., price).

In formal semantics, many in-depth analyses use a semantics of gradable expressions that relies on the notion of degree (Cresswell, 1976; Kennedy, 1997; Heim, 2000; Lassiter, 2017, among others). Despite this, meaning representations and inferences for gradable expressions have not been well developed from the perspective of computational semantics in previous research (Pulman, 2007). Indeed, a number of logic-based inference systems have been proposed for the task of Recognizing Textual Entailment (RTE), a task to determine whether a set of premises entails a given hypothesis (Bos, 2008; MacCartney and Manning, 2008; Mineshima et al., 2015; Abzianidze, 2016; Bernardy and Chatzikyriakidis, 2017). However, these logic-based systems have performed relatively poorly on inferences with gradable constructions, such as those collected in the FraCaS test suite (Cooper et al., 1994), a standard benchmark dataset for evaluating logic-based RTE systems (see §5 for details).

There are at least two obstacles to developing a comprehensive computational analysis of gradable constructions. First, the syntax of gradable constructions is diverse, as shown in (2):

(2) a. Ann is tall. (Positive)
b. Ann is taller than Bob. (Phrasal)
c. Ann is taller than Bob is. (Clausal)
d. Ann is as tall as Bob. (Equative)
e. Ann is $2''$ taller than Bob. (Differential)

In the examples above, (2c) is a clausal comparative

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 47-56, Hakodate, Japan, September 13-15, 2019

in which *tall* is missing from the subordinate *than*-clause. (2e) is an example of a differential comparative in which a measure phrase, $2''$ (*2 inches*), appears. The diversity of syntactic structures makes it difficult to provide a compositional semantics for comparatives in a computational setting.

Second, gradable constructions give rise to various inference patterns that require logically complicated steps. For instance, consider (3):

(3)
P_1:	Mary is taller than 4 feet.
P_2:	Harry is shorter than 4 feet.
H:	Mary is taller than Harry.

To logically derive H from P_1 and P_2, one has to assign the proper meaning representations to each sentence, and those representations include numeral expressions (*4 feet*), antonyms (*short/tall*), and their interaction with comparative constructions.

For these reasons, gradable constructions pose an important challenge to logic-based approaches to RTE, serving as a testbed to act as a bridge between formal semantics and computational semantics.

In this paper, we provide (i) a compositional semantics to map various gradable constructions in English to semantic representations (SRs) and (ii) an inference system that derives logical inference with gradable constructions in an effective way. We will mainly focus on gradable adjectives and their comparative forms as representatives of gradable expressions, leaving the treatment of other gradable constructions such as verbs and adverbs to future work.

We use Combinatory Categorial Grammar (CCG) (Steedman, 2000) as a syntactic component of our system and the so-called *A-not-A analysis* (Seuren, 1973; Klein, 1980, 1982; Schwarzschild, 2008) to provide semantic representations for comparatives (§2, §3). We use ccg2lambda (Martínez-Gómez et al., 2016) to implement compositional semantics to map CCG derivation trees to SRs. We introduce an axiomatic system COMP for inferences with comparatives in typed logic with equality and arithmetic operations (§4). We use a state-of-the-art prover to implement the COMP system. We evaluate our system[1] on the two sections of the FraCaS test suite (ADJECTIVE and COMPARATIVE) and show that it can handle various complex inferences with gradable adjectives and comparatives.

[1]All code is available at:
https://github.com/izumi-h/fracas-comparatives_adjectives

2 Background

2.1 Comparatives in degree-based semantics

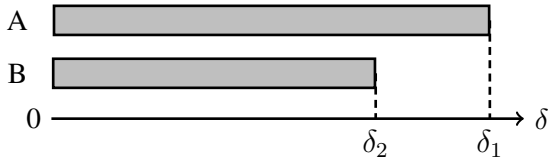
To analyze gradable adjectives, we use the two-place predicate of entities and degrees as developed in degree-based semantics (Klein, 1982; Kennedy, 1997; Heim, 2000; Schwarzschild, 2008). For instance, the sentence *Ann is 6 feet tall* is analyzed as $\mathbf{tall}(\mathbf{Ann}, 6\text{ feet})$, where $\mathbf{tall}(x, \delta)$ is read as "x is (at least) as tall as degree δ".[2]

In degree-based semantics, there are at least two types of analyses for comparatives. Consider (4), a schematic example for a comparative construction.

(4) A is taller than B is.

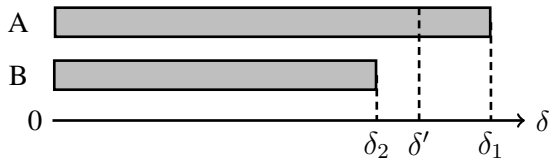
The first approach is based on the maximality operator (Stechow, 1984; Heim, 2000). Using the maximality operator ($\max$) as illustrated in (5), the sentence (4) is analyzed as a statement asserting that the maximum degree δ_1 of A's tallness is greater than the maximum degree δ_2 of B's tallness.

(5) $\max(\lambda\delta.\mathbf{tall}(A, \delta)) > \max(\lambda\delta.\mathbf{tall}(B, \delta))$



The other approach is the A-not-A analysis (Seuren, 1973; Klein, 1980, 1982; Schwarzschild, 2008). In this type of analysis, (4) is treated as stating that there exists a degree δ' of tallness that A satisfies but B does not, as shown in (6).

(6) $\exists\delta\,(\mathbf{tall}(A, \delta) \wedge \neg\,\mathbf{tall}(B, \delta))$



[2]For simplicity, we do not consider the internal structure of a measure phrase like *6 feet*. For an explanation of why $\mathbf{tall}(x, \delta)$ is not treated as "x is *exactly* as tall as δ", see, e.g., Klein (1982).

Table 1: Semantic representations of basic comparative constructions

Type	Example	SR
Increasing Comparatives	Mary is taller than Harry.	$\exists\delta\,(\mathbf{tall}(\mathbf{m},\delta) \wedge \neg\,\mathbf{tall}(\mathbf{h},\delta))$
Decreasing Comparatives	Mary is less tall than Harry.	$\exists\delta(\neg\mathbf{tall}(\mathbf{m},\delta) \wedge \mathbf{tall}(\mathbf{h},\delta))$
Equatives	Mary is as tall as Harry.	$\forall\delta(\mathbf{tall}(\mathbf{h},\delta) \rightarrow \mathbf{tall}(\mathbf{m},\delta))$

Table 2: Semantic representations of complex comparative constructions

Type	Example	SR
Subdeletion Comparatives	Mary is taller than the bed is long.	$\exists\delta(\mathbf{tall}(\mathbf{m},\delta) \wedge \neg\,\mathbf{long}(\mathbf{the}(\mathbf{bed}),\delta))$
Measure phrase comparatives	Mary is taller than 4 feet.	$\exists\delta(\mathbf{tall}(\mathbf{m},\delta) \wedge (\delta > 4'))$
Differential Comparatives	Mary is 2 inches taller than Harry.	$\forall\delta(\mathbf{tall}(\mathbf{h},\delta) \rightarrow \mathbf{tall}(\mathbf{m},\delta + 2''))$
Negative Adjectives	Mary is shorter than Harry.	$\exists\delta\,(\mathbf{short}(\mathbf{m},\delta) \wedge \neg\,\mathbf{short}(\mathbf{h},\delta))$

Although the two analyses are related as illustrated in the figures (5) and (6), we can say that the A-not-A analysis is less complicated and easier to handle than the maximality-based analysis from a computational perspective, mainly because it only involves constructions in first-order logic (FOL).[3] We thus adopt the A-not-A analysis and extend it to various types of comparative constructions for which inference is efficient in our system.

2.2 Basic syntactic assumptions

There are two approaches to the syntactic analysis of comparative constructions. The first is the *ellipsis* approach (e.g. Kennedy, 1997), in which phrasal comparatives such as (2b), are derived from the corresponding clausal comparatives, such as (2c). The other is the *direct* approach (e.g. Hendriks, 1995), which treats phrasal and clausal comparatives independently and does not derive one from the other. An argument against the ellipsis approach is that it has difficulties in accounting for coordination such as that in (7) (Hendriks, 1995).

(7) a. Someone at the party drank more vodka than wine.
b. Someone at the party drank more vodka than someone at the party drank wine.

Here, (7a), a phrasal comparative with an existential NP *someone*, does not have the same meaning as the corresponding clausal comparative (7b); the person who drank vodka and the one who drank wine do not have to be the same person in (7b), whereas they must be the same person in (7a).[4] In this study, we adopt the direct approach and use CCG to formalize the syntactic component of our system.

3 Framework

3.1 Semantic representations

Table 1 shows the SRs for basic constructions of comparatives under the A-not-A analysis we adopt. Using this standard analysis, we also provide SRs for more complex constructions, including subdeletion, measure phrases, and negative adjectives. Table 2 summarizes the SRs for these constructions.

Some remarks are in order about how our system handles various linguistic phenomena related to gradable adjectives and comparatives.

Antonym and negative adjectives *Short* is the antonym of *tall*, which is represented as $\mathbf{short}(x,\delta)$ and has the meaning "the height of x is less than or equal to δ". Thus, we distinguish between the monotonicity property of positive adjectives such as *tall* and *fast* and that of negative adjectives such as *short* and *slow*. For positive adjectives, if $\mathbf{tall}(x,\delta)$ is true, then x satisfies all heights below δ; by contrast, for negative adjectives, if $\mathbf{short}(x,\delta)$ is true, then x satisfies all the heights above δ.

In general, for a positive adjective F^+ and a negative adjective F^-, (8a) and (8b) hold, respectively.

(8) $\forall\delta_1\forall\delta_2 : \delta_1 > \delta_2 \rightarrow$
a. $\forall x(F^+(x,\delta_1) \rightarrow F^+(x,\delta_2))$
b. $\forall x(F^-(x,\delta_2) \rightarrow F^-(x,\delta_1))$

[3] See van Rooij (2008) for a more detailed comparison of the two approaches.

[4] See Hendriks (1995) and Kubota and Levine (2015) for other arguments against the ellipsis approach.

Positive form and comparison class As mentioned in §1, the positive form of an adjective is regarded as involving comparison to some threshold that can be inferred from the context of the utterance. We write $\theta_F(A)$ to denote the contextually specified threshold for a predicate F given a set A, which is called COMPARISON CLASS (Klein, 1982). When a comparison class is implicit, as in (9a) and (10a), we use the universal set U as a default comparison class[5]; we typically abbreviate $\theta_F(U)$ as θ_F. Thus, (9a) is represented as (9b), which means that the height of Mary is more than or equal to the threshold θ_{tall}. Similarly, the SR of (10a) is (10b), which means that the height of Mary is less than or equal to the threshold θ_{short}.

(9) a. Mary is tall.
b. $\mathbf{tall}(\mathbf{m}, \theta_{\text{tall}})$

(10) a. Mary is short.
b. $\mathbf{short}(\mathbf{m}, \theta_{\text{short}})$

A threshold can be explicitly constrained by an NP modified by a gradable adjective. Thus, (11a) can be interpreted as (11b), relative to an explicit comparison class, namely, the sets of animals.[6]

(11) a. Mickey is a small animal. (FraCaS-204)
b. $\mathbf{small}(\mathbf{m}, \theta_{\mathbf{small}}(\mathbf{animal})) \wedge \mathbf{animal}(\mathbf{m})$

Numerical adjectives We represent a numerical adjective such as *ten* in *ten orders* by the predicate $\mathbf{many}(x, n)$, with the meaning that the cardinality of x is at least n, where n is a positive integer (Hackl, 2000). For example, *ten orders* is analyzed as $\lambda x.(\mathbf{order}(x) \wedge \mathbf{many}(x, 10))$. The following shows the SRs of some typical sentences involving numerical adjectives.

(12) a. Mary won ten orders.
b. $\exists x(\mathbf{order}(x) \wedge \mathbf{won}(\mathbf{m}, x) \wedge \mathbf{many}(x, 10))$

(13) a. Mary won many orders.
b. $\exists \delta \exists x(\mathbf{order}(x) \wedge \mathbf{won}(\mathbf{m}, x) \wedge \mathbf{many}(x, \delta) \wedge (\theta_{\text{many}} < \delta))$

(14) a. Mary won more orders than Harry.
b. $\exists \delta(\exists x(\mathbf{order}(x) \wedge \mathbf{won}(\mathbf{m}, x) \wedge \mathbf{many}(x, \delta)) \wedge \neg \exists y(\mathbf{order}(y) \wedge \mathbf{won}(\mathbf{h}, y) \wedge \mathbf{many}(y, \delta)))$

3.2 Compositional semantics in CCG

Here we give an overview of how to compositionally derive the SRs for comparative constructions in the framework of CCG (Steedman, 2000). In the CCG-style compositional semantics, each lexical item is assigned both a syntactic category and an SR (represented as a λ-term). In this study, we newly introduce the syntactic category D for degree and assign $S\backslash NP\backslash D$ to gradable adjectives. For instance, the adjective *tall* has the category $S\backslash NP\backslash D$ and the corresponding SR is $\lambda\delta.\lambda x.\mathbf{tall}(x, \delta)$.

Table 3 lists the lexical entries for representative lexical items used in the proposed system. We abbreviate the CCG category $S\backslash NP\backslash D$ for adjectives as AP and $S/(S\backslash NP)$ (a type-raised NP) as $NP^{\uparrow}$.[7]

The suffix *-er* for comparatives such as *taller* is categorized into four types: clausal and phrasal comparatives (-er$_{\text{simp}}$), subdeletion comparatives (-er$_{\text{sub}}$), measure phrase comparatives (-er$_{\text{mea}}$), and differential comparatives (-er$_{\text{diff}}$). We assume that equatives are constructed from as$_{\text{simp}}$ and as$_{\text{cl}}$; for instance, the equative sentence in Table 1 corresponds to *Mary is as$_{simp}$ tall as$_{cl}$ Harry*. For measure phrase comparatives, such as *Mary is taller than 4 feet*, we use than$_{\text{deg}}$; and for comparatives with numerals, such as (14a), we use more$_{\text{simp}}$.

On the basis of these lexical entries, we can compositionally map various comparative constructions to suitable SRs. Some example derivation trees for comparative constructions are shown in Figure 1 and 2. An advantage of using CCG as a syntactic theory is that the *function composition* rule ($>\mathbf{B}$) can be used for phrasal comparatives such as that in Figure 1, where the VP *is tall* is missing from the subordinate *than*-clause. For positive forms, we use the empty element *pos* of category $S\backslash NP/(S\backslash NP\backslash D)$, as shown in Figure 2.[8]

[5] In this case, we do not consider the context-sensitivity of the implicit comparison class. See Narisawa et al. (2013) for work on this topic in computational linguistics.

[6] Here and henceforth, when an example appears in the FraCaS dataset, we refer to the ID of the sentence in the dataset.

[7] We also abbreviate $\lambda X_1 \ldots \lambda X_n.M$ as $\lambda X_1 \ldots X_n.M$.

[8] Note that the role played by the empty element *pos* here can be replaced by imposing a unary type-shift rule from $S\backslash NP\backslash D$ to $S\backslash NP$.

Table 3: Lexical entries in CCG-style compositional semantics

PF	CCG categories	SR
tall	AP	$\lambda\delta x.\mathbf{tall}(x,\delta)$
Mary	NP	$\mathbf{mary}$
is	$S\backslash NP/(S\backslash NP)$	id
$4'$	D	$4'$
$\text{than}_{\text{simp}}$	S/S	id
than_{deg}	D/D	id
than_{gq}	$S\backslash NP\backslash(S\backslash NP/NP^{\uparrow})/NP^{\uparrow}$	$\lambda QWx.Q(\lambda y.W(\lambda P.P(y))(x))$
pos	$S\backslash NP/AP$	$\lambda A.A(\theta_A)$
-er_{simp}	$S\backslash NP/NP^{\uparrow}\backslash AP$	$\lambda AQx.\exists\delta(A(\delta)(x)\wedge\neg Q(A(\delta)))$
-er_{sub}	$S\backslash NP/(S\backslash D)\backslash AP$	$\lambda AKx.\exists\delta(A(\delta)(x)\wedge\neg K(\delta))$
-er_{mea}	$S\backslash NP/D\backslash AP$	$\lambda A\delta' x.\exists\delta(A(\delta)(x)\wedge(\delta>\delta'))$
-er_{diff}	$S\backslash NP/NP^{\uparrow}\backslash D\backslash AP$	$\lambda A\delta' Qx.\forall\delta(Q(A(\delta))\rightarrow A(\delta+\delta')(x))$
as_{simp}	$S\backslash NP/NP^{\uparrow}/AP$	$\lambda AQx.\forall\delta(Q(A(\delta))\rightarrow A(\delta)(x))$
as_{cl}	S/S	id
more_{num}	$S\backslash NP/NP^{\uparrow}\backslash(S\backslash NP/NP)/N$	$\lambda NGQz.\exists\delta(\exists x(N(x)\wedge G(\lambda P.P(x))(z)\wedge\mathbf{many}(x,\delta))$ $\wedge\neg\exists y(N(y)\wedge Q(G(\lambda P.P(y)))\wedge\mathbf{many}(y,\delta)))$
more_{is}	$S\backslash NP/NP^{\uparrow}\backslash(S\backslash NP/NP)/N/AP$	$\lambda ANGQz.\exists\delta(\exists x(N(x)\wedge G(\lambda P.P(x))(z)\wedge A(\delta)(x)))$ $\wedge\neg Q(\lambda y.(N(y)\wedge A(\delta)(x)))$
more_{has}	$S\backslash NP/NP^{\uparrow}\backslash(S\backslash NP/NP)/N/AP$	$\lambda ANGQz.\exists\delta(\exists x(N(x)\wedge G(\lambda P.P(x))(z)\wedge A(\delta)(x)))$ $\wedge\neg\exists y(N(y)\wedge Q(G(\lambda P.P(y)))\wedge A(\delta)(x))$

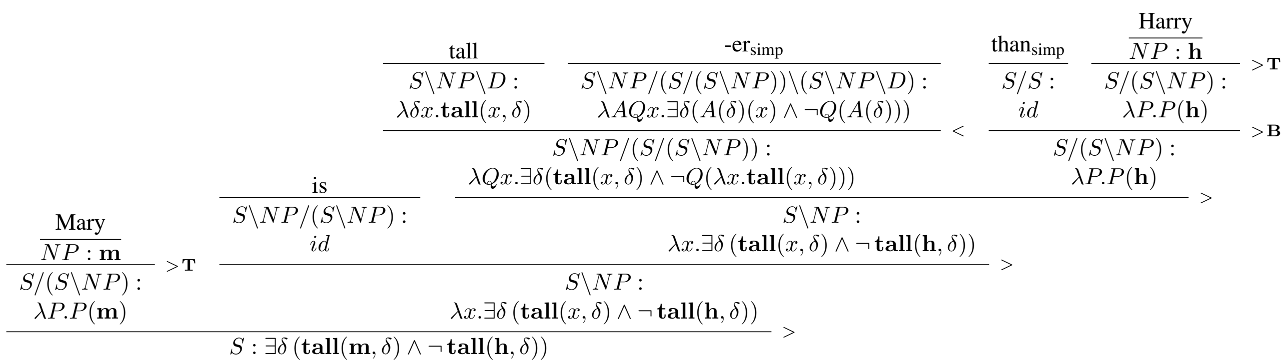


Figure 1: Derivation tree of *Mary is taller than Harry*

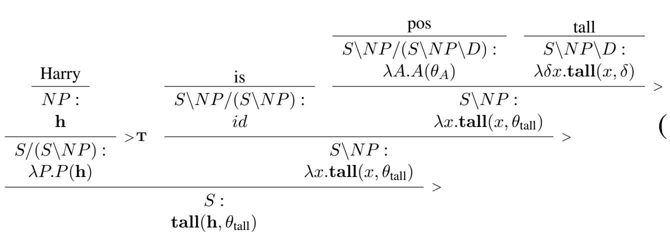


Figure 2: Derivation tree of *Harry is tall*

Quantification When determiners such as *all* or *some* appear in *than*-clauses, we need to consider the scope of the corresponding quantifiers (Larson, 1988). As examples, (15a) and (16a) are assigned the SRs in (15b) and (16b), respectively.

(15) a. Mary is taller than everyone.

b. $\forall y(\mathbf{person}(y)$
$\rightarrow\exists\delta\,(\mathbf{tall}(\mathbf{m},\delta)\wedge\neg\,\mathbf{tall}(y,\delta)))$

(16) a. Mary is taller than someone.

b. $\exists y(\mathbf{person}(y)$
$\wedge\,\exists\delta\,(\mathbf{tall}(\mathbf{m},\delta)\wedge\neg\,\mathbf{tall}(y,\delta)))$

Figure 3 shows a derivation tree for (15a). Here, *everyone* in *than*-clause takes scope over the degree quantification in the main clause. For this purpose, we use the lexical entry for than_{gq} in Table 3, which handles these cases of generalized quantifiers.

Conjunction and disjunction Conjunction (*and*) and disjunction (*or*) appearing in a *than*-clause show different behaviors in scope taking, as pointed out by Larson (1988). For instance, in (17a), the con-

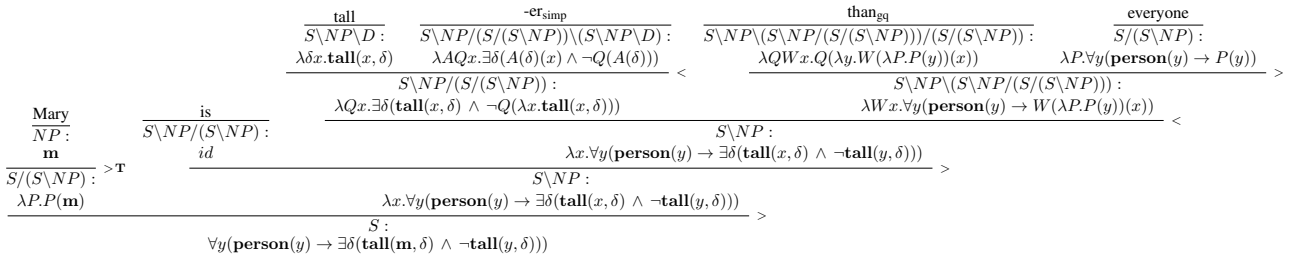


Figure 3: Derivation tree of *Mary is taller than everyone*

junction *and* takes wide scope over the main clause, whereas in (18a), the disjunction *or* can take narrow scope; thus, we can infer *Mary is taller than Harry* from both (17a) and (18a). These readings are represented as in (17b) and (18b), respectively.

(17) a. Mary is taller than Harry and Bob.
b. $\exists\delta\,(\mathbf{tall}(\mathbf{m},\delta) \wedge \neg\,\mathbf{tall}(\mathbf{h},\delta))$
$\wedge\, \exists\delta\,(\mathbf{tall}(\mathbf{m},\delta) \wedge \neg\,\mathbf{tall}(\mathbf{b},\delta))$

(18) a. Mary is taller than Harry or Bob.
b. $\exists\delta(\mathbf{tall}(\mathbf{m},\delta)$
$\wedge \neg(\mathbf{tall}(\mathbf{h},\delta) \vee \mathbf{tall}(\mathbf{b},\delta)))$

The difference in scope for these sentences can be derived by using $\text{than}_{\text{simp}}$ and than_{gq}: $\text{than}_{\text{simp}}$ derives the narrow-scope reading (cf. the derivation tree in Figure 1) and than_{gq} derives the wide-scope reading (cf. the derivation tree in Figure 3).

Attributive comparatives The sentence *APCOM has a more important customer than ITEL* (FraCaS-244/245) can have two interpretations, i.e., (19a) and (20a), where the difference is in the verb of the *than*-clause.

(19) a. APCOM has a more important customer than ITEL is. (FraCaS-244)
b. $\exists\delta(\exists x(\mathbf{customer}(x)$
$\wedge\, \mathbf{has}(\mathbf{a},x) \wedge \mathbf{important}(x,\delta))$
$\wedge \neg(\mathbf{customer}(\mathbf{i}) \wedge \mathbf{important}(\mathbf{i},\delta)))$

(20) a. APCOM has a more important customer than ITEL has. (FraCaS-245)
b. $\exists\delta(\exists x(\mathbf{customer}(x) \wedge \mathbf{has}(\mathbf{a},x)$
$\wedge\, \mathbf{important}(x,\delta))$
$\wedge \neg\exists y(\mathbf{customer}(y) \wedge \mathbf{has}(\mathbf{i},y)$
$\wedge\, \mathbf{important}(y,\delta)))$

We use more_{is} or more_{has} in Table 3 to give the compositional derivations of the SRs in (19b) and (20b), respectively.

4 Inferences with comparatives

We introduce an inference system COMP for logical reasoning with gradable adjectives and comparatives based on the SRs under the A-not-A analysis presented in §3. Table 4 lists some axioms of COMP for inferences with comparatives. Here, F is an arbitrary gradable predicate, F^+ a positive adjective, and F^- a negative adjective.[9]

$(\mathbf{CP})$ is the so-called Consistency Postulate (Klein, 1982), an axiom asserting that if there is a degree satisfied by x but not by y, then every degree satisfied by y is satisfied by x as well. By $(\mathbf{CP})$, we can derive the following inference rule.

$$(\mathbf{CP}\star)\ \frac{\exists\delta\,(\mathbf{F}(x,\delta) \wedge \neg\,\mathbf{F}(y,\delta))}{\forall e(\mathbf{F}(y,e) \rightarrow \mathbf{F}(x,e))}$$

Using this rule, the inference from *Mary is taller than Harry* and *Harry is tall* to *Mary is tall* can be derived as shown in Figure 4.

$$\dfrac{\dfrac{\dfrac{\exists\delta\,(\mathbf{tall}(\mathbf{m},\delta) \wedge \neg\,\mathbf{tall}(\mathbf{h},\delta))}{\forall e(\mathbf{tall}(\mathbf{h},e) \rightarrow \mathbf{tall}(\mathbf{m},e))}\,(\forall\mathbf{E})}{\mathbf{tall}(\mathbf{h},\theta_{\text{tall}}) \rightarrow \mathbf{tall}(\mathbf{m},\theta_{\text{tall}})}\,(\mathbf{CP}\star) \quad \mathbf{tall}(\mathbf{h},\theta_{\text{tall}})}{\mathbf{tall}(\mathbf{m},\theta_{\text{tall}})}\,(\rightarrow\mathbf{E})$$

Figure 4: Example of a proof

$(\mathbf{Ax_1})$ and $(\mathbf{Ax_2})$ are axioms for positive and negative adjectives described in (8). The axioms from $(\mathbf{Ax_3})$ to $(\mathbf{Ax_6})$ formalize the entailment relations between antonym predicates. For instance, the inference of (3) mentioned in §1 is first mapped to the following SRs.

[9] We also use an axiom for privative adjectives such as *former*, drawn from Mineshima et al. (2015).

Table 4: Axioms of COMP

$(\mathbf{TH})$	$\theta_{F^+} > \theta_{F^-}$
$(\mathbf{CP})$	$\forall x \forall y (\exists \delta\, (F(x,\delta) \wedge \neg F(y,\delta)) \rightarrow (\forall e (F(y,e) \rightarrow F(x,e))))$
$(\mathbf{Ax_1})$	$\forall e \forall x (F^-(x,e) \leftrightarrow \forall \delta ((\delta \geq e) \rightarrow F^-(x,\delta)))$
$(\mathbf{Ax_2})$	$\forall e \forall x (F^+(x,e) \leftrightarrow \forall \delta ((\delta \leq e) \rightarrow F^+(x,\delta)))$
$(\mathbf{Ax_3})$	$\forall e \forall x (F^-(x,e) \leftrightarrow \forall \delta ((\delta > e) \rightarrow \neg F^+(x,\delta)))$
$(\mathbf{Ax_4})$	$\forall e \forall x (F^+(x,e) \leftrightarrow \forall \delta ((\delta < e) \rightarrow \neg F^-(x,\delta)))$
$(\mathbf{Ax_5})$	$\forall e \forall x (\neg F^-(x,e) \leftrightarrow \forall \delta ((\delta \leq e) \rightarrow F^+(x,\delta)))$
$(\mathbf{Ax_6})$	$\forall e \forall x (\neg F^+(x,e) \leftrightarrow \forall \delta ((\delta \geq e) \rightarrow F^-(x,\delta)))$

(21)
$$\begin{array}{ll} P_1\colon & \exists \delta (\mathbf{tall}(\mathbf{m},\delta) \wedge (\delta > 4')) \\ P_2\colon & \exists \delta (\mathbf{short}(\mathbf{h},\delta) \wedge (\delta < 4')) \\ \hline H\colon & \exists \delta\, (\mathbf{tall}(\mathbf{m},\delta) \wedge \neg\, \mathbf{tall}(\mathbf{h},\delta)) \end{array}$$

Then, it can be easily shown that H follows from P_1 and P_2, using the axioms $(\mathbf{Ax_2})$ and $(\mathbf{Ax_3})$.

5 Implementation and evaluation

To implement a full inference pipeline, one needs three components: (a) a syntactic parser that maps input sentences to CCG derivation trees, (b) a semantic parser that maps CCG derivation trees to SRs, and (c) a theorem prover that proves entailment relations between these SRs. In this study, we use manually constructed CCG trees as inputs and implement components (b) and (c).[10] For component (b), we use ccg2lambda[11] as a semantic parser and implement a set of templates corresponding to the lexical entries in Table 3. The system takes a CCG derivation tree as an input and outputs a logical formula as an SR. For component (c), we use the off-the-shelf theorem prover *Vampire*[12] and implement the set of axioms described in §4.

Suppose that the logical formulas corresponding to given premise sentences are $P_1, \ldots, P_n$ and that the logical formula corresponding to the hypothesis (conclusion) is H. Then, the system outputs *Yes* if $P_1 \wedge \cdots \wedge P_n \rightarrow H$ can be proved by a theorem prover, and outputs *No* if the negation of the hypothesis (i.e., $P_1 \wedge \cdots P_n \rightarrow \neg H$) can be proved. If both of them fail, it tries to construct a counter model; if a counter model is found, the system outputs *Unknown*. Since the main purpose of this implementation is to test the correctness of our semantic analysis and inference system, the system returns *error* if a counter model is not constructed with the size of an allowable model restricted.

We evaluate our system on the FraCaS test suite. The test suite is a collection of semantically complex inferences for various linguistic phenomena drawn from the literature on formal semantics and is categorized into nine sections. Out of the nine sections, we use ADJECTIVES (22 problems) and COMPARATIVES (31 problems). The distribution of gold answers is: (yes, no, unknown) = (9, 6, 7) for ADJECTIVES and (19, 9, 3) for COMPARATIVES. Table 6 lists some examples.

Table 5 gives the results of the evaluation. We compared our system with existing logic-based RTE systems. B&C (Bernardy and Chatzikyriakidis, 2017) is an RTE-system based on Grammatical Framework (Ranta, 2011) and uses the proof assistant Coq for theorem proving. The theorem proving part is not automated but manually checked. Nut (Bos, 2008) and MINE (Mineshima et al., 2015) use a CCG parser (C&C parser; Clark and Curran, 2007) and implement a theorem-prover for RTE based on FOL and higher-order logic, respectively. LP (Abzianidze, 2016) is a system, LangPro, that uses two CCG parsers (C&C parser and EasyCCG; (Lewis and Steedman, 2014)) and implements a tableau-based natural logic inference system. M&M (MacCartney and Manning, 2008)

[10] CCG parsers for English, such as C&C parser (Clark and Curran, 2007) based on CCGBank (Hockenmaier and Steedman, 2007), are widely used, but there is a gap between the outputs of these existing parsers and the syntactic structures we assume for the analysis of comparative constructions as described in §3. We leave a detailed comparison between those structures to another occasion. We also have to leave the task of combining our system with off-the-shelf CCG parsers for future research.

[11] https://github.com/mynlp/ccg2lambda

[12] https://github.com/vprover/vampire

Table 5: Accuracy on FraCaS test suite. '#All' shows the number of all problems and '#Single' the number of single-premise problems.

Section	#All	Ours	B&C	Nut	MINE	LP	M&M (#Single)
ADJECTIVES	22	**1.00**	.95	.32	.68	.73	.80* (15)
COMPARATIVES	31	**.94**	.56	.45	.48	-	.81* (16)

Table 6: Examples of entailment problems from the FraCaS test suite

FraCaS-198 (ADJECTIVES) Answer: No	
Premise 1	John is a former university student.
Hypothesis	John is a university student.
FraCaS-224 (COMPARATIVES) Answer: Yes	
Premise 1	The PC-6082 is as fast as the ITEL-XZ.
Premise 2	The ITEL-XZ is fast.
Hypothesis	The PC-6082 is fast.
FraCaS-229 (COMPARATIVES) Answer: No	
Premise 1	The PC-6082 is as fast as the ITEL-XZ.
Hypothesis	The PC-6082 is slower than the ITEL-XZ.
FraCaS-231 (COMPARATIVES) Answer: Unknown	
Premise 1	ITEL won more orders than APCOM did.
Hypothesis	APCOM won some orders.
FraCaS-235 (COMPARATIVES) Answer: Yes	
Premise 1	ITEL won more orders than APCOM.
Premise 2	APCOM won ten orders.
Hypothesis	ITEL won at least eleven orders.

uses an inference system for natural logic based on monotonicity calculus. M&M was only evaluated for a subset of the FraCaS test suite, considering single-premise inferences and excluding multiple-premise inferences. These four systems, Nut, MINE, LP, and M&M, are fully automated.

Although direct comparison is impossible due to differences in automation and the set of problems used for evaluation (single-premise or multiple-premise), our system achieved a considerable improvement in terms of accuracy. It should be noted that by using arithmetic implemented in Vampire our system correctly performed complex inferences from numeral expressions such as that in FraCaS-235 (see Table 6). Because we did not implement a syntactic parser and used gold CCG trees instead, the results show the upper bound of the logical capacity of our system. Note also that the five systems (B&C, MINE, LP, M&M, and ours) were developed in part to solve inference problems in FraCaS, where there is no separate test data for evaluation. Still, these problems are linguistically very challenging; from a linguistic perspective, the point of evaluation is to see *how* each system can solve a given inference problem. Overall, the results of evaluation suggest that a semantic parser based on degree semantics can, in combination with a theorem prover, achieve high accuracy for a range of complex inferences with adjectives and comparatives.

There are two problems in the COMPARATIVES section that our system did not solve: the inference from P to H_1 and the one from P to H_2, both having the gold answer *Yes*.

P: ITEL won more orders than the APCOM contract.
H_1: ITEL won the APCOM contract. (FraCaS-236)
H_2: ITEL won more than one order. (FraCaS-237)

To solve these inferences in a principled way, we will need to consider a more systematic way of handling comparative constructions that expects at least two patterns with missing verb phrases.

6 Conclusion

We proposed a CCG-based compositional semantics for gradable adjectives and comparatives using the A-not-A analysis studied in formal semantics. We implemented a system that maps CCG trees to suitable SRs and performs theorem proving for RTE. Our system achieved high accuracy on the sections for adjectives and comparatives in FraCaS.

In future work, we will further extend the empirical coverage of our system. In particular, we will cover deletion operations like Gapping in comparatives, as well as gradable expressions other than adjectives. Combining our system with a CCG parser is also left for future work.

Acknowledgement This work was supported by JSPS KAKENHI Grant Number JP18H03284.

References

Abzianidze, L. (2016). Natural solution to FraCaS entailment problems. In *Proceedings of the Fifth Joint Conference on Lexical and Computational Semantics*, pages 64–74. Association for Computational Linguistics.

Bernardy, J.-P. and Chatzikyriakidis, S. (2017). A type-theoretical system for the FraCaS test suite: Grammatical framework meets Coq. In *Proceedings of the 12th International Conference on Computational Semantics (IWCS)*.

Bos, J. (2008). Wide-coverage semantic analysis with Boxer. In *Semantics in Text Processing. STEP 2008 Conference Proceedings*, pages 277–286.

Clark, S. and Curran, J. R. (2007). Wide-coverage efficient statistical parsing with CCG and log-linear models. *Computational Linguistics*, 33(4):493–552.

Cooper, R., Crouch, R., van Eijck, J., Fox, C., van Genabith, J., Jaspers, J., Kamp, H., Pinkal, M., Poesio, M., Pulman, S., et al. (1994). FraCaS–a framework for computational semantics. *Deliverable*, D6.

Cresswell, M. J. (1976). The semantics of degree. In *Montague Grammar*, pages 261–292. Elsevier.

Hackl, M. (2000). *Comparative Quantifiers*. PhD thesis, Massachusetts Institute of Technology.

Heim, I. (2000). Degree operators and scope. In *Semantics and Linguistic Theory*, volume 10, pages 40–64.

Hendriks, P. (1995). *Comparatives and Categorial Grammar*. PhD thesis, University of Groningen dissertation.

Hockenmaier, J. and Steedman, M. (2007). CCGbank: A corpus of CCG derivations and dependency structures extracted from the Penn Treebank. *Computational Linguistics*, 33(3):355–396.

Kennedy, C. (1997). *Projecting the Adjective: The Syntax and Semantics of Gradability and Comparison*. PhD thesis, University of California, Santa Cruz.

Klein, E. (1980). A semantics for positive and comparative adjectives. *Linguistics and philosophy*, 4(1):1–45.

Klein, E. (1982). The interpretation of adjectival comparatives. *Journal of Linguistics*, 18(1):113–136.

Kubota, Y. and Levine, R. (2015). Against ellipsis: arguments for the direct licensing of 'noncanonical' coordinations. *Linguistics and Philosophy*, 38(6):521–576.

Larson, R. K. (1988). Scope and comparatives. *Linguistics and Philosophy*, 11(1):1–26.

Lassiter, D. (2017). *Graded Modality: Qualitative and Quantitative Perspectives*. Oxford University Press.

Lewis, M. and Steedman, M. (2014). A* CCG parsing with a supertag-factored model. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 990–1000. Association for Computational Linguistics.

MacCartney, B. and Manning, C. D. (2008). Modeling semantic containment and exclusion in natural language inference. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling)*, pages 521–528.

Martínez-Gómez, P., Mineshima, K., Miyao, Y., and Bekki, D. (2016). ccg2lambda: A Compositional Semantics System. In *Proceedings of ACL 2016 System Demonstrations*, pages 85–90.

Mineshima, K., Martínez-Gómez, P., Miyao, Y., and Bekki, D. (2015). Higher-order logical inference with compositional semantics. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2055–2061.

Narisawa, K., Watanabe, Y., Mizuno, J., Okazaki, N., and Inui, K. (2013). Is a 204 cm man tall or small? Acquisition of numerical common sense from the web. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 382–391. Association for Computational Linguistics.

Pulman, S. (2007). Formal and computational semantics: a case study. In *Proceedings of the Seventh International Workshop on Computational Semantics (IWCS)*, pages 181–196.

Ranta, A. (2011). *Grammatical Framework: Programming with Multilingual Grammars*. CSLI Publications.

Schwarzschild, R. (2008). The semantics of comparatives and other degree constructions. *Language and Linguistics Compass*, 2(2):308–331.

Seuren, P. A. (1973). The comparative. In *Generative Grammar in Europe*, pages 528–564. Springer.

Stechow, A. v. (1984). Comparing semantic theories of comparison. *Journal of Semantics*, 3(1-2):1–77.

Steedman, M. J. (2000). *The Syntactic Process*. The MIT Press.

van Rooij, R. (2008). Comparatives and quantifiers. *Empirical Issues in Syntax and Semantics*, 7:423–444.

A Type-Theoretical Approach to Register Classification

Hou Renkui
Guangzhou University,
China
hourk0917@163.com

Huang Chu-Ren
The Hong Kong Polytechnic University,
Hong Kong;
churen.huang@polyu.edu.hk

Abstract:

We propose to differentiate different registers based on the distribution of different Parts of Speeches. Based on a type-theoretical approach, grammatical categories are defined by their combinatory and mapping functions. With noun as the basic category representing entities, verbs are functions taking them as arguments; and adverbs are functions taking verbs as arguments. Based on this different functional mapping relations, we hypothesis that their ratio, like unit-constituency ratios, can differentiate different types of texts, and especially registers. We calculated the ratios between grammatical categories based on their function mapping relations. For example the ratio between verbs and nouns, and adverbs and verbs. The boxplots was used to show the distribution of the ratios between these parts of speeches in each register. The linear regression was used to verify the differences of these ratios in different registers. The text clustering result showed that these ratios can differ conversational and written registers.

Keywords: Chinese register, Parts of speeches, Linear regression, Text clustering

1 Introduction

Grammatical categories, also known as parts of speech (PoS), are ubiquitous attributes that can be assigned to each word in a text. Biber and Conrad (2009) pointed out that most people do not notice common language features such as nouns and pronouns; they are pervasive features that are so common that speakers normally do not even notice their existence. Hence grammatical categories are not often used in register studies. Even when they are used, they are used among a bundle of features and no clear explanatory model has been proposed to link grammatical categories to genres. Yet, given that register is often considered as the most important perspective on text varieties (Biber & Conrad 2009) and that grammatical categories are the most fundamental morpho-syntactic attributes, is there no direct relation between them. Recent work (Hou et al. 2019a, Hou et al. 2019b) showed that the relation between different linguistic units and levels offers powerful tools to capture textual variations, following the spirit of the Menzarath-Altmann law as the modeling the relation between linguistics units and their constituents. Following this rationale, we adopt a type-theoretical view (Steedman 1989) to view grammatical categories as defined as mappings between other more basic categories, starting from sentence being defined as truth-value, and nouns as entities. In this view, nouns are basic categories; verbs are first order predicates as functions mapping nouns to sentences; adjectives also the first order predicates as functions mapping nouns to nouns; and adverbs as second-order predicates as the function mapping verbs (as first-order predicate) to verbs. Hence, we can view these four grammatical categories (or PoS's) as linguistic units linking to each other in a hierarchical mapping relations. Given these functional mapping relations, we can view them as another layer of text-internal relations

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 57-67, Hakodate, Japan, September 13-15, 2019

that are subject to self-adaptation and hence as good features for register classification. In this paper, based on the type-theoretical view on grammatical categories, we propose to use the ratios between two pairs of PoS's with direct functional mapping relations (i.e. nouns/verbs, and verbs/adverbs) to model different registers.

Words can be classified in different ways if one uses at least a comparative property. Traditionally, grammatical categories of words are defined purely on either syntactic or semantic criteria (Huang, Hsieh and Chen 2017). In grammar the usual inherited from Latin is to classify words into parts of speeches like nouns, verbs, adjectives etc. The purpose of this kind of classification is to portray different usage of different words and grammatical structures of sentences. There are many studies to explore the taxonomy and usage of Chinese parts of speeches. It is generally thought that words can be classified as the content words and function words rich in grammatical meaning. With lexical and grammatical meaning, content words can independently act as syntactic components. Content words can be divided as the nouns, verbs and adjectives, etc. (Huang, Hsieh and Chen 2017). One cannot speak about the "correctness" of a classification but rather of its aim and conceptual background. This paper explored the ratios between thematic words including nouns, first-order predicates which modify the nouns and adverbs which modify the first-order predicates in various registers.

1.1Related literature

Huang, Hsieh and Chen (2017) addressed the need to provide users and public with a full account of the Part of Speech classification framework and its criteria. Lexical analysis including word segmentation and parts of speech tagging is an important task in Chinese natural language processing. We selected the existing parts of speech system as the classification framework of words.

Biber (1994) used register as a cover term for all language variations associated with different situations and purposes. Linguistic characteristics are one of major perspectives in register analysis. For example, many previous studies provided the assumption that function words offer the best evidence to differentiate various registers (e.g., Zeng 2008, Zhang 2012). Biber (1993) proved that there are differences in parts of speech and syntax in different registers. Some studies showed that different register texts also differ in some textual features, such as order sequence of the most frequently used words (Hoover 2002), part of speech histogram (Feldman et al. 2009, Hou and Jiang 2016). Shah and Bhattacharyya (2002) concluded that the content words, i.e., nouns, verbs, adjectives and adverbs, can differ different type texts effectively through studying the five types of texts in BNC. Zhang (2012) demonstrated that the different text types prefer to use different PoS.

Kõler (2012) advocated incorporation of quantitative mathematical approaches in linguistic studies. Cramer (2005) proposed that investigating the statistical aspects of language advances natural language processing research, as well as basic linguistic research. Register can also be studied using such mathematical methods. Biber (1986, 1988) is generally credited for introducing quantitative methods to the register study. Biber (1995) restated and underlined the role of computational, statistical, and interpretive techniques using multi-dimensional analysis. He proposed that any text characteristic that is encoded in language and can be reliably identified and counted is a candidate for inclusion. Research on register characteristics has also been undertaken from the perspective of quantitative linguistics. For example, Hou, Huang, and Liu (2017) fitted the distribution of Chinese sentence lengths using nonlinear regression and used the fitted

parameters as quantitative features of the corresponding Chinese registers. Hou et al (2019a) fitted the relationship between the Chinese clause length and word length based on Menzerath-Altmann law and showed that the fitted parameters can differ various Chinese registers. Hou et al (2019b) used these fitted parameters to calculate the formality degree and the distance between different Chinese registers.

Biber's (1994) observation of the lack of agreement on the definition and taxonomy of register also applies to the study of registers in Chinese. Feng (2010), on the other hand, proposed that register is a polarized opposite continuum, with the formal written and daily colloquial registers being the two poles, and others lying in between. He thought the register is generated in interpersonal communication and that the essence of register is to adjust the psychological distance between communicators. We adopt Biber's (1994) position to reconcile the above differences: that registers are varieties in a continuum, but which can be analytically identified as different categories.

1.2 Research question and methodology

This paper explored the ratios between parts of speeches in various registers. The occurrence frequencies of these parts of speeches including nouns, first-order predicates and adverbs should be calculated. Then the different distribution of these ratios are manifested visually using boxplots. The linear regression, ratio as the dependent variable and register as the independent variable, are used to compare the group mean of these ratios in various registers. Then the text clustering analysis showed that the texts from the spoken and written registers are distinguished when the text is represented by these four ratios. The open source programming language and environment R (R Core Team 2011) was used to realize the linear regression, and text clustering.

2. Corpus establishment

Effective register analyses are always comparative. It is virtually impossible to know what is distinctive about a particular register without comparing it to other registers (Biber and Conrad 2009: 36). Hence, we selected the texts from three registers to establish the corpus and to study the differences of linguistic characteristics of these three registers.

News Co-Broadcasting, as a program of Central China TV, mainly gives a brief introduction to important state policies and events taking place at home and abroad. It is characterized by the formal, serious and solemn use of language and can represent *News Broadcasting* register. It objectively reports the news facts, and should not and rarely used exaggerated words to describe news events. *Behind the headline with Wentao*, as a program of Phoenix satellite TV, the host discusses some current hot issues together with guests in TV. They talk freely face to face, chatting so as to deliver recreational information, create fun and discriminate truth from falsehood, not focusing on the "right answers" to the issues. They do not read the scripts edited ahead of time. The conversation are produced from the host and guests in the TV immediately after thinking. They share the communication environment and time. The speaker is able to use the similar context of utterances with the hearers. It can represent *TV talk show* register. The *Science* papers report the scientific facts, new findings, etc. and interpret this finding, or formulize the new theory. The objectivity and precise are the most important key points of the scientific papers. Scientific register are more explicit and abstract, and have less interpersonal and affective content and fewer narrative concerns than spoken register (Gardner, Nesi and Biber 2018). The paper are not produced and read by

scientists and readers at the same time. Generally speaking, the readers and writers of scientific papers are professional in the same domains. So, they share the same professional background of knowledge. This will influence the produce of papers for writers.

The number of the collected texts are 100 in both *News Co-broadcasting* and *Science* registers, and 101 in *Behind the Headlines with Wentao*. The texts from these three registers were segmented and Parts of Speech tagged using the Chinese Lexical Analysis System created by the Institute of Computing Technology of the Chinese Academy of Sciences (ICTCLAS). The parts of speech tag set from ICTCLAS has been revised in order to be used in natural language processing because some words have different grammatical functions.

3 Experiment

Themes of texts are represented by the nouns and first-order predicates which are used to modify the nouns and composed of verbs and adjectives. The adverbs, as second-order predicates, can modify the verbs and adjectives. We firstly computed the ratios between first-order predicates and nouns respectively. Then we computed the ratios between adverbs, as the second-order predicate, and the first-order predicates.

3.1 Ratios between the first-order predicates and the nouns

The ratios between verbs and nouns are computed in texts from various registers. The nouns and verbs are the two major parts of speeches in human languages. The former represents the concrete and abstract concepts in the world. The latter represents the action from the subjects or on the objects. The subject-predicate and verb-object structures are also produced when these two parts of speeches are combined. The boxplot was used to visually manifest the distribution of this ratio in various registers, as shown in Figure 1. In Figure 1, the *n* represents the nouns, the *v* and *a* represent verbs and adjectives respectively, the *d* represent the second-order predicate, names adverbs.

The high of the square in the Figure 1 can represent the dispersion of the ratio distribution. The bottom and top lines of the square represent the 25% and 75% quartile of the frequency distribution of the ratios respectively. The more the difference between these two values is, the more disperse of the data is.

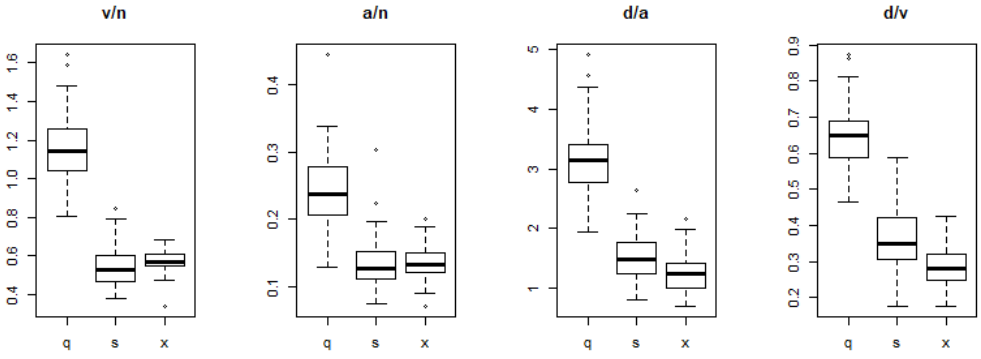


Figure 1：The distribution of ratios between various parts of speeches across registers (“q” represents *TV Talk Show*, “s” represents *Science* texts, “x” represents *News Broadcasting*)

From the left panel 1 in Figure 1, we can see that the distribution of ratios between verbs and nouns in *TV Talk Show* is significantly larger than the *Science* and *News Broadcasting*. And, the dispersion of this ratio distribution in *TV Talk Show* is larger than the other registers. This means that the ratios between verbs and nouns are dispersive maybe because the different guests in different time have different language usage characteristics. For example, some guests are used to omit the subjects in talking and others not. The small dispersion in *News Broadcasting* shows that there are high consistency degree of these ratios in programs. The dispersion of the ratios in *Science* shows that there are some differences in different scientific fields.

The linear regression can fit relationship between one ordinal variable and one category variable. We can use the linear regression, *lm*() function in R programming, to test whether there are significant differences of group means of the ratios between verbs and nouns in various registers using the register as the independent variable. The regression result is shown in Table 1.

Table 1: The regression result between ratios and registers

	estimate	std. error	t value	pr (>| t |
（Intercept）	1.157	0.011	105.61	< 2e-16
types	-0.615	0.015	-39.62	< 2e-16
typex	-0.560	0.015	-37.34	< 2e-16

In Table 1, intercept estimate, 1.157, represents the group mean of ratios between the number of verbs and nouns in *TV Talk Show*. The *p*-value for the intercept showed that the estimate is unlikely to be zero. There are two additional coefficients, "types" for the contrast between the group mean of ratios in *Science* and *TV Talk Show*, and "typex" for the contrast between the group mean of ratios in *News Co-broadcasting* and *TV Talk Show* respectively. The two *p*-values demonstrate that these two contrasts are significant. Hence, we can reconstruct the other two group means from this regression result. The group mean ratio of verbs on nouns in *Science* texts, smaller than in *TV Talk Show*，is 1.157-0.615=0.542. The group mean ratio of verbs on nouns in *News Co-broadcasting* is 1.157-0.56=0.597. From the above analysis, there is one comparison that is left out to be examined. When a factor (register) has 3 levels, i.e. *TV Talk Show*, *Science* and *News Co-Broadcasting*, there will be one comparison that does not appear in the Table 1, the contrast between group means of ratios in *Science* and *News Co-broadcasting*. Similarly, we use regression analysis to examine that comparison. The regression result, as shown in Table 2, demonstrated that the mean ratio in *News Co-Broadcasting* is little larger than in *Science*, the difference is 0.055. The *p*-value shows that this difference is significant. From Table 2, the intercept is 0.542 which is the group mean of ratio in *Science* texts.

Table 2: The result of the linear regression of the ratios between verbs and nouns in *Science* and *News Broadcasting*

	Estimate	std. error	t value	pr (>| t |
（Intercept）	0.542	0.008	68.98	< 2e-16
typex	0.055	0.011	3.195	0.00163

TV Talk Show is the face to face conversation register. The nouns acting as the subject or object are often omitted. This leads to the higher occurrence frequency of the verbs than the nouns. The speaker is producing language at the same time that he is thinking about what he wants to say. Some context, background of knowledge between speaker and listener are identical. According to the cooperation principle in discourse analysis, the speaker omits the components in the sentence that he think the hearer should understand in order to improve the communication proficiency. The identical context between the communicators can help the listener supplement the missing components.

In *Science* and *News Broadcasting*, the writers and readers/hearers locates different places and there is no interactions between them. They don't share the same environment and the same context. All the language components which may impact communication cannot be omitted. In these two registers, the sentences are often the complete subject-predicate-object construction, hence the occurrence frequencies of nouns are more than the occurrence frequencies of verbs. The aim of *News Broadcasting* is to narrate and report past events and describe some state of affairs. There are some interviews in some *News Broadcasting* texts in which the occurrence frequencies of verbs are larger than the nouns because of the omitting of some subjects or objects. The aims of *Science* papers are explaining and interpreting information, arguing or persuading, providing procedural information about how to perform certain activities. The nouns representing the information, argument, etc. are plenty even if some nouns are omitted because of the shared knowledge between writers and readers. This leads to the low ratio between the number of verbs and nouns.

The distributions of ratios between the occurrences of adjectives and nouns in each register are shown in panel 2 on the left in Figure 1. The ratios between the occurrence frequencies of adjectives and nouns in *TV Talk Show* are also greater than that in *News Broadcasting* and *Science* obviously. Linear regression was used to fit the relationship between ratios and the registers. The result of regression, as shown in Table 3, demonstrated that the group mean of ratios in *TV Talk Show* is 0.242. The negative estimates of "types" and "typex" showed that these group mean of ratios in *Science* and *News Co-broadcasting* are smaller than in *TV Talk Show*. The *p*-values showed that these two contrasts, *TV talk Show* and *Science*, *TV Talk Show* and *News Co-broadcasting* are significant.

Table 3: The result of linear regression of relationship between ratios and the registers

	estimate	std. error	t value	pr (>| t |
（Intercept）	0.242	0.004	65.86	< 2e-16
types	-0.111	0.005	-21.19	< 2e-16
typex	-0.107	0.005	-20.43	< 2e-16

One of the aims of *TV Talk Show* is entertaining the addressee and revealing personal feelings or attitudes. Speakers often use adjectives to modify their attitudes, feeling in order to meet the need of communication in that context. This leads to the high ratio between the occurrences frequencies of adjectives and nouns in *TV Talk Show*. However, the description of event in *News Broadcasting* and the explaining and the interpretation of information in *Science* papers

should be objective. The few adjectives are selected to modify the nouns in these two registers. This leads to the small ratios between occurrences frequencies of the adjectives and the nouns.

Table 4: The result of linear regression of the relationship between ratios and register in *News Broadcasting* and the *Science*

	estimate	std. error	t value	pr (>| t |
（Intercept）	0.132	0.003	46.51	< 2e-16
typex	0.004	0.004	0.997	0.32

The distribution of ratios between the occurrences frequencies of adjectives and nouns in *News Broadcasting* and *Science* are similar from the boxplot in Figure 1. The linear regression, as shown in Table 4 in which *p*-value is larger than 0.05, also demonstrated that there are not significant differences of group means of ratios between the occurrence frequencies of adjectives and nouns in *Science* and *News Co-broadcasting*.

3.2 Ratios between adverbs and the first-order predicates

Then, we discussed the ratios between occurrences frequencies of the adverb and the first-order predicates, adjectives and nouns, as shown on the panel 3 and 4 from the left in the Figure 1.

In one text, the number of adverbs is constant when computing the ratios between adverbs and adjective, between adverbs and verbs. In *TV Talk Show*, as mentioned above, the adverbs are used more frequently to modify the adjective and verbs because of its communicative purpose, for example degree adverbs and negative adverbs. So these two ratios in *TV Talk Show* are higher than in other two registers. The regression analysis also showed that the group mean of this ratio in *TV Talk Show* is higher than in other two registers as can be seen from Table 5 and 6.

Table 5: Regression result of ratios between adverbs and adjective in three registers

	estimate	std. error	t value	pr (>| t |
（Intercept）	3.146	0.040	78.11	< 2e-16
types	-1.623	0.057	-28.43	< 2e-16
typex	-1.912	0.057	-33.50	< 2e-16

Table 6: Regression result of ratios between adverbs and verbs in three registers

	estimate	std. error	t value	pr (>| t |
（Intercept）	0.646	0.007	88.54	< 2e-16
types	-0.279	0.010	-26.98	< 2e-16
typex	-0.363	0.010	-35.12	< 2e-16

Table 5a: Regression result of ratios between adverbs and adjectives in *News Broadcasting* and *Science*

	estimate	std. error	t value	pr (>| t |
（Intercept）	1.523	0.033	46.586	< 2e-16
typex	-0.289	0.046	-6.258	2.36e-09

Table 6a: Regression result of ratios between adverbs and verbs in *News Broadcasting* and *Science*

	estimate	std. error	t value	pr (>| t |
（Intercept）	0.367	0.0072	50.612	< 2e-16
typex	-0.084	0.0102	-8.205	2.89e-14

These two ratios in *News Co-broadcasting* and *Scienc*e are significantly different, especially for the ratio between adverbs and verbs as shown in Figure 1. The linear regression result of ratio in these two registers were shown in Table 5a and 6a. In *News Co-Broadcasting*, the positive adjectives are more frequently used. Maybe this is the reason which lead to the small ratio between adverbs and adjectives. In *Science*, the authors focus on the new scientific facts rather than on the action. The few verbs usages lead to the relative high ratios between adverbs and verbs in the *Science* papers. One of the important aims of the *News Broadcasting* is to report the events happed that day. This leads to the high frequency of verbs and the smaller ratios between the number of adverbs and verbs.

From the above analysis, these ratios in *TV Talk Show* are significantly higher than in other two registers.

3.3 Text clustering

The texts were represented by these four ratios in each register. In text clustering, the Euclidean distance was used to calculate the distances between the texts. Ward's method (Error Sum of Square Criterion) was selected to calculate the distance between clusters. If the clustering result is good, we can say that these four ratios can differ the selected registers and can be used as the register characteristics. So, text clustering analysis is our approach instead of purpose in another words. The clustering result is shown in Figure 2 and Table 7.

From the Table 7, we can see that the texts in cluster 1 are from *TV Talk Show* register, the cluster 2 and cluster 3 are both composed of *Science* and *News Broadcasting* texts respectively. The texts from *News Broadcasting* and *Science* are merged into one cluster when we cut the dendrogram into 2 clusters. The distances between texts or clusters are measured by the height of their common ancestor, the higher the common ancestor, the far of the distance, and vice versa. Figure 2 shows that the texts from *TV Talk Show*, left branch, has a far distance from the other register texts, right branch. The texts from *News Broadcasting* and *Science* are close and not separated each other. Hence, we can say these four ratios can differ conversation register and written formal register including *News Broadcasting* and *Science* and cannot differ the subset of written register. They can be used as the distinctive characteristic of conversation and written formal registers. The advantage of this characteristic of register is that it is calculated easily and not affected by the text length.

Table 7: The agglomerative hierarchical clustering result of texts

	Cluster 1	Cluster 2	Cluster 3
TV Talk Show	101	0	0
Science	1	55	44
News Co-broadcasting	0	90	10

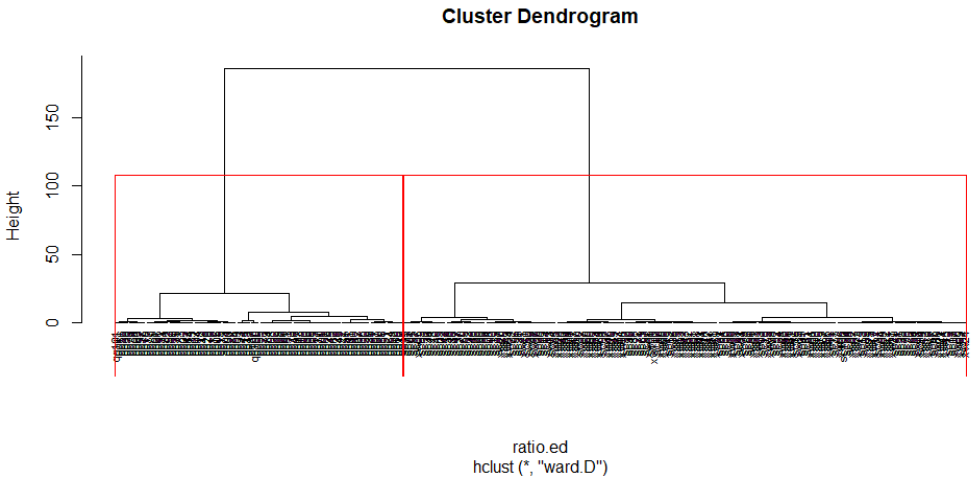


Figure 2：The agglomerative hierarchical clustering result of texts (left brunch is *TV Talk Show*)

4 Conclusion

This paper explored the linguistic characteristics of Chinese registers which are *News Broadcasting*, *Science* and *TV Talk Show* based on the Parts of Speeches of content words. Firstly, we discussed the ratios between different parts of speeches including nouns, first-order predicates and their modifier, i.e. adverbs because the words of these parts of speeches can represent the contents of the texts. The experiments showed that these ratios in *TV Talk Show* are more different from the other two registers. The boxplot showed these differences visually. The linear regression, in which register is used as independent variable and the ratio as dependent variable, verified this point. The clustering analysis showed that the clusters of *TV Talk Show* are far away from another clusters including texts from *Science* and *News Broadcasting*.

From this, we can say the ratios between these parts of speeches can differ conversation register and written formal registers. This characteristic avoid the influence of text length and text number from various registers. Furthermore, these ratios are easy to calculate.

Acknowledgements. We would like to thank the anonymous reviewers for their insightful and helpful comments.

Funding support. Research on this paper was funded by National Social Science Fund in China (Grant No. 16BYY110), the Hong Kong Polytechnic University Grant 4-ZZFE.

References

Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. Language. 62(2):384-413.

Biber, D. (1988). Variation across Speech and Writing. England Cambridge: Cambridge University Press.

Biber, D. (1993). Using register-diversified corpora for general language studies. Computational linguistics. 19(2): 219-241.

Biber, D. (1994). An analytical framework for register studies. In D, Biber and E. Finegan Eds. *Sociolinguistic perspectives on register*, pp.31-56.

Oxford: Oxford University Press.

Biber, D. (1995). On the role of computational, statistical, and interpretive techniques in multi-dimensional analyses of register variation: A reply to Watson. *Text-Interdisciplinary Journal for the Study of Discourse*, *15*(3), 341-370.

Biber, D., and Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge University Press.

Cramer, I. (2005). The parameters of the Altmann-Menzerath law. Journal of Quantitative Linguistics, 12, 41–52.

Feldman, S., M. A. Marin, M. Ostendorf and M. R. Gupta. (2009). Part-of-speech histograms for genre classification of text. In *Proceedings of the 2009 IEEE International Conference on Acoustics, Speech and Signal Processing*. Washington, DC. pp 4781–4784.

Feng, S (2010). On mechanisms of Register System and its grammatical property. Studies of the Chinese Language. 400-412.

Gardner, S., Nesi, H., & Biber, D. (2018). Discipline, level, genre: Integrating situational perspectives in a new MD analysis of university student writing. *Applied Linguistics*. https://doi.org/10.1093/applin/amy005

Hoover, D. L. (2002). Frequent word sequences and statistical stylistics. Literary and Linguistic Computing. 17(2): 157-180.

Hou, Renkui, Jiang Yang and Minghu Jiang. (2014). A Study on Chinese Quantitative Stylistic Features and Relation Among Different StylesBased on Text Clustering. Journal of Quantitative Linguistics. (21)3: 246-280.

Hou, R., & Jiang, M. (2016). Analysis on Chinese quantitative stylistic features based on text mining. Digital Scholarship in the Humanities, 31(2): 357-367.

Hou, R., Huang, C., & Liu, H. (2017). A study on Chinese register characteristics based on regression analysis and text clustering. Corpus Linguistics and Linguistic Theory. doi:10.1515/cllt-2016-006.

Hou, R., Huang, C.-R., San DoH., and Liu, H. (2017). A study on correlation between Chinese sentence and constituting clauses based on the Menzerath-Altmann Law. Journal of Quantitative Linguistics, 24(4): 350–66.

Hou, Renkui, Chu-Ren Huang, Kathleen Ahrens and Yat-Mei Sophia Lee. (2019a). Linguistic characteristics of Chinese register based on the Menzerath-Altmann law and text clustering. Digital Scholarship in the Humanites. Doi:10.1093/llc/fqz005.

Hou, Renkui, Chu-Ren Huang, Mi Zhou and Menghan Jiang. (2019b). Distance between Chinese Registers Based on the Menzerath-Altmann Law and Regression Analysis. Glottometrics. Vol. 45: 24-56.

Huang, Chu-Ren, Shu-Kai Hsieh and Keh-Jian Chen. (2017). Mandarin Chinese Words and Parts of Speech. London, England: Routledge.

Köhler, R. (2012). Quantitative syntax analysis (Vol. 65). Berlin: Walter de Gruyter.

Lv, S. (1992). Studies on Chinese grammar through comparison. Foreign Language Teaching and Research. (2).

R Development Core Team (2011). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org/.

Shah, C & P. Bhattacharyya. (2002). A Study for Evaluating the Importance of

Various Parts of speech (POS) for Information Retrieval(IR). Presented at International Conference on Universal Knowledge and Languages, Goa, India.

Steedman, M. J. (1989). Constituency and coordination in a combinatory grammar. In Baltin, M. R. and Kroch, A. S. (Eds.), Alternative Conceptions of Phrase Structure, pp. 201-231. University of Chicago, Chicago.

Zeng, Y. (2008). An analysis on the Type Differentiation of Language. Journal of Fujian Normal University (Philosophy and Social Sciences Edition). No.2: 34-40.

Zhang, Z. (2012).A Corpus Study of Variation in Written Chinese. Corpus Linguistics and Linguistic Theory. Vol.8, No.1, pp.209-240.

Modeling the Idiomaticity of Chinese Quadra-syllabic Idiomatic Expressions

Shu-Kai Hsieh **Yu-Hsiang Tseng** **Chiung-Yu Chiang**
Graduate Institute of Linguistics
National Taiwan University
shukaihsieh@ntu.edu.tw

Abstract

This paper proposes a computational model of idiomaticity for Chinese Quadra-syllabic idiomatic expressions based on variations, compoundness and compositeness measure. Two classification experiments are conducted to test the model, together with linguistic analysis of the connection to wordnet. The result is promising and we believe that it will shed more light on our understanding of cognitive dynamics that underlies multiword expressions processing.

1 Introduction

Multiword expressions (MWEs) as the *habitual recurrent word combinations* in our daily language use have been regarded as the bottleneck in current NLP technology. In this paper, we will focus on a special type of idiomatic expressions of even length in Chinese called Quadra-syllabic Idiomatic Expressions (QIEs), which have pervasive presence in the Sinosphere (e.g., Japan, Korea, Vietnam, and other ethnic groups like the Naxi) due to the influence of emblematic logographic writing systems (Tsou, 2012).

Traditionally, idioms/idiomatic expressions are defined as MWEs for which the semantic interpretation is not a compositional function of their composing units. Over the past years, a rich amount of analytic works on them for mainly European languages has been proposed. Main efforts have been made to their linguistic and statistic characteristics, and the computational treatment as well. However, due to the lack of cross-language comparative work, QIEs as an idiosyncratic and indispensable part in Chinese and other languages haven't been well studied in the area of current MWE paradigm.

From the usage-based emergentist perspective, as one type of MWEs, Chinese QIEs are characterized by a holistic storage format that reveals high-level entrenchment and constructionist accounts of complex linguistic strings in the minds of language users. However, it is also noted that corpus evidence and acceptability ratings support that idioms are subject to variation too (Geeraert et al., 2017). This paper takes the challenge in modeling the QIE's idiomatic behaviour along three crucial dimensions, and explores their mapping to the synset of Chinese wordnet.

2 Chinese QIEs

The notion of *idiomaticity* has been proposed since (Chafe, 1968) and the issues debated in NLP have been well-recognized (Sag et al., 2002).

Quadra-syllabic Idiomatic Expressions (QIEs) in Chinese can be considered as a special type of idiomatic expressions of even length (i.e., four characters). In this paper, we further divide QIEs into two main types: idioms ('chengyu') and prefabs (Hsieh et al., 2017). Idiom-QIEs often involves Locus Classicus and awareness of cultural background with classical Chinese, they are formed through ages of constant use, well-compiled in dictionary and learned in school (e.g., 化險為夷 hua4 xian3 wei2 yi2,'turn danger to safety'). With their archaic origins, idiom-QIE in particular, are still prevalent in modern use and behaves more vividly than its synonyms represented by common lexemes.

tsou2012 observes some defining characteristics of QIEs which cannot find direct equivalents in En-

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 68-75, Hakodate, Japan, September 13-15, 2019

glish: they consist of four syllables or logographs, have relatively fixed structure and patterns, and carry figurative meaning and semantic opacity. Prefabs-QIEs, on the other hands, are more compositionally dependent, direct results of language use. They are mainly conventional combination of four morphemes taken up and reproduced by speakers they heard before. It can be understood as the *variations-tolerant* lexical bundles composing of four characters/morphemes (e.g., 好久不見 hao2 jiu3 bu2 jian4 'long time no see').

3 QIEs model of idiomaticity

This section introduces our proposed computational model of idiomaticity for Chinese QIEs. The model is based upon idiomaticity theories in linguistics and leveraged resources in Chinese Wordnet (CWN).

Idioms are complex linguistic and psychological configurations. Researchers proposed various theories and frameworks to describe aspects of linguistic construct and processing of idioms (Healy, 1994; Fernando, 1996), where different definitory dimensions are used to capture the nature of idiomaticity (Langlotz, 2006). Basing on previous literature, this paper described idiomaticity of Chinese QIEs along three dimensions:

1. **Variation** indicates the degree of conventionalization of QIEs. Idioms have gone through a socio-linguistic process through which the speakers became familiar and conventionalize the expression. The resulting construct became unitized (Healy, 1994) or frozen ("recalcitrance to undergo transformations") (Fraser, 1970). That is, the constituents of a QIE cannot be replaced or altered in actual usage.

2. **Compoundness** denotes the degree of idiosyncrasies in QIEs' compound structure. Past studies argued English idioms showed constructional idiosyncrasies, such as *trip the heavy fantastic*, which is otherwise ungrammatical (Langlotz, 2006). Similarly, Chinese idioms etymologically came from classical Chinese, their morphology and grammatical rules are different from contemporary Mandarin Chinese when compounding single-character words into QIEs.

3. **Compositeness** represents the extent of semantic un-compositionality, namely *opaqueness*, of QIEs. The uncompositional nature is the defining feature of idiom, that is the meaning of the idioms is not the compositional results of their constituent parts. Therefore two levels of meanings are to be distinguished: the literal meaning (the sum of constituent meanings) and the idiomatic meaning (the lexicalized meaning of the idiom). The more distinct these two levels of meanings of an idiom has, the more *opaque* an idiom is.

The computational model of idiomaticity formalized variation, compoundness, and compositeness with three indices respectively. These indices not only shed light on the nature of any given QIEs, but facilitate QIE candidates selection when incorporating QIEs into CWN.

3.1 Variation

Variation measures the extent of lexicogrammatically restriction of a QIE. The restriction is operationalized as usage variation frequency of a QIE in a corpus. These variations were further defined as two types: (1) substitution, where the second or the third character of a QIE was replaced by another character; and (2) insertion, where characters were placed between the spaces of the four characters in a QIE. The frequency of these variation patterns were identified and summed together, along with the frequency of the QIE itself, the index of variation can be computed as:

$$\text{variation} = \log \frac{\text{variations frequency} + 1}{\text{QIE frequency}} \qquad (1)$$

Higher variation values indicate more substitution or insertion patterns could be found for a given QIE, therefore the QIE is less likely to be *frozen*. For example, 狂風驟雨 kuáng fēng zòu yǔ "raining cats and dogs" is less conventionalized (variation $= 1.58$), since it is frequently found with the third character replaced with 暴 bào "fiercely" without changing the meaning. On the contrary, a low variation value implied a QIE is more likely to be conventionalized, thus fewer variation can be observed in corpus. For instance, 刮目相看 guā mù xiāng kàn "revere with

respect" has no variation form observed in corpus (variation $= -5.73$).

3.2 Compoundness

Compoundness indicates how probable the compound structure of a QIE follows morphological or grammatical rules in contemporary Mandarin Chinese. For instance, the idiom 虎頭蛇尾 hǔ tóu shé wěi "working industriously at first but carelessly in the end", followed common morphological rules in Mandarin Chinese. The first two characters 虎頭 literally mean "the head of a tiger", and the last two characters 蛇尾 is "the tail of a snake". The two parts both follow the same common, or probable, compound structure in Chinese word morphology. By contrast, the idiom 來龍去脈 lái lóng qù mài "the preceding and succeeding contexts of a subject matter" does not follow a common Chinese word compound rule. The first character 來 is often used as an adverb, but it seldom precedes a noun such as 龍 lóng "dragon". Similarly, the third character 去, which has comparable grammatical role as 來, is not commonly followed by a noun 脈 mài "context". Therefore this idiom has less probable compound structure.

To capture the common morphological rules in Chinese words, we constructed a morphological graph between Chinese words and characters. The graph incorporated the productive word morphology in Chinese, along with lexical and semantic relations encoded in CWN. From the morphological graph, we computed the embeddings of each character nodes (Grover and Leskovec, 2016), basing on which we devised a probability index to signify the compoundness of a QIE.

3.2.1 Morphological graph

The purpose of the morphological graph was to represent (1) the morphological relations between Chinese words and characters, and (2) the lexical and semantic relations between these words. The graph included only single- or two-character words, in the consideration that (1) Chinese words are predominantly bi-syllabic, words with one or two characters already account for 93.39% of word frequencies in a corpus; and (2) words longer than two characters potentially contaminated the graph with QIE compound information when modeling QIE compound structure.

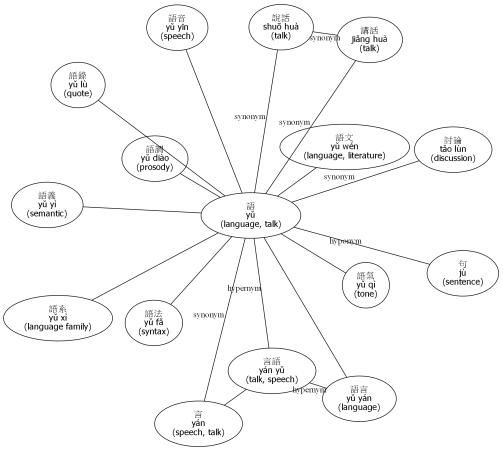


Figure 1: A sample morphological graph including 語 (yǔ , 'language, talk') and its immediate neighbors

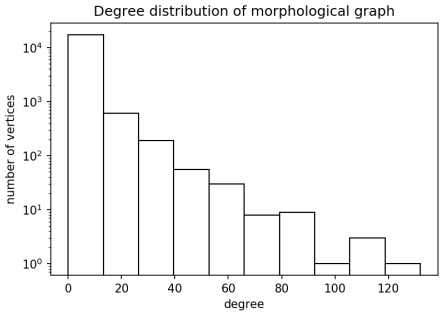


Figure 2: Degree distribution of the morphological graph.

The morphological graph was constructed from CWN. It had 18,251 vertices (including Chinese characters and two-character words) and 27,932 undirected edges (including morphological relations and CWN relations). Among the edges were 13,480 morphological relations, where characters were linked with their composed words. A sample graph was shown in 1. The degree distribution of the graph was shown in 2. There were 87% of vertices having 5 or fewer neighbors, while the most connected 20 vertices accounted for 70% connections in the graph.

Word morphology and semantic relations encoded in the graph allowed us to investigate the relations between characters, even ones not explicitly encoded in the graph. To efficiently explore the relations between characters in the graph, we computed a latent, low-dimensional node embeddings representation in

the graph. The index of compoundness was then defined basing on the embedding vectors.

3.2.2 Morphological vectors of Chinese characters

We used node2vec (Grover and Leskovec, 2016) to compute vector representations for each of the nodes in morphological graph. node2vec found a mapping $f : V \rightarrow \mathbb{R}^d$ from each vertices to a vector representation, and the mapping was optimized to maximize the log-probability of observing its neighbors in the graph given the vector. The mapping f was defined as:

$$\max_f \sum_{c \in \mathbb{C}} \log p(N(c)|f(c)) \tag{2}$$

where c is each of characters, $\mathbb{C}$, in the graph, and $N(c)$ denoted the neighbors of the character c in the graph. node2vec provided parameters to fine tune the random walk strategies when learning latent representations. In order to stress the homophily among characters, we chose $p = 2$ and $q = 0.5$ as random walk parameters. Since the probability of compounding would be evaluated on the character level, only embedding vectors of single characters were considered in following steps. We defined these vectors of characters as morphological vectors, $\mu_i = f(c_i)$, where subscript i denoted each character in the morphological graph.

Basing on morphological vectors μ_i, we first defined the compoundness of two characters as a conditional probability observing the second character given the first character. The conditional probability is based on the cosine similarity among morphological vectors, normalized to a categorical distribution with the softmax function, which could be formulated as follows:

$$p(c_2 \mid c_1) = \frac{\exp\left(\phi(\mu_1, \mu_2)\right)}{\sum_{i \in \mathbb{C}} \exp\left(\phi(\mu_1, \mu_i)\right)} \tag{3}$$

where $\phi(x, y)$ was cosine similarities between two vectors, and $\mathbb{C}$ denoted all characters in the morphological graph.

The compoundness of a QIE was defined through conditional probabilities. We assumed a linear dependency structure within QIE, that is, each character only dependent on its immediate predecessor. The compoundness of QIE then factored into a series of conditional probabilities between neighboring characters:

$$\begin{aligned} \text{compoundness} &= \log p(c_1, c_2, c_3, c_4) \\ &= \log p(\mu_1) p(\mu_2 \mid \mu_1) \\ &\quad\; p(\mu_3 \mid \mu_2)\, p(\mu_4 \mu_3) \end{aligned} \tag{4}$$

where μ_1, μ_2, μ_3, μ_4 denoted four morphological vectors of characters in the QIE. Higher probability signified stronger compoundness, i.e., the QIE followed a more common compound rules, such as the idiom 虎頭蛇尾 (compoundness $= -22.59$). Lower probabilities signified low compoundness, i.e. the QIE followed less common compound patterns, such as the idiom 來龍去脈 (compoundness $= -24.06$).

3.3 Compositeness

Semantic non-compositionality, or opaqueness, is the defining feature of idiomaticity. For example, 滄海桑田, cāng hǎi sāng tián, "drastic change of circumstances over time" is an opaque idiom. Each of its constituent characters: 滄, cāng, "blue", 海, hǎi, "ocean", 桑, sāng, "mulberry", 田, tián, "farm" bears no indication of the idiomatic meaning. As opposed to a more *transparent* idiom, 盡善盡美, jìn shàn jìn měi, "as perfect as possible" is more related to its constituents' meanings: 盡, jìn, "try to" 善, shàn, "good", 美, měi, "beauty".

In order to model compositeness, this paper took advantage of recent development of contextualized embeddings models, and the example sentences in CWN as a sense disambiguated lexical resources. We first constructed *sense vectors* from contextualized embeddings, and upon which we formalized idiomatic meaning and literal meaning of QIEs.

3.3.1 Idiomatic meaning of QIEs

Vector semantics received wide attentions in recent years, especially word embedding models such as word2vec (Mikolov et al., 2013). However, models of word semantics represented word meaning on lemma levels, which conflated different senses of a single word form (Camacho-Collados and Pilehvar, 2018). Recent advancement of contextualized embeddings, such as BERT model (Devlin et al., 2018) used a cloze task in training, allowing model to encode sentential contexts of the target word. Previ-

ous studies demonstrated that these contextualized embeddings, when combined with a set of disambiguated sense example sentences from CWN, resulted in sense vectors which can serve as a representation of CWN senses. These sense vectors, guided by linguistics constraints, help lexicographers find potential semantic relations in CWN (Tseng and Hsieh,).

Following the proposal of sense vectors, and the fact QIEs are predominately monosemic, we defined idiomatic meaning of a QIE as its contextualized embedding in the sentences. That is, a sense vector of QIE, σ_q, can be estimated by sampling sentences it occurred in, which can be formulated as an expectation over a set of sentences:

$$\sigma_q = \mathop{\mathbb{E}}_{\mathbf{w}\in\mathbb{W}} \left[\mathrm{CE}(\mathbf{w}) \cdot \mathrm{I_q}(\mathbf{w})\right] \tag{5}$$

where **w** was the list of words in a sentence, which was sampled from all the sentences the target QIE q occurred in, $\mathbb{W}$. $\mathrm{CE}(\mathbf{w})$ denoted the contextualized embeddings of the sentences, and $\mathrm{I_{target}}(\mathbf{w})$ was the indicator function to select out the embeddings of the target QIE.

In contrast of idiomatic meaning, the formalization of literal meanings was complicated by the fact most of the Chinese characters are polysemous (or homonymic). That is, to construct the sense vectors of a literal meaning, the character senses from which the literal meaning were composed should be first independently determined.

3.3.2 Literal meanings of QIEs

The task of determining character senses participated in QIE literal meanings, can be framed as finding the most probable sequence of sense composition. This view drew support from the semantic description view of Chinese word morphology, which argued meaning of the whole word came from the meaning of its constituent parts (Packard, 2000). That is, the compositionality of different senses should manifest itself on how surface word form compound to each other. In other words, the morphological vector space constructed in 3.2 could be regarded as an approximate estimate of sense composition space. The sense composition could be estimated by first projecting the sense vector into morphological vector space with projection matrix P:

$$\mathbf{P} = (\mathbf{S}^\intercal\mathbf{S})^{-1}\,\mathbf{S}^\intercal\mathbf{M} \tag{6}$$

where **M** was the morphological matrix with its rows being morphological vectors of each character, and **S** was the matrix with its rows being first sense vectors of each character in CWN (basing on the heuristic that the first sense of each character was the most frequently used sense). After obtaining the projection matrix P, we defined a function g mapping from sense vectors σ_{x_i} to an estimated morphological vector $\hat{\mu}_{x_i}$:

$$\hat{\mu}_{\mathrm{x_i}} = g(\sigma_{x_i}) = \mathbf{P}\cdot\sigma_{\mathrm{x_i}} \tag{7}$$

The joint probability of a given sense assignment can be computed based on the projected vector $\hat{\mu}_{\mathrm{x_i}}$, as defined in compoundness:

$$\begin{aligned} p(x_1,x_2,x_3,x_4) =& p(\hat{\mu}_{\mathrm{x_1}})p(\hat{\mu}_{\mathrm{x_2}}|\hat{\mu}_{\mathrm{x_1}}) \\ & p(\hat{\mu}_{\mathrm{x_3}}|\hat{\mu}_{\mathrm{x_2}})p(\hat{\mu}_{\mathrm{x_4}}|\hat{\mu}_{\mathrm{x_3}}) \end{aligned} \tag{8}$$

The most probable sense sequence in a given QIE q is then the sense assignment, $\mathbf{x}_q = (x_1, x_2, x_3, x_4)$ that maximize the joint probability:

$$\mathbf{x}_q = \mathop{\mathrm{argmax}}_{\mathrm{x}\in\mathrm{S}(q)} p(x_1,x_2,x_3,x_4) \tag{9}$$

where $\mathrm{S}(q)$ denotes all possible sense assignments in the given QIE q.

Basing on the probability, the most probable sense sequence $\mathbf{x}_q$ can then be decoded with beam search.

Equipped with the most probable sense sequence decoded in QIE, we defined index of compositeness as sum of (square root) distances between each of character sense vectors (literal meaning) and the QIE sense vector (idiomatic meaning). The index was calculated by:

$$\text{compositioness} = \sum_{x_i\in\mathbf{x}(q)} \frac{\sqrt{\|\sigma_{x_i} - \sigma_q\|^2}}{d} \tag{10}$$

where d was the dimension of sense vectors. Higher compositeness indicated literal meanings further away from idiomatic meaning, i.e., QIE was more opaque, such as the idiom 滄海桑田

(compositeness = 0.0997). Lower compositeness indicated literal meanings closer to idiomatic meaning, i.e. the QIE was more transparent, such as the idiom 盡善盡美 (compositeness = 0.0892).

4 Experiment

We presented two experiments, where three dimensions in model of idiomaticity were used as features to classify idioms and proper nouns from general QIEs.[1]

4.1 Idiom classifications

The purpose of this experiment was to illustrate the nature of QIEs, including prefabs and idioms. While Chinese idioms themselves were not a homogeneous class of linguist construct, prefabs, as a dynamic phenomena of language usage, should exhibit more variant behaviors with respect of variation, compoundness, and compositeness.

The experiment analyzed QIEs in a corpus of 1.2 billion characters, which included texts from news and online forum. In the corpus, we first extracted 319,201 quadgrams that occurred more than 32 times. Among these quadgrams, we selected 2,478 different prefabs that (1) were frequently occurred in the corpus, (2) has high PMI score (i.e. the four characters did not collocate by chance), and (3) did not frequently occurred in a fixed five-grams. Along with the prefabs, we referenced the idioms dictionary from Ministry of Education, Taiwan (MOE) to select a list of idioms as analyzing materials. Among 5,106 idioms included in the dictionary, we only included 1,518 idioms occurred more than 50 times.

Each of these 3,996 prefabs and idioms were computed for three features in model of idiomaticity. These three features were then used as classifying features in a gaussian-kernel SVM. The results classification was evaluated with a 5-fold cross validation, with mean accuracy of 70.80%, *SD* = 0.0146. Due to unequal number of prefabs and idioms, the random baseline was 62.01%.

Features distribution were shown in 3. In index of variation, the mean of idioms(*M* = -2.69, *SD* = 1.49) was lower than prefabs (*M* = -1.48, *SD* = 1.51), which was consistent to the observation that idioms were more conventionalized, therefore more resistant to usage variation. The compoundness distribution of prefabs (*M* = -23.25, *SD* = 0.37) had higher value than ones in idioms (*M* = -23.39, *SD* = 0.30), and exhibited fatter tail. It was consistent to the idea that idioms, comes from classic Chinese, followed a less common compound rule. However, compositeness distribution of prefabs and idioms showed greater overlap, and values of prefabs (*M* = 0.095, *SD* = 1.98e−3) were slightly higher than idioms (*M* = 0.094, *SD* = 2.02e−3).

One possible reason for higher compositeness (hence more opaque) of prefabs was some of which were proper nouns, such as names of locations or person. Since these proper nouns were often translated names, the characters meaning has no relations to the names they referring to, i.e., they are more opaque. Therefore, Proper nouns would serve as a clear materials to test the model of idiomaticity. Specifically, proper nouns should be opaque (high in composite index), and they would not follow morphological rules (hence low in compoundness index), and cannot allowed variations (low on variation index).

To test the hypothesis above, we conducted another experiment with proper nouns and other general prefabs.

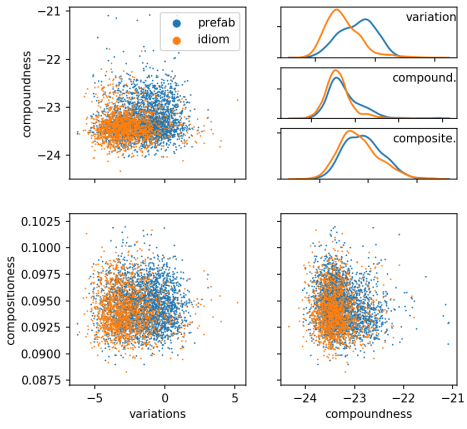


Figure 3: Distribution and scatter plots of three features in idioms and prefabs

[1] We intend to make code publicly available via github after the reviewing process.

4.2 Proper noun classification

This experiment was aimed to investigate the proper nouns with model of idiomaticity. The proper nouns were manually identified from the prefabs used in previous experiment. There were 108 proper nouns selected for this experiment, which were largely translated person names (e.g., 哈利波特, hā lì bō tè, "Harry Potter"), or locations (e.g., 巴基斯坦, bā jī sī tǎn, "Pakistan"). We randomly selected another 108 items (none of them were proper nouns) as general prefabs.

A proper noun classification task was performed and the results classification was also evaluated with a 5-fold cross validation. The mean accuracy was 71.29%, *SD* = 6.19%. The chance (baseline) level was 50.0%.

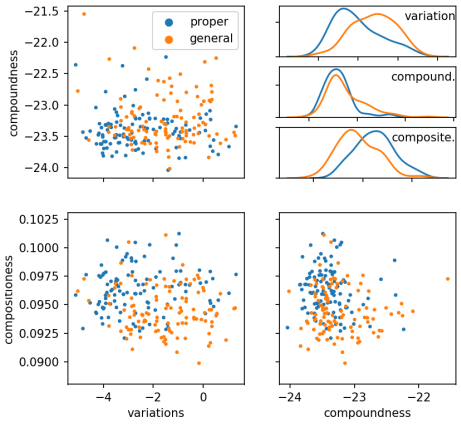


Figure 4: Distribution and scatter plots of three features in proper nouns and general prefabs.

4 showed the feature distribution of proper nouns and general prefabs. The overall patterns were consistent with the prior hypothesis. In index of variation, proper nouns (*M* = -2.54, *SD* = 0.14) were less likely to have variation forms, compared to general prefabs (*M* = -1.52, *SD* = 0.13). Proper nouns also had lower compoundness (*M* = -23.40, *SD* = 0.029) than general ones (*M* = -23.24, *SD* = 0.039). Compositeness showed a clear difference between proper nouns (*M* = 0.096, *SD* = 1.89e−4) and general prefabs (*M* = 0.095, *SD* = 1.95e−4).

The results of these two experiments demonstrated model of idiomaticity can be useful to shed light on properties, namely the variation, compoundness, and compositeness of Chinese QIEs.

4.3 Encoding QIEs in CWN

2478 QIEs-prefabs and 1518 idiom-QIEs are explored in this study. In considering the inclusion of wordnet , Idiom-QIEs are excluded, as they are well-studied in Chinese lexicography. What interests us more is the prefabs-QIEs and how we encode them into the organization of Chinese Wordnet.

We select top 200 prefabs-QIEs for manually clustering and determining their mapping to the current synsets with possible relations. Among these 200 QIEs, 109 QIEs could be justified as established concepts to incorporate into CWN (e.g., 移送法辦 'bring to justice'), 48 QIEs are more likely quasi-compounds with high frequency (e.g., 競選總部 'campaign headquarter'), and 43 QIEs are hard to be mapped into CWN as they carry mainly the pragmatic/discourse meaning (換句話說 'in other words').

5 Conclusion

In this paper, we demonstrate a proposed approach in modeling the idiomaticity of a special yet recurrent type of idiomatic expressions called QIE in Chinese. In contrast with English idioms, Chinese QIEs are different in that they are phonologically composed of four syllables, syntactically fixed structure, and semantically intransparent. Three dimensions are considered in modeling QIE's behaviour, and two classification experiments are conducted to test the model. In addition, the consequences of encoding QIEs in Chinese Wordnet is discussed.

References

Jose Camacho-Collados and Mohammad Taher Pilehvar. 2018. From word to sense embeddings: A survey on vector representations of meaning. *Journal of Artificial Intelligence Research*, 63:743–788, December.

Wallace Chafe. 1968. Idiomaticity as an anomaly in the chomskyan paradigm. *Foundations of Language*, 4:109–127.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.

C. Fernando. 1996. *Idioms and Idiomaticity*. Describing English language. Oxford University Press.

Bruce Fraser. 1970. Idioms within a transformational grammar. *Foundations of Language*, 6(1):22–42.

Kristina Geeraert, John Newman, and R Harald Baayen. 2017. Idiom variation: Experimental data and a blueprint of a computational model. *Topics in cognitive science*, 9(3):653–669.

Aditya Grover and Jure Leskovec. 2016. Node2vec: Scalable feature learning for networks. In *Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, pages 855–864, New York, NY, USA. ACM.

Alice F. Healy. 1994. Letter detection: A window to unitization and other cognitive processes in reading text. *Psychonomic Bulletin & Review*, 1(3):333–344, Sep.

Shu-Kai Hsieh, Chiung-Yu Chiang, Yu-Hsiang Tseng, Bo-Ya Wang, Tai-Li Chou, and Chia-Lin Lee. 2017. Entrenchment and creativity in chinese quadrasyllabic idiomatic expressions. In *Workshop on Chinese Lexical Semantics*, pages 576–585. Springer.

A. Langlotz. 2006. *Idiomatic Creativity: A Cognitive-linguistic Model of Idiom-representation and Idiom-variation in English*. Human cognitive processing. J. Benjamins.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc.

J.L. Packard. 2000. *The Morphology of Chinese: A Linguistic and Cognitive Approach*. Cambridge University Press.

Ivan A. Sag, Baldwin Timothy, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for nlp. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics*, pages 1–15.

Y. H. Tseng and S. K. Hsieh. Augmenting chinese wordnet semantic relations with contextualized embeddings. submitted.

Benjamin K Tsou. 2012. Idiomaticity and classical traditions in some east asian languages. In *Proceedings of the 26th Pacific Asia Conference on Language, Information, and Computation*, pages 39–55.

V-*gei* Double Object Construction and Extra Argument in Mandarin

Yu-Yin Hsu
Hong Kong Polytechnic University
Hung Hom, Hong Kong
yyhsu@polyu.edu.hk

Teng Qu
Hong Kong Polytechnic University
Hung Hom, Hong Kong
teng0319.qu@connect.polyu.hk

Abstract

This paper examines the syntax of a morphologically complex double object construction in Mandarin, V-*gei* structure, and uses the results as the basis for a new account of a special phenomenon: sentences with an extra experiencer. Following Pylkkänen's (2002) work on applicative phrases, we argue that different interpretations of the indirection object in double object construction can be accounted for by the differences between high and low applicatives. We adopt Paul and Whitman's (2010) raising applicative hypothesis to account for double object construction, and argue that the indirect object moves to the specifier of low applicative projection to be licensed with the goal reading. Further, we argue that this indirect object may optionally raise to the high applicative phrase to obtain the benefactive thematic role. This helps to explain the phenomenon of indirect objects not always carrying a benefactive reading. We then propose that an argument may directly merge with the high applicative head as its specifier, resulting in sentences with an (unexpected) extra argument expressing either a benefactive or a malfactive reading. Lastly, the structural properties of both the high applicative projection and the low applicative projection will be discussed in relation to passivization and the causative *ba* construction in Mandarin.

1 Introduction

This paper proposes to extend the essential components of Pylkkänen's (2002) high and low applicative analysis to two phenomena in Mandarin: the V-*gei* 'V-*give*' double object construction, and a special type of sentence that carries extra arguments.

V-*gei* sentences (e.g., (1a)) are interesting because Mandarin has a set of bare ditransitive verbs that do not include the morpheme *gei* (e.g., *song* in (1b) vs. *xie-gei* in (1a)), but which nonetheless sometimes occur with the morpheme *gei* (e.g., (1c)). It is therefore worth asking what the semantic and structural functions of *gei* in sentences like (1) are.

(1) a. Tony xiě-gěi-le Mălì yī-fēng-xìn.
Tony write-give-Asp Mali 1-CL-letter
'Tony wrote Mali a letter.'
b. Wǒ sòng-le Mălì yī-gè shǒubiǎo .
1SG send-ASP Mali one-CL watch
'I sent Mali a watch.'
c. Wǒ song-gěi-le Mălì yī-gè shǒubiǎo.
1SG send-give-ASP Mali one-CL watch
'I sent Mali a watch.'

Just like their English counterparts, typical Mandarin verbs take two arguments if transitive (like *he* 'drink' in (2a)) and one if intransitive (like *ku* 'cry' in (2b)).

(2) a. Zhāngsān hē-le sān-píng-jiǔ.
Zhangsan drink-Asp three-bottle-wine.
'Zhangsan drank three bottles of wine.'
b.Mălì kū-de xīn fán.
Mali cry-De upset
'Mali cried and felt upset.'

However, sometimes we see an extra argument in such sentences, e.g., the word *Lǐsì* of examples (3a) and (3b).

(3) a. Zhāngsān hē-le Lǐsì sān-píng-jiǔ.
Zhangsan drink-Asp Lisi three-bottle-wine.
'Zhangsan drank Lisi's three bottles of wine.'
b. Mălì kū-de Lǐsì xīnfán.
Mali cry-De Lisi upset
'Mali's crying made Lisi upset.'

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 76-84, Hakodate, Japan, September 13-15, 2019

In this paper, we will argue that the differences between sentences like (2) and (3) can be accounted for by a modified version of Pylkkänen's (2002) high applicative analysis; and that V-*gei* double object construction can be explained by extending Paul and Whitman's (2010) modification of Pylkkänen's low applicative projections to the Mandarin context. Specifically, in the spirit of Larson's (1988) VP shell hypothesis, and following Paul and Whitman's raising applicative analysis, we assume that *gei* in the V-*gei* construction is the head of a low applicative projection ($Appl_LP$). However, we depart from Pylkkänen's original proposal by arguing, like Paul and Whitman, that such an $Appl_L$ selects a VP and attracts the indirect object (IO) to its specifier. The lexical verb then undergoes head-movement to gei_{ApplL} to yield the V-*gei* complex.

We will then extend Pylkkänen's (2002) high applicative analysis to account for Mandarin's additional benefactive reading of IO in ditransitive constructions, and for the non-canonical extra arguments like those in (3). We will show how combining this high applicative projection with Paul and Whitman's (2010) raising applicative structure can account for a wider range of Chinese data than either of them by itself. Our proposal will be unlike Kuo's (2016) insofar as it eliminates empty movements and extra functional projections.

This paper is organized as follows. In section 2, we discuss the two competing accounts of low applicative phrases in Chinese, i.e., Paul and Whitman's (2010) raising applicative hypothesis, and Kuo's (2016) light applicative projection. In section 3, we present our proposal regarding V-*gei* double object construction and how it can explain sentences with an extra argument. We provide empirical support for the predictive value of the current proposal in section 4, and then briefly sum up our findings and their implications in section 5.

2 The V-*gei* Construction

2.1 Applicatives in Chinese Double Object Construction

A few studies have recently discussed the application of applicative projections in Mandarin for some syntactic phenomena. A high applicative phrase ($Appl_HP$) introduces a (benefactive) argument above the VP (4): e.g., the Luganda example 'Katonga' in (5) (Pylkkänen 2002: 25).

(4) High Applicative
$[_{vP}\ [_{ApplHP}\ NP_{BENEFACITVE}\ [_{ApplH'}\ Appl_H\ [_{VP}\ V\ NP]]]]$

(5) Mukasa ya-tambu-le-dde Katonga.
Mukasa PAST-walk-APPL-PAST Katonga
'Mukasa walked for Katonga.'

Unlike $Appl_HP$, a low applicative phrase ($Appl_LP$) merges under a VP (6) (Pylkkänen 2002: 24) and introduces a source/recipient argument (e.g., *him* in (7)), such that the event encoded by the VP denotes a transfer of possession.

(6) Low Applicative
$[_{VP}\ V\ [_{ApplLP}\ NP_{SOURCE/RECIPIENT}\ [_{ApplL'}\ Appl_L\ NP]]]$

(7) I baked him a cake.

While (6) can account for the thematic relations in sentences with typical ditransitive verbs, Paul and Whitman (2010) point out that such a structure cannot be directly applied to Chinese V-*gei* double object construction, because the potential head-raising of $Appl_L$ to the verb would produce an ungrammatical **gei*-V complex, e.g., **gei-song* 'give-send' in (8). Therefore, they propose a raising applicative analysis, as in (9), where an applicative projection dominates a VP with a double object, and the goal argument raises to Spec,ApplP.

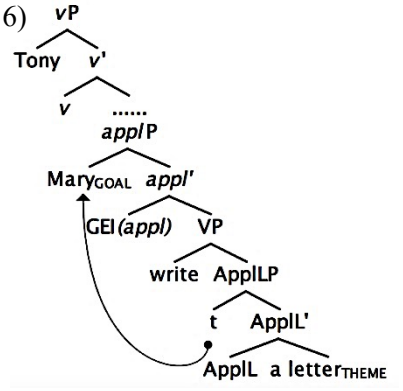
(8) * $[_{VP}$ *gei-song* $[_{ApplLP}\ NP\ [_{ApplL'}\ \sout{gei}_{ApplL}\ NP]]]$

(9) Raising Applicative
$[_{APPLP}\ DP_{Goal}\ [_{APPL'}\ Appl\ [_{VP}\ \sout{DP_{Goal}}\ [_{V'}\ V\ DP_{Theme}]]]]$

Much like Paul and Whitman (2010), Kuo (2016) argues that the IO in sentences like (1) raises from the VP to a higher position above it. Following Citko (2011), Kuo further proposes a light applicative projection (applP) associated with low applicatives in Mandarin. As shown in (10), the low applicative accounts for the basic IO in the double object construction; but to derive the V-*gei* complex, and interpretation of *Mali* as being the recipient and/or the benefactive, both verb and IO move. However, Kuo's (2016:60) analysis requires an extended light applicative projection of the low

applicative projection, and that this extended light projection be separated by another lexical head verb which is not directly related to applP.

(10) Kuo (2016)



Both the above accounts point out that the notion of a low applicative in V-*gei* sentences is supported by such sentences passing Pylkkänen's (2002) diagnostics for the low applicative projection. For example, since low applicatives express transfer of possession within an event, they are incompatible both with stative verbs (which by their nature cannot describe a transferring event) and intransitive verbs; and Mandarin's V-*gei* construction exhibits the same type of incompatibility (e.g., (11) and (12)), suggesting that a low applicative structure is involved in V-*gei* sentences. Importantly, as noted by Pylkkänen (2002), the high applicative – unlike the low one – are compatible with these types of verbs: a point we will return to in section 3.

(11) Intransitive verb
*Mǎlì kū-gěi Lǐsì.
Mali cry-give Lisi
'Lisi was upset by Mali's crying.'

(12) Stative predicate
*Zhāngsān ná-gěi-zhe Mǎlì bāo.
Zhangsan hold-give-Asp Mali bag
'Zhangsan held the bag for Mali.'

Additionally, Kuo (2016) points out that not all double object constructions require the morpheme *gei*, and thus, some ditransitive sentences show variations, e.g., (13).

(13) Tony sòng(-gěi)-le Mǎlì yī-gè-shǒu.biǎo.
Tony send(-give)-Asp Mali 1-CL-watch
'Tony gave Mali a watch.'

We agree with Kuo that sentences like (13) do not simply contain an optional *gei* 'give', and we will argue that two different structures are involved: that is, a simple ditransitive structure (with *song* 'send/give' alone) or a V-*gei* complex verb, especially if we observe the fact that some *gei* cannot be omitted from certain V-*gei* sentences, e.g., (13) vs. (14).

(14) a. Tony mǎi(-gěi)-le Mǎlì yī-gè-shǒu.biǎo.
Tony buy-give-Asp Mali 1-CL-watch
'Tony bought a watch to/for Mali.'
b.*Tony mǎi-le Mǎlì yī-gè-shǒu.biǎo.
Tony buy-Asp Mali 1-CL-watch
'Tony bought a watch to Mali.'

Moreover, we suggest that, even without the functional light applicative projection proposed by Kuo (2016), bare ditransitive sentences like those with *song* 'give/send' in (13) and canonical transitive sentences like (15) can still be derived.

(15) Tony mǎi-le yī-gè-shǒu.biǎo.
Tony buy-Asp 1-CL-watch
'Tony bought a watch.'

That is, following Paul and Whitman's (2010) analysis that *gei* is the head of a low applicative projection that selects a VP as its complement to express transfer of possession, we propose that sentences like (15), with a bare canonical transitive VP, do not have a low applicative projection, and cannot express a goal/recipient IO (e.g., (14b)).

2.2 IO-raising in V-*gei* Construction

Both the raising applicative and light applicative structures require an IO-raising mechanism. Kuo (2010) argues that Paul and Whitman (2010) failed to prove that the IO must be moved from the VP, and therefore that an alternative is needed.

According to Paul and Whitman, distributive quantifiers such like *meiren* 'each' occur to the right of the IO, as shown in (16).

(16) Lǐsì sòng-gěi háizi-men měirén
Lisi send-give children-PL each
yī-bǎi kuài.
100-CL money
'Lisi gave the children each 100 dollars.'

They argue that *meiren* adjoins to the VP, and the IO raises from Spec,VP to Spec,ApplP. There are three reasons for this derivation: first, that the order of the distributive quantifiers and the frequency adverb is fixed; so if *meiren* were inside the VP, (17b) would be acceptable.

(17) a. lǎoshī sòng-gěi háizi-men
teacher send-give children-PL
měirén sān-cì lǐwù.
each 3-times gift
'The teacher gave every child a gift three times.'
b. *lǎoshī sòng-gěi háizi-men
teacher send-give children-PL
sān-cì měirén lǐwù.
3-times each gift
'The teacher gave every child a gift three times.'

Second, *meiren* cannot form a constituent with a noun phrase (NP), so the sentences in (18), which have *meiren*-NP as the IO and direct object (DO), are both ungrammatical.

(18) a. *lǎoshī sòng-gěi [měirén háizi-men]
Teacher send-give each children-PL
yī-jiàn lǐwù
1-CL gift
'The teacher gave the children each a gift.'
b. *lǎoshī mà-le [háizi-men měisrén].
teacher scold-Asp children-PL each
'*The teacher scolded the children each.'

Third, when a different distributive quantifier, *yiren*, is added to the NP, the quantifier and the NP still do not form a constituent.

(19) *xiàozhǎng fen-gěi [yīrén lǎoshī]
principal allot-give each teacher
shí-gè xuéshēng
10-CL students
'The principal allotted ten students to each teacher.'

Instead, Kuo (2016) argues that there could be different ways to explain ungrammatical sentences like (18a): namely, that distributive quantifiers may not occur in a pre-nominal position but must be post-nominal, as shown in (20).

(20) a. háizi-men měirén mǎi-le
children-PL each buy-Asp
yī-běn shū
1-CL book
'The children each bought a book.'
b.*měirén háizi-men mǎi-le yī-běn
each children-PL buy-Asp 1-CL
shū.
book
'The children each bought a book.'

However, this view may not be tenable. That is, if quantifiers like *meiren* only occur after the NP, sentences like (18b) should be acceptable, contrary to the facts.

3 Our Proposal

3.1 New High and Low Applicative Analyses

Following Paul and Whitman's (2010) raising-applicative analysis, we propose that when the head of a low applicative is not overt, it produces sentences like (21).

(21) Tony sòng-le Kaite yī-jiàn lǐwù.
Tony send-Asp Kaite 1-CL gift
'Tony gave a gift to Kaite.'

We also propose that when *gei* 'give' (the head of the low applicative) is overt, it yields sentences like (22).

(22) Tony sòng-gěi-le Kaite yī-jiàn lǐwù.
Tony send-give-Asp Kaite 1-CL gift
'Tony gave a gift to Kaite.'

IO in English double object construction can ambiguously have either a pure goal reading or an extra benefactive reading (23), and its Chinese counterparts (21-22) exhibit the same type of ambiguity: that is, the IO (e.g., *Kaite* in both (21) and (22)) can be a benefactive or just a goal.

(23) Tony baked Kaite a cake.
a. Tony bake a cake to Kaite (as goal).
b. Tony bake a cake for Kaite (as benefactive).

We propose that the low applicative introduces the goal/recipient argument in Chinese double object construction, denoting transfer of possession. However, when an IO carries a

benefactive reading, it is because that IO has raised to a high applicative projection inside of *v*P, as shown in the structure in (24). In other words, sentences whose IOs have goal readings involve a low applicative, while IOs with benefactive readings involve low-to-high applicative raising.

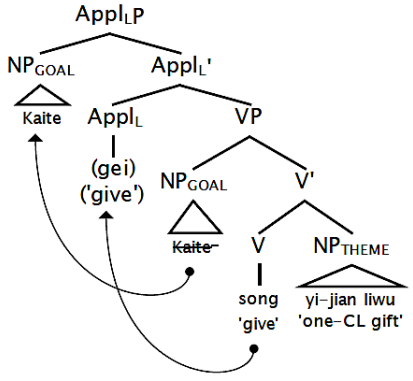
(24) Proposal: New High Applicative Analysis
[$_{vP}$*v* [$_{ApplHP}$ NP Appl$_H$ [$_{ApplLP}$ ~~NP~~ [$_{ApplL'}$ Appl$_L$ VP]]]]

This proposal can also explain the availability of passivization in double object construction. Looking again at the Chinese examples (21) and (22), only sentences like the former allow passivization of IO *Kaite* (25), while the V-*gei* sentences do not, e.g., (26).

(25) Kaite bèi Tony sòng-le yī-jiàn lǐwù.
Kaite BEI Tony send-Asp 1-CL gift
'Kaite was sent a gift by Tony.'
(26) *Kaite bèi Tony sòng-gěi-le yī-jiàn lǐwù.
Kaite BEI Tony send-give-Asp 1-CL gift
'Kaite was sent a gift by Tony.'

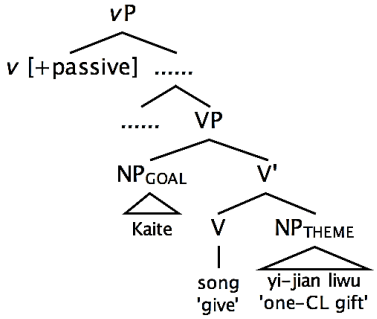
Given that passivization suppresses one internal argument inside a VP, our analysis predicts that sentences like (22) with V-*gei* structures in which the IO has already undergone raising to the spec,Appl$_H$P, outside of its base VP (e.g., (27)), the VP's internal lower copy cannot participate in the other syntactic operation.

(27) argument raising to Appl$_L$P

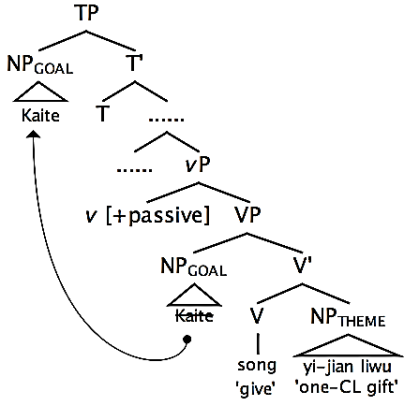


Supposedly, if the IO always moves to Appl$_H$P, sentences like (25) should be ungrammatical. We accept that it is indeed the case, and that sentences like (25) must be grammatical for some other reason. We propose that, while the IO *Kaite* still remains in the VP of the ditransitive verb *song* 'give' (28a), when a passive *v* is merged, the IO inside the VP is not moved out (unlike in (27)), and this IO can therefore still be passivized later (e.g., (28b)), yielding sentences like (25).

(28) a. No A-movement type in ditransitive VP



b. Passivization of ditransitive VP of (28a)



3.2 Application of New High Applicatives

Our proposed structure (24) can account for double object construction and sentences with extra arguments in a uniform way.

It has been noticed that sometimes, transitive and intransitive sentences have an extra argument. The examples of this phenomenon from (3) are repeated below.

(3) a. Zhāngsān hē-le Lǐsì sān-píng-jiǔ.
Zhangsan drink-Asp Lisi three-bottle-wine.
'Zhangsan drank Lisi's three bottles of wine.'
b. Mǎlì kū-de Lǐsì xīn-fán.
Mali cry-DE Lisi upset
'Mali's crying made Lisi upset.'

Similar phenomena have been discussed in Liu and Shi (2018). They proposed that such extra arguments directly merged with Appl project to express either a benefactive or malfactive role.

They claim that since such extra arguments are not subcategorized by the main verb, this type of arguments cannot form a causative *ba* sentence; Liu and Shi (2018) do not discuss the availability of passivization for this kind of sentences.

As we will show immediately below, we think the applicative analysis provided in Liu and Shi (2018) is tenable, but some of their interpretations of examples and their explanations to the applicative structure and other associated constructions seem to be oversimplified. We will present data showing that, contrary to the claim in Liu and Shi (2018), NPs that directly merge with high applicative projection (see (24)) and receive benefactive or malfactive role can participate in the later syntactic derivation such as passivization or to form a causative *ba* construction.

For examples like those in (3), we suggest that the extra argument *Lisi* in both sentences in (3) directly merges with the high applicative projection, and takes a specific thematic role: either benefactive or malfactive. As an extra argument introduced by the high applicative, this argument is indirectly related to the event described by the verb. Also, because it is not introduced by low applicative, this argument is not relevant to the concept of transfer of possession.

Accordingly, we predict that if an argument does not undergo a prior movement inside *v*P, it may undergo passivization (cf. (28)). This prediction is borne out in sentences with extra arguments. For instance, the extra argument *Lisi* in (3) that we proposed to be directly merged with the high applicative can be passivized, as in (29).

(29) a. Lǐsì bèi Zhāngsān hē-le
Lisi BEI Zhangsan drink-Asp
sān-píng-jiǔ.
three-bottle-wine
'He drank three bottles of wine on me.'
b. Lǐsì bèi Mǎlì kū-de xīn-fán.
Lisi BEI Mali cry-DE upset
'I was upset by Mali's crying.'

We also note that the same argument can occur in the *ba* construction, in a preverbal position.

(30) a.Zhāngsān bǎ Lǐsì hē-le
Zhangsan BA Lisi drink-Asp
sān-píng-jiǔ.
three-bottle-wine
'Zhangsan drank three bottles of wine on Lisi. (He still complained that it's not enough.)'
b. Mǎlì bǎ Lǐsì kū-de xīn-fán.
Mali BA Lisi cry-De upset
'Mali's crying made Lisi upset.'

Interestingly, sentences like (1) with V(-*gei*) – which we previously said involved raising of a goal argument from Spec,VP to Spec,Appl$_L$P – do not allow the IO to occur in a *ba* construction.

(31) a.*Tony ba Mǎlì xiě-gěi-le yī-fēng-xìn.
Tony BA Mali write-give-Asp 1-CL-letter
'Mali got to be given a letter by Tony.'
b.* Tony ba Kaite sòng-le yī-gè lǐwù.
Tony BA Kate send-ASP one-CL gift
'Kate got to be sent a gift from Tony.'

The contrast between (30) and (31) reflects the structural differences between high (30) and low (31) applicatives. We argue that the grammatical differences between (30) and (31) can be explained derivationally: i.e., if an argument has already gone through movement, its lower copy at the original site cannot participate in other later derivation in the same phase domain (e.g., *v*P in this case).

Nonetheless, some apparent counter-examples have caught our attention. The sentences in (32) have transitive verbs (*da* 'hit' and *bo* 'peel') yet seem to take two internal arguments, *and* to allow the seeming IOs – i.e., *Lisi* in the (a) sentences and *juzi* 'orange' in the (b) sentences – to occur in the causative *ba* construction, as in (33).

(32) a. Zhāngsān dǎ-le Lǐsì yī-gè ěrguāng.
Zhangsan hit-ASP Lisi one-CL slap
'Zhangsan gave Lisi a slap on his face.'
b. Zhāngsān bō-le júzi pí.
Zhangsan peel-ASP orange skin
'Zhangsan peeled an orange.'
(33) a. Zhāngsān bǎ Lǐsì dǎ-le yī-gè ěrguāng.
Zhangsan BA Lisi hit-ASP one-CL slap
b. Zhāngsān bǎ júzi bō-le pí.
Zhangsan BA orange peel-ASP skin

Given what we propose, Appl$_L$P accounts for the goal IO, which assumes an operation involving raising of a canonical goal IO to Spec,Appl$_L$P; and this derivation should block later raising. So, how does syntax derive grammatical sentences like (33)?

We note that the types of internal arguments in sentences like (32-33) express inalienable possession, rather than a goal-theme relation. We thus argue that the seeming IOs *Lisi* and *juzi* are possessors of the DO, located inside of the nominal domain of DO (Hsu 2008, 2019), and were not introduced by $Appl_L$. Therefore, no raising to Spec,$Appl_L$P is involved, and such a possessor can further undergo raising from a nominal internal position to form a causative *ba* construction, resulting in sentences like (33).

We are grateful to a reviewer of an earlier version of this paper for pointing out that, in addition to inalienable possession, IOs expressing alienable possession (e.g., *Lisi* and *Zhangsan* in (34)) can also form causative *ba* sentences (e.g., (35)).

(34) a. Zhèngfǔ chāi-le
government pull.down-Asp
Lǐsì yī-tào fángzǐ.
Lisi 1-CL house
'The government pulled down a house of Lisi.'
b. Lǐsi zhāi-diào-le Zhāngsān màozǐ.
Lisi take off-Asp Zhangsan hat
'Lisi took off Zhangsan's hat.'

(35) a. Zhèngfǔ bǎ Lǐsì chāi-le
government BA Lisi pull.down-Asp
yī-tào fángzǐ.
1-CL house
'The government pulled down a house of Lisi.'
b. Lǐsì bǎ Zhāngsān zhāi-diào-le màozǐ.
Lisi BA Zhangsan take-off-Asp hat
'Lisi took off Zhangsan's hat.'

The same account can be applied to objects involved with kinship terms, such as in (36). If one accepts our proposal that a possessor can merged directly with high applicative from its nominal-internal position, and that from the high applicative position, it can undergo passivization (recall (29)). This prediction is borne out, as shown in (37).

(36) a. Zhāngsān dǎ-le Lǐsì érzǐ.
Zhangsan hit-Asp Lisi son
'Zhangsan hit Lisi's son.'
b. Tǔfěi dǎ-sǐ-le Lǐsì bàbà.
bandit beat-death-Asp Lisi father
'The bandits beat Lisi's father to death.'

(37) a. Lǐsì bèi Zhāngsān dǎ-le érzǐ.
Lisi BEI Zhangsan hit-Asp son
'Zhangsan hit Lisi's son.'
b. Lǐsì bèi tǔfěi dǎ-sǐ-le bàbà.
Lisi BEI bandit beat-death-Asp father
'The bandits beat Lisi's father to death.'

Nonetheless, we note some semantic restriction of the verb *da* 'hit' when it comes to forming the causative *ba* construction. The causative *ba* versions of the sentences in (36) do not receive the same level of acceptance as the originals, as shown in (38).

(38) a. *Zhāngsān bǎ Lǐsì dǎ-le érzǐ.
Zhangsan BA Lisi hit-Asp son
'Zhangsan hit Lisi's son.'
b. ?Tǔfěi bǎ Lǐsì dǎ-sǐ-le bàbà.
bandit BA Lisi beat-death-Asp father
'The bandits beat Lisi's father to death.'

We speculate that the difference between (38a) and (38b) is not due to derivational restriction, but rather to the semantics of the action verb *da* 'hit' which does not encode a result-state as required by the causative construction. Therefore, simply using the verb *da* 'hit' cannot form an acceptable causative sentence (e.g., (38a)); but the inclusion of a clear consequence to the hitting event, such as in (32a) and (38b), makes such *ba* sentences acceptable.

Before we move on, we would like to comment on some sentences' ambiguous readings. If one considers that *nawei shifu* 'that master' in (39) is the possessor of *bushao juezhao* 'many tricks', and predicts that *nawei shifu* should be able to form a causative *ba* construction, that prediction is not borne out, as shown in (40).

(39) Zhāngsān xué-le nàwèi shīfù
Zhangsan learn-Asp the master
bú shǎo jué zhāo.
many tricks
'Zhangsan learned many tricks from the master.'

(40) *Zhāngsān bǎ nàwèi shīfù xué-le
Zhangsan BA the master learn-A
bú shǎo jué zhāo.
not few tricks
'Zhangsan learned many tricks from the master.' (Liu and Shi 2018)

We suggest, however, that this contradiction is only apparent. Due to the main verb *xue* 'learn' in (39), the interpretation of *nawei shifu* and *bushao juezhao* in (39) is not simply a possessive relation, but a source-theme relation in terms of the learning event. That is, the structure of (39), unlike the possessive sentences we discussed previously, should be seen as parallel to the ditransitive construction, in which the IO *nawei shifu* 'that master' raised to the low applicative projection as the source, and cannot further raise to form a causative *ba* sentence (recall (21) and (31b)).

4 Some Extension

In light of our analysis that an argument can be directly merged as the specifier of $Appl_HP$ to obtain an indirectly associated thematic role (either benefactive or malfactive), we predict that this syntactic derivation should be compatible not only with intransitive and transitive verbs (e.g., (3)), but also with typical double object VPs (e.g., (41)). That is, the current proposal predicts that the specifiers of $Appl_HP$ and $Appl_LP$ can be occupied by different NPs. Though such sentences may require specific contexts to be uttered, the prediction is borne out.

(41) Zhāngsān sòng-le Lǐsì yī-jiàn lǐwù.
Zhangsan send-Asp Lisi 1-CL gift
'Zhangsan gave a gift to Lisi.'

Let us consider a scenario in which Mali promises to do Zhangsan a favor on the condition that the he gives a gift (possibly a bribe) to Lisi. Zhangsan then does give Lisi a gift, but Mali does not help Zhangsan as promised. One could comment on this situation with a sentence like:

(42) Zhāngsān bái-gěi Mǎlì sòng-le
Zhangsan in.vain-give Mali send-Asp
Lǐsì yī-jiàn lǐwù.
Lisi 1-CL gift
'Zhangsan gave a gift to Lisi for Mali but got nothing in return.'

In (42), Mali plays the benefactive role in this gift-giving event. The *gei* is the high applicative head to introduce *Mali* as a benefactive to be associated with the event described.

Since this argument is associated with the predicate through its direct merge at the high applicative, rather than from inside the low applicative or the VP, we predict that it can be passivized; and this prediction is borne out.

(43) ?Mǎlì bèi Zhāngsān bái-gěi
Mali BEI Zhangsan bai-gei
sòng-le Lǐsì yī-jiàn lǐwù.
give-Asp Lisi 1-CL gift
'Mali got benefit from Zhangsan's giving a gift to Lisi.'

5 Concluding Remarks

In this study, we examined the V-*gei* double object construction in Mandarin under the applicative framework. Following the insights of Paul and Whitman (2010) and Kuo (2016), we proposed a revised implementation of Pylkkänen's (2002) high and low applicatives. We went on to demonstrate that our proposal can account for the V-*gei* phenomenon and its associated structures (e.g., passivization and causative *ba* construction) in a simpler way, i.e., without relying on extra functional projections and empty movement of either the verb or the argument proposed in Kuo (2016).

We also tested how our proposal could account for some interesting sentence variances in Mandarin, and showed that an extra argument – either benefactive or malfactive – can be introduced by the high applicative on top of various types of verbal structures, including intransitive, transitive, and ditransitive ones. The restrictions of deriving the causative *ba* construction and passivization were also discussed with respect to sentences involving goal/source-theme relations and possessive relations, as well as sentences with extra arguments.

Our next step will be to extend our survey to additional phenomena involving various types of dislocation, to further test the validity of the current proposal and its explanatory power. However, due to limitations of space, we will do so via separate papers in the future.

Acknowledgments

This research was supported by the conference grant funded by the Department of Chinese and

Bilingual Studies at the Hong Kong Polytechnic University. We would like to thank the three PACLIC-33 anonymous reviewers for their insightful comments. Mistakes remaining are exclusively our own.

References

Citko, Barbara. 2011. *Symmetry in Syntax: Merge, Move, and Labels*. New York: Cambridge University Press.

Hsu, Yu-Yin. (2008) Possessor Extraction in Mandarin Chinese. *University of Pennsylvania Working Papers in Linguistics* 15:1 , Article 12.

Hsu, Yu-Yin. (2019) Possessor, appositive, and Beyond: A study of nominal-initial pronouns and proper names in Chinese. In: George Fowler, James Lavine, and Ronald F. Feldstein (eds.) *Festschrift for Steven Franks.* Bloomington, IN: Slavica Publishers.

Kuo, Pei Jung. (2016). Applicative and the double object construction in Mandarin Chinese. *Taiwan Journal of Linguistics*, 14(2), 33-76.

Larson, Richard. (1988). On the double object construction. *Linguistic Inquiry* 19, 335-391.

Liu Na, and Shi Dingxu. (2018) The syntactic properties of outer objects and the transitivity of V-O complex. *Language Teaching and Linguistic Studies.* 2018 Vol. 2, No. 190:32-43.

Pylkkänen Liina. (2002). *Introducing Arguments*. Cambridge: MIT dissertation.

Paul, Waltraud and John Whitman. (2010). Applicative Structure and Mandarin ditransitives. In: Maia Duguine, Susana Huidobro and Nerea Madariaga (eds.) *Argument Structure and Syntactic Relations. A cross-linguistic Perspective*. Amesterdam:John Benjamins, pp. 261-282.

Re-examining Syntactic, Semantic and Pragmatic Properties of Long-Distance Bound *Caki-casin* in Korean: An Experimental Study

Ji-Hye Kim
Department of English Education
Korea National University of Education
Cheongju, Korea
jkim@knue.ac.kr

Yong-hun Lee
Department of English Language & Literature
Chungnam National University
Daejeon, Korea
yleeuiuc@hanmail.net

Abstract

The present study investigated the syntactic, semantic and pragmatic properties of Korean Long-Distance Binding of anaphor *caki-casin*. Based on some previous experimental studies on this matter, we attempted to re-examine the properties of LD-bound local anaphor *caki-casin* as understood by Korean native speakers. A type of replication and modifications from some previous studies were made for the current study in terms of experimental design and statistical analyses using the data of the responses from 43 Korean native speakers. The results mostly reconfirmed the findings of previous studies, showing that Korean local anaphor *caki-casin* can be LD-bound with relevant syntactic/semantic/pragmatic properties. Detailed discussions will follow.

1 Introduction: Exempt Anaphor and Long-Distance Binding in Korean

In Standard Binding Theory (Chomsky 1981, 1986), the Binding Domain (BD) where reflexives should find their antecedents was defined as conjunction of Tensed S Condition (TSC) and Specified Subject Condition (SSC). As shown in (1) below, binding outside a finite clause (cf. 1b) as well as binding across a specified Subject (cf. 1c) yield ungrammaticality, compared to (1a).

(1)
a. John$_i$ blamed himself$_i$.
b.*John$_i$ said [that himself$_i$ was to blame].
c. *John$_i$ saw [Bill$_j$'s article about himself$_i$].
d. John$_i$ said [that the article about himself$_i$ was published in Times].

However, Pollard & Sag (1992) argued that TSC did not necessarily define the Binding Domain (BD) in English, while SSC still did. According to Pollard & Sag (1992), (1b) is ungrammatical because nominative anaphors are not allowed in English, since TSC-violation in (1d) above is still acceptable. While Chomsky (1986) tried to allow TSC violation like (1d) with the notion of 'Accessible Subject', Pollard and Sag (1992) explained the case by using the term 'exempt anaphors' – a term that is distinguished from core anaphors (i.e., grammatical anaphors that are constrained by syntactic properties). They claimed that TSC-violating reflexive in (1d) is exempt from syntactic Binding Theory, and is licensed extra-grammatically.

The properties of exempt anaphors introduced by Pollard & Sag (1992) are the following: i) they can be bound Long-Distance (LD) outside the local domain or be discourse-bound (cf. 2a, b); ii) they do not need c-commanding antecedents (cf. 2c).

(2)
a. Bill$_i$ remembered that [the Times had printed [a picture of himself$_i$] in its Sunday edition].
b. Physicists like yourself$_i$ are a godsend.
c. [Incriminating pictures of himself$_i$ published in the Times] worry Bill$_i$.
(Pollard & Sag 1992)

Claiming the distinction between core vs. exempt anaphors are necessary, Pollard & Sag (1992) explained as follows: anaphors that have a potential antecedent within the BD are constrained by syntactic Binding Theory, whereas anaphors that do not have potential antecedent within the BD are exempt from syntactic Binding Theory. This

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 85-94, Hakodate, Japan, September 13-15, 2019

can be interpreted that exempt binding is applied exclusively for the anaphors without a potential antecedent within a local domain.

The definition of BD (e.g., TSC & SSC) and core vs. exempt anaphor distinction were originally for languages like English; however, Korean is a language different from English, in that the anaphor inventory is composed of multiple anaphors. There are morphologically simple anaphors – *caki*, and *casin* – which have been known as Long-Distance Anaphors (LDAs), and complex anaphors – *caki-casin* and *pronoun-casin* – which have been discussed predominatly as locally bound anaphors (Moon 1995, Kang 1998, J-M Yoon 1989, etc.). Another difference is that in Korean both local and LD anaphors can violate TSC (cf. 3a), unlike in English. Furthermore, LDAs in Korean can even violate SSC as well (cf. 3b).

(3)
a. John$_i$-un [caki$_i$/casin$_i$/caki-casin$_i$-ka(i)
J-TOP self-NOM
choyko-la]-ko sayngkakhan-ta.
the best-be-REL-COMP think-DECL
'John$_i$ thinks that self$_i$ is the best.'
b. John$_i$-un [Mary-ka caki$_i$/casin$_i$/caki-casin$_i$ –ul
J-TOP M-NOM self-ACC
silhehanta]-ko sayngkakhan-ta.
hate-COMP think-DECL
'John$_i$ thinks that Mary does not like self$_i$.'

The cross-linguistic differences in BD and types of anaphors mentioned above were explained earlier in terms of GC-parametrization (Yang 1983, Manzini & Wexler 1987) as follows. While SSC defines the BD for local anaphors in Korean, LDAs has different BD – which is, the root clause. As for TSC, it is ineffective for defining Korean BD.

However, such analysis of dichotomy between local vs. LD anaphors in Korean were later challenged to be revised - by a series of experimental syntactic studies conducted sequentially. Kim & Yoon (2008) and Kim, Montrul & Yoon (2009) reported that Korean native speakers did not completely reject the sentences with LD-bound *caki-casin*; though less in degree compared to the other LDAs in Korean, the native speakers considered LD-bound *caki-casin* acceptable – even in the presence of a local antecedent. In addition, Kim & Yoon (2009), with their experimental results with Korean native speakers, demonstrated that the local anaphor *caki-casin* could be bound violating SSC as exempt anaphor when the anaphor was forced to be LD-bound; and when they were LD-bound, they behaved like exempt anaphors. Kim & Yoon (2013) and Kim (2013), with follow-up experimental studies including both LDAs and local anaphors in Korean, further discussed the possibility that even TSC-violating Korean anaphors – both local and LDAs – could be bound as exempt anaphors. This seems to support Exempt Binding approach over standard Binding Theory or GC-parameterization approach.

However, those previous studies mentioned had some limitations with respect to experimental design as well as methodological limitations in statistical analyses; and this calls for another study with revisions and proper methods of dealing with experimental data. Recently, E. Kim & Yoon (forthcoming) reconfirmed the findings of Kim & Yoon (2009) by using different experimental methods covering up some design flaws of the previous study. This means, despite the methodological weaknesses in experimental design or analyses in the series of previous studies, what seems still obvious is the possibility LD-binding of Korean local anaphor *caki-casin*.

Therefore, setting aside some unsolved theoretical issues in Korean binding (e.g., what comprises BD, etc.), the current study attempts to focus merely on apparent cases of LD-binding of *caki-casin* that violates SSC, to re-examine properties of LD-bound *caki-casin* in different linguistic levels – syntax, semantics and pragmatics. The study is based on the original study of Kim and Yoon (2009) by modifying their materials and changing the method of result analysis under more proper statistical methods that have not been covered in the previous experimental studies. The research questions of the current study are the following:

1) Does Korean LD-bound *caki-casin* show preference for Subject antecedent compared to non-Subject antecedent?
2) Is Korean LD-bound local anaphor *caki-casin* interpreted more like a co-referential pronoun than locally bound anaphor?

3) Does Korean LD-bound local anaphor *caki-casin* show preference for logophoric antecedents compared to those with less logophoricity?

The following several sub-sections will briefly introduce syntactic, semantic as well as pragmatic properties of LD-bound *caki-casin* that are tested in the present study.

1.1 Syntactic Properties of LD-bound *caki-casin*

Though the issue as to whether the Binding Domain of Korean local anaphors is TSC or SSC was debatable (Kim & Yoon 2009, 2013, Kim 2013), what is still undebatable is that Korean anaphors – be local or LDAs – can violate SSC and be bound LD as exempt anaphor. The sentences in (4) show the examples of SSC-violating exempt anaphors in English (cf. 4a) and Korean (cf. 4b).

(4)
a. John$_i$ believes that Mary despises [everyone but himself$_i$]
b. John$_i$-un [tongchanghoy-ka caki-casin$_i$ –lul
J-TOP alumni association-NOM self-ACC
sokyessta]-ko sayngkakhan-ta.
deceived-COMP think-DECL
'John$_i$ thinks that the alumni association deceived self$_i$.'

When anaphors are bound violating SSC, another related issue that has often come up to the discussion is the structural properties of the antecedents. It is well-known that the antecedent's structural prominence plays an important role in the well-formedness of LD-binding. O'Grady (1987) and Kim (2000) proposed structural hierarchy of the LD-antecedent focusing on Korean LDA *caki*, suggesting that structurally prominent antecedent such as Subject or Topic makes sentences with LDAs more well-formed. Choi and Kim (2007), in their experimental study of Korean anaphor processing with *caki* and *casin,* demonstrated that sentences where *caki* was bound by the matrix Subject were preferred by Korean native speakers. Han & Storoshenko (2012) viewed *caki* as both local and LDAs by applying core vs. exempt anaphor analysis and mentioned about Subject orientation of the antecedent. E. Kim et al (2013) conducted a sentence processing study with Korean multiple anaphors and reported that *caki-casin* showed preference for local Subject, rather than the matrix Subject with bi-clausal sentences. However, their study did not provide contextual information and presented the local Subject as a potential binder for *caki-casin*, thus resulted in investigating *caki-casin* as core anaphor only. Though the issue of Subject orientation has been discussed mostly for LDAs like *caki*, we also wanted to test whether LD-bound *caki-casin* also shows sensitivity to such structural factors.

As for the LD-binding of local anaphors, Cole et al. (2001) argued that Chinese local phrasal anaphor *ziji* can turn into exempt anaphors (logophors) and showed Subject-oriented property. Likewise, Kim & Yoon (2009) showed that grammatical–structural factors (subject vs. non-subject antecedents) also affected the acceptability of the long-distance binding of Korean local anaphor *caki-casin*. That is, though exempt anaphors are judged not to be constrained by structural factors such as grammatical relation, such structural factors may also influence determining the well-formedness of exempt binding. In line with those previous studies, the present study also tests whether the structural prominence of the LD- antecedent - such as being a Subject can facilitate the acceptability of LD-bound *caki-casin,* when compared to the sentences with non-Subject antecedent. The relevant pair of the sentences contrasted with Subject vs. non-Subject are shown in (5) below.

(5)
a. **John$_i$**-un [tongchanghoy-ka **caki-casin$_i$**–i
J-TOP alumni-NOM self-NOM
swumki-n pimil-ul alanayssta-ko]
hid-REL secret-ACC found-out- COMP
malhayss-ta.
said- DECL (Subject antecedent)
'John$_i$ said that the alumni found out the secret that self$_i$ hid.'
b. Na-nun **John$_i$**-hantheyse [tongchanghoy-ka
I-TOP John-from alumni-NOM
caki-casin$_i$–i swumki-n pimil-ul
self-NOM hid-REL secret-ACC
alanayssta-ko] tuless-ta.
found-out- COMP heard- DECL
'I heard from John$_i$ that the alumni found out the secret that self$_i$ hid.' (Non-subject antecedent)

1.2 Semantic Properties of LD-bound *caki-casin*

According to Pollard & Sag (1992), exempt anaphors are co-referential rather than referentially dependent on the antecedent, as shown in (6). In (6a) below, the anaphor *himself* is bound within the local BD and the underlined elliptical VP in the sentence *Bill did (so), too* is interpreted sloppily (i.e. Bill defended Bill…) in neutral contexts without special pragmatic information. On the other hand, in case of exempt binding as in (6b), the possibility of sloppy reading is considerably reduced; instead, the strict reading (i.e. Bill thinks that an article written by John…) becomes more possible and even preferred, compared to the cases like (6a). Huang and Liu (2001) argued that this can serve as a diagnostic for discriminating core vs. exempt anaphor; Runner et al. (2006) verified this in their experimental study of anaphor processing.

(6)
a. John$_i$ defended himself$_i$ against the committee's accusations.
Bill did (so), too (=Bill defended Bill >John…).
b. John$_i$ thinks [that an article written by himself$_i$ caused the uproar].
Bill does (so), too (= Bill thinks that an article written by John >Bill…).
(Kim & Yoon 2009)
c. **John$_i$**-un [tongchanghoy-ka **caki-casin$_i$**–i
J-TOP alumni-NOM self-NOM
swumki-n pimil-ul alanayssta-ko]
hid-REL secret-ACC found-out- COMP
malhayss-ta.
said- DECL
'John$_i$ said that the alumni found out the secret that self$_i$ hid.'
Bill-to kulessta.
B- too so-DECL
'Bill does (so), too.'

The underlined part of the sentence in (6c) shows the elliptical VP in Korean with LD-bound *caki-casin*. Kim and Yoon (2009) originally found that the rate of strict readings were significantly higher than sloppy readings in such case of LD-bound *caki-casin*. However, given that local binding dominantly yields sloppy readings rather than strict readings, we have to check if the sloppy vs. strict reading decreases/increases as the possibility of LD-binding increases. Therefore, in the present study, within-speaker responses between acceptability scores and the choice of sloppy vs. strict readings will be measured related to each other to find decreasing vs. increasing patterns of sloppy vs. strict readings according to their acceptability of LD-bound *caki-casin*.

1.3 Pragmatics of LD-bound *caki-casin*

It is well-known that LD-bound exempt anaphors are sensitive to pragmatic/logophoric factors (Kuno 1987, Sells 1987, Huang & Liu 2001, Oshima 2007, etc.). Instead of being constrained by syntactic binding conditions, exempt anaphors should meet pragmatic conditions to be considered legitimate. For example, exempt binding is well-formed if the LD-antecedent has a canonical role in a discourse context. In the theory of logophoricity proposed in Sells (1987), logophoricity is divided into three component roles as follows: i) SOURCE: the agent communicating the propositional content; ii) SELF: one whose mental state or attitude the content of the proposition describes; iii) PIVOT: one with respect to whose (space-time) location the content of the proposition is evaluated.

Sells (1987) further claimed that canonical order for the above three roles are the following: SOURCE>SELF>PIVOT. This can be shown in (7) below: The structural distance between the antecedent and anaphor and structural relation between them (i.e., no c-command) are identical in (7a) and (7b), but there is a clear degree of contrast in terms of acceptability of the sentences.

(7)
a. [Incriminating pictures of himself published in the Times] have been worrying John for some time.
b.*? [Incriminating pictures of himself published in the Times] accidentally fell on John's head.

The judgments in (7) reflect that *John* can be identified as a logophoric center – by being a SELF (and thus also a PIVOT) - in (7a), whereas in (7b) it can only be a PIVOT.

As for the canonical order among the three logophoric centers, it has been reported differently across languages: Huang & Liu (2001) argued that SELF seemed to play more crucial role than the others in Chinese, while Kim & Yoon (2009)

reported that the similar hierarchy as in Sells (1987) was found in Korean. However, the later study of Kim & Yoon (2013b) demonstrated that SELF got a slightly higher acceptability than SOURCE. Since previous experiments yielded conflicting results, this study pursues to re-examine the logophoric hierarchy of different logophoric centers with LD-bound *caki-casin*. The test sentences representing different logophoric roles[1] in Korean are shown in (8).

(8)

a. **John$_i$**-un [tongchanghoy-ka **caki-casin$_i$**–i
J-TOP alumni-NOM self-NOM
swumki-n pimil-ul alanayssta-ko]
hid-REL secret-ACC found-out- COMP
malhayss-ta.
said- DECL
'John$_i$ said that the alumni found out the secret that self$_i$ hid.' (SOURCE)

b. **John$_i$**-un [tochanghoy-ka **caki-casin$_i$**–i
J-TOP alumni-NOM self-NOM
swumki-n pimil-ul alanayssta-ko]
hid-REL secret-ACC found-out- COMP
sayngkakhayss-ta.
thought- DECL
'John$_i$ thought that the alumni found out the secret that self$_i$ hid.' (SELF)

c. [Mary-ka **caki-casin$_i$**-ul chaca oass-ul ttay],
M-NOM self-ACC search-come-REL when
John$_i$-un (pro -ul) pankapkey maca cwuess-ta.
J-TOP gladly greeted- DECL
'When Mary came to see self, John greeted (her) gladly.' (PIVOT)

2 Research Method

2.1 Hypotheses & Predictions

Our specific hypotheses and predictions based on the research questions introduced earlier are the following:

1) Korean native speakers will regard the sentences where LD-bound *caki-casin* has Subject antecedent more acceptable than those with non-Subject antecedent.
2) Korean native speakers will interpret elliptical VPs with more preference for strict readings compared to sloppy readings, as LD-bound *caki-casin* is considered more acceptable.
3) Korean native speakers will regard sentences with LD-bound *caki-casin* more acceptable, especially when the LD antecedent has canonical logophoric roles compared with the cases of less logophoric antecedents.

2.2 Participants

Forty-three Korean native speakers (Age range = 36~57) residing in and near Seoul, South Korea, who were raised monolingually, participated in the experiment.

2.3 Task, Materials, and Procedure

The main task was an acceptability judgment task using 5-point Likert scales, accompanied by preferential interpretation task. The stimuli for the acceptability judgment was constructed based on 4 logophoric conditions (SOURCE, SELF, PIVOT, less logophoric antecedent) and 2 GRs of the LD antecedents (Subject, non-Subject)[2], composed of 100 Korean sentences - 40 target items representing LD-bound *caki-casin* and 60 fillers (35 ungrammatical distractors and 25 sentences with other purposive fillers (e.g., local binding, TSC-only violation, multiple potential antecedents, backward binding, etc.)).

Each test sentence was presented with immediately following paired-elliptical VP (marked with underline); and the participants had to judge the acceptability of the given test sentence and then choose the preferred interpretation of the underlined elliptical VP.[3] For the elliptical VP,

[1] The sentences with PIVOT were constructed following Kim & Yoon (2009). Also, the sentences with distinct logophoric antecedents were compared to those with less logophoric antecedent which has a similar form as (8c), but with a matrix Subject that has different logophoric role from the Subject in the adjunct clause, as Kim & Yoon (2009) originally constructed. This type of sentences yielded significantly less acceptability scores than the sentences with canonical roles shown in (8).

[2] For constructing the sentences with non-Subject antecedents, we designed the sentences violating SSC that represent SOURCE and SELF only for logophoric roles. Other types of sentences with different logophoric roles were not counted for the comparison of GRs, so as to avoid further influence from confounding factors and interactions.

[3] We asked them to respond to VP ellipsis regardless of the acceptability, since we wanted to see whether and how the participants' responses with acceptability and those with the interpretations under VP ellipsis are consistent to each other.

three interpretation choices (A: Sloppy reading, B: Strict reading, C: Neither) were provided for the participants. The test items have the basic format of the following.

(9) Test Item Format
The Target sentence with LD-bound *caki-casin*
[Unacceptable 1 2 3 4 5 Acceptable]

John-to Kulessta (The sentence with VP-ellipsis)

Interpretation of the underlined part:
(A) The sentence representing sloppy reading
(B) The sentence representing strict reading
(C) Neither

Though there were other trials in the previous studies asking for possibility of strict vs. sloppy interpretations in VP-ellipsis or sometimes asking possibilities together with preferential choice in each items (Kim & Yoon 2013, 2013b, Kim 2013), we stick to the method of preferential choice, since preferential choice can represent stronger interpretation pattern out of ambiguity among dual or multiple possibilities - without confusing the participants.

2.4 Statistical Analysis

In order to investigate acceptability of the sentences with GRs of the antecedent as well as with distinct logophoric roles, the ordered logistic regressions[4] were conducted. As for the frequency of responses from the preferential choices (sloppy vs. strict) and their relation to the acceptability score of each item, a χ^2-Test and Rank Biserial Correlation were conducted. For the responses of acceptability with sloppy vs. strict readings, the acceptability scores 1 and 2 were coded as 'Low', while the scores 4 and 5 were coded as 'High' to see the correlation with sloppy vs. strict reading preference. The responses with medial acceptability (score 3) were dropped for the correlational analysis, since the responses with neutral (or uncertain) acceptability of the target sentences do not represent relevant interpretations in the elliptical VP.

3 Results

Overall results with the sentences of LD-bound *caki-casin* are the following: First of all, the sentences with LD-bound SSC-violating *caki-casin* were regarded as acceptable in majority of cases. Out of 1464 responses, the frequency of responses for each score was as follows: Score 1 = 109, Score 2 = 87, Score 3 = 125, Score 4 = 253, and Score 5 = 900). This seems to show that even the local anaphor *caki-casin* in Korean can have LD-antecedent if necessary. The details of the results related to properties of distinct linguistic levels are presented in the following several sub-sections.

3.1 Syntax of LD-bound *caki-casin*: Subject vs. Non-subject Antecedent

The results with different GRs for the LD-antecedent of *caki-casin* demonstrated the following. The sentences with Subject antecedent got more responses with higher acceptability scores, compared to those with non-Subject antecedent. The pattern of the results with Subject vs. non-Subject antecedents is shown in Table 1 and Figure 1 below.

Score / Antecedent	1	2	3	4	5
Subject	138	122	165	327	1139
Non-Subject	19	18	26	86	274
Total	157	140	191	413	1413

Table 1: Acceptability by GRs of Antecedent

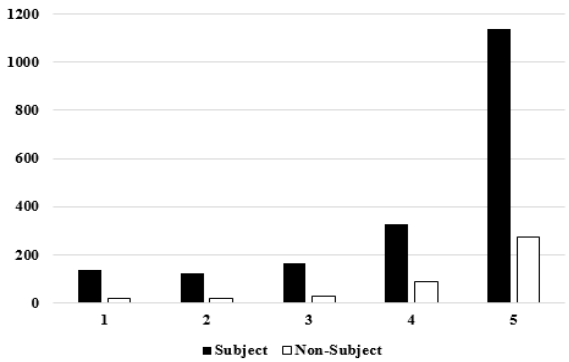


Figure 1: Acceptability by GRs of Antecedent

[4] The acceptability scores (from 1 to 5) are ordinal, but the variables such as logophoric roles, GRs of the antecedents are categorical. Therefore, it is unreasonable to use parametric tests (such as *t*-tests or ANOVAs). That is why we chose the ordered logistic regressions, χ^2-Test and Rank Biserial Correlation for the analysis.

The ordered logistic regressions revealed that the Subject vs. non-Subject differences significantly influenced the acceptability scores (t=25.501, $p<1.9\times10^{-143}$).

3.2 Semantics of LD-bound *caki-casin*: Interpretation with VP-ellipsis

The results with sloppy vs. strict readings in the elliptical VP showed that the responses with high acceptability scores (4 and 5) showed more responses of strict reading preference (Responses for sloppy reading: 445/1153, Responses for strict reading: 708/1153). On the other hand, those with low acceptability scores (1 and 2) showed more responses for sloppy reading preference overall (Responses for sloppy reading: 104/196; Strict reading: 92/196).

The result of χ^2-test revealed that the relation between the frequency of sloppy-strict readings and acceptability scores is significant (χ^2=10.484, df=1, $p<0.00121$). Ranked Biserial Correlation analysis further demonstrated that the relation between the two factors was significant (R=0.07812, $p<0.0011$). The mosaic plot from the correlation analysis is given in Figure 2.

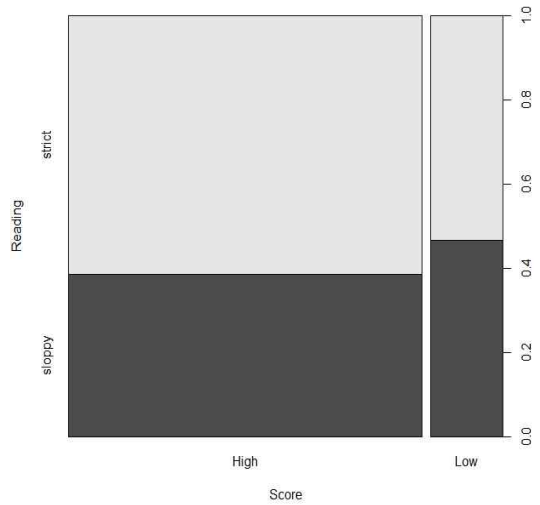


Figure 2: Preferential Interpretations in VP-ellipsis by Acceptability scores

As shown in Figure 2 above, the sentences with LD-bound *caki-casin* got much more high scores than low scores (as the width of the bars between High and Low shows). When the sentences got 4 and 5 for acceptability (i.e., High), strict reading choice was dominant, compared to the sentences with Low acceptability scores (1 and 2). However, it is notable that even with the highly acceptable sentences, there were still robust portion of sloppy readings. Also, even with low scores, strict reading rate seems not lower than that of sloppy readings.

3.3 Pragmatics of LD-bound *caki-casin*: Logophoric Roles of the Antecedents

As for logophoric roles of the LD-antecedents, the results patterned with those of Kim & Yoon (2009). The sentences with canonical logophoric roles got significantly higher frequency of higher acceptability scores than those with less logophoric antecedents. Furthermore, the canonical hierarchy of Sells (1987) was reconfirmed with SOURCE getting the highest frequency of the responses for higher acceptability scores than the other roles, while PIVOT getting the lowest among the three logophoric roles. The frequency of acceptability scores by different logophoric roles is shown in Table 2 and Figure 3 below.

Score / Logophoric roles	1	2	3	4	5
SOURCE	23	18	45	113	433
SELF	45	45	49	103	392
PIVOT	51	36	42	75	216
Less logophoric	25	27	22	50	84
Total	144	126	158	341	1125

Table 2: Acceptability by Logophoric Roles

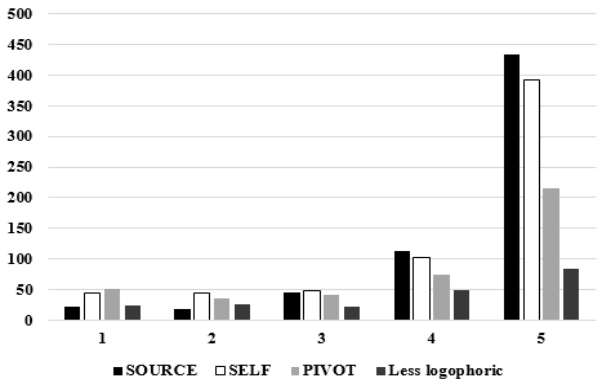


Figure 3: Acceptability by Logophoric Roles

The analysis using ordered logistic regressions showed that the differences found in the comparisons of between-logophoric roles in terms of acceptability scores were all significant (t=10.313, $p<2.611\times10^{-13}$).

4 Discussion

Through the result patterns of our experiment, the first hypothesis about syntactic properties of LD-bound *caki-casin* was supported: Korean native speakers indeed considered the sentences where *caki-casin* is LD-bound; and the sentences with Subject antecedent were regarded as more acceptable than those with non-Subject antecedent. Though *caki-casin* is not a typical LDA such as *caki* or *casin*, syntactic properties such as structural prominence of the LD-antecedent facilitated acceptability of the sentences.

Secondly, Korean native speakers interpret elliptical VPs with more preference for strict readings, compared to sloppy readings, especially when *caki-casin* is bound LD and considered more well-formed. However, though choice of the strict readings was more dominant with higher acceptability scores, it is noteworthy that the choice for sloppy readings was also found to a robust degree. This result seems to imply that we may have to reconsider our assumption about sloppy vs. strict readings in VP-ellipsis as a valid diagnostic that distinguish exempt anaphors from core anaphors. Charnavel and Sportiche (2016) noted that using diagnostic properties to argue for exempt status is risky, since there are few properties that actually distinguish core and exempt anaphors categorically. If this is true, we should seek for alternative ways to figure out types of binding the native speakers apply in each item.

On the other hand, we can interpret the result in such a way that the domain for core vs. exempt binding may overlap in Korean. In other words, regardless whether the anaphor is bound within the BD or outside the legitimate BD, Korean native speakers are open to the possibilities of interpreting the anaphor either using syntactic constraints or pragmatic conditions. If this chances to be the case, the explanation seems to go with the argument by Pollard & Xue (2001) for Chinese anaphor. However, to verify this claim, we need another follow-up experiment that include test materials testing core and exempt binding possibilities in one sentence simultaneously – with dual potential antecedents – one resolved by core binding and the other by exempt binding. Further discussions and applications should follow for this matter in the future study.

Finally, Korean native speakers considered sentences with LD-bound *caki-casin* with more responses of higher acceptability when the LD antecedent has canonical logophoric roles compared to the cases with less logophoric antecedent (i.e., where the antecedent in the matrix clause and the Subject in the embedded clause had different logophoric roles). Also, the hierarchy proposed by Sells (1987) was again reconfirmed as in Kim and Yoon (2009). This seems to support our third hypothesis.

5 Conclusion

The current study investigated what comprises the syntactic, semantic and pragmatic properties of LD-binding of Korean local anaphor *caki-casin*. Throughout the results of the current study, we can tentatively conclude that despite all the weak points found in the previous studies with respect to experimental designs as well as methodological problems in the analysis of the results, it is true that local anaphor *caki-casin* in Korean can be bound LD if extra-grammatical conditions are met.

The current study reconfirmed the findings of Kim and Yoon (2009) by recapitulating the results that acceptability of LD-bound *caki-casin* is affected syntactically by different GRs and pragmatically by distinct logophoric roles. Finally, when LD-bound with highly acceptable degree, more coreferential readings seem to be involved as if it indicates that the anaphor is bound outside the domain of syntax.

Nevertheless, due to some unexpected pattern of the results (e.g., robust degree of sloppy readings found with highly acceptable cases of LD-binding), we need to re-consider and re-examine the issue as to whether sloppy vs. strict readings in the elliptical VP can still be a valid diagnostic for the discrimination between core vs. exempt binding; and if not, we have to seek for what can be an alternative diagnostic that can work for the core vs. exempt distinction. Furthermore, it is not enough to investigate the exempt binding of local anaphor after the resolution of the antecedents: We need more studies of examining sentence processing patterns of the native speakers that can show us the resolution procedures of various properties of the antecedents (e.g., in terms of logophoricity) and how the participants respond to such sentences.

References

Isabelle Charnavel and Dominique Sportiche. 2016. Anaphor binding: What French inanimate anaphors show. Linguistic Inquiry, 47(1): 35–87.

Noam Chomsky. 1980. On binding. Linguistic Inquiry, 11, 1-46.

Noam Chomsky. 1981. Lectures in Government and Binding. Foris, Dordrecht.

Noam Chomsky. 1986. Knowledge of Language: Its Nature, Origin and Use. New York: Praeger.

Kwangill Choi and Youngjin Kim. 2007. Caykwitaymyengsa-uy Ta.uyseng Hayso-Kwaceng: Ankwu-Wuntong Pwunsek (Ambiguity Resolution Processes of Reflexives: Eye-tracking Data), The Korean Journal of Experimental Psychology, 19(4): 263-277.

Peter Cole, Gabriella Hermon and C.-T. James Huang. 2001. Introduction. Long-distance reflexives: The State of the Art. Syntax and Semantics, 33: xiii-xlvii.

Chung-hye Han and Dennis Ryan Storoshenko. 2012. Semantic binding of long-distance anaphor caki in Korean. Language, 88(4): 764-790.

C.-T. James Huang and Chen-Sheng Luther Liu. 2001. Logophoricity, attitude, and ziji at the interface. Syntax and Semantics, 33: 141-195.

Beom-Mo Kang. 1998. Mwunpep-kwa Ene Sayong: Khophes-ey Kipanhan Caykwisa 'caki', 'casin', 'caki-casin'-uy Kinung Pwunsek-ul Cwungsim-ulo (Grammar and the use of language: Korean reflexives 'caki, 'casin', and caki-casin). Kwuk-e-hak (Korean Linguistics), 31: 165-204.

Ji-Hye Kim, Silvina Montrul and James Hye-Suk Yoon. 2009. Binding Interpretations of Anaphors by Korean Heritage Speakers. Language Acquisition, 16: 3-35.

Ji-Hye Kim, and James Hye-Suk Yoon. 2008. An Experimental Syntactic Study of Binding of Multiple Anaphors in Korean. Journal of Cognitive Science, 9(1): 1-30.

Ji-Hye Kim and James Hye-Suk Yoon. 2009. Long-Distance Bound Local Anaphors in Korean – An Empirical Study of the Korean Anaphor Caki-casin, Lingua, 119: 733-755.

Ji-Hye Kim and James Hye-Suk Yoon. 2013. The tensed-S condition (TSC) and the determination of the binding domain of anaphors in Korean. In Proceedings of Japanese/Korean Linguistics, 22 (JK22).

Ji-Hye Kim and James Hye-Suk Yoon. 2013b. TSC (Tensed S Condition) and Korean Anaphors: An Experimental Study of Binding Domain and VP Ellipsis in Koreatongcahngn, Paper Presented at the 19th International Congress of Linguists, University of Geneva

Eunah Kim, Myeong Hyeon Kim and James Hye-Suk Yoon. 2013. An experimental investigation of online and offline binding properties of Korean reflexives. In Proceedings of Japanese/Korean Linguistics, 22 (JK22).

Eun Hee Kim and James Hye-Suk Yoon. Forthcoming. Experimental evidence supporting the overlapping distribution of core and exempt anaphors: Reexamination of long-distance bound caki-casin in Korean. Lingua.

Ji-Hye Kim. 2013. Demarcating Local vs. Long-Distance Binding: Re-examining Tensed S Condition (TSC) in Korean. Linguistic Studies, 27: 135-154.

Soo-Yeon Kim. 2000. Acceptability and preference in the interpretation of anaphors. Linguistics, 38(2): 315-353.

Susumu Kuno. 1987. Functional syntax. Chicago: University of Chicago Press.

Rita Manzini, and Ken Wexler. 1987. Parameters, binding theory and learnability. Linguistic Inquiry, 18: 413-444.

Seung-Chul Moon. 1995. An Optimality Approach to Long-Distance Anaphors. Doctoral dissertation, University of Washington, Seattle.

William O'Grady. 1987. The interpretation of Korean anaphora: the role and representation of grammatical relations. Language 63: 251–277.

David Y. Oshima. 2007. On empathic and logophoric binding, Research on Language and Computation, 5(1):19-35.

Carl Pollard and Ivan A. Sag. 1992. Anaphors in English and the scope of Binding Theory. Linguistic Inquiry, 23: 261-303.

Carl Pollard and Ping Xue. 2001. Syntactic and non-syntactic constraints on long-distance reflexives. In Peter Cole, Gabriella Hermon & James Huang (eds.) Long distance reflexives: Syntax and semantics series. 317-342. Academic Press: New York.

Jeffrey Runner, Rachel Sussman and Michael Tanenhaus, M. K. 2006. Processing reflexives and pronouns in picture noun phrases. Cognitive Science, 30: 193-241.

Peter Sells. 1987. Aspects of logophoricity. Linguistic Inquiry, 18 (3): 445-479.

Dong-Whee Yang. 1983. The extended binding theory of anaphors. Language Research, 19: 169-192.

Jeong-Mi Yoon. 1989. Long-distance Anaphors in Korean and their Crosslinguistic Implications. Chicago Linguistic Society, 25: 479-495.

The *persuade*-construction in Korean controls nothing

Juwon Lee
Jeonju University
Department of English Education
303, Cheonjam-ro, Wansan-gu,
Jeollabuk-do, Korea
juwonlee@khu.ac.kr

Sanghoun Song
Korea University
Department of Linguistics,
145, Anam-ro, Seongbuk-gu,
Seoul, Korea
sanghoun@korea.ac.kr

Abstract

Quite a few studies of the control constructions in Korean have assumed that *seltukha*-'persuade' in Korean serves as an object control verb like its corresponding translation *persuade* in English. However, this study shows that the claim is based on dubious theoretic and empirical premises. In particular, we argue that the *seltukha*-construction in Korean is not an object control providing several pieces of empirical evidence. The evidence shows that the object in the matrix clause and the subject of the embedded clause can simultaneously appear in *seltukha*-construction and they are not necessarily co-indexed with each other. Building upon the non-control analysis, we suggest the Anti-redundancy Hypothesis; two NPs referring to the same entity or having the same form tend not to appear right next to each other in order to avoid redundancy. Finally, we discuss some possible extensions of the non-control analysis to other related constructions.

1 Introduction

In many prior syntactic studies in Korean, the verb *seltukha*- 'persuade' has been considered as involving an object control. However, the present study argues that the verb is tangential to a syntactic control in spite of the correspondence in translation.

First, consider some canonical examples of English subject and object controls given in (1) (see e.g. Farkas 1988; Borsley 1999; Jackendoff & Culicover 2003; Sag *et al*. 2003; *inter alia*). The missing subjects of the embedded clauses are marked with the gap:

(1) a. John$_i$ tried [____$_i$ to leave].
b. John$_i$ promised Mary$_j$ [____$_i$ to leave].
c. John$_i$ persuaded Mary$_j$ [____$_j$ to leave].

The control constructions, irrespective of subject or object control, have the two defining properties in common across languages. The first property is that the subject of the embedded clauses must be silent. As illustrated in (2), no explicit NP can appear in the subject position of the *to*-infinitive clause.

(2) a. John tried [(*John/*he/*him/*himself) to leave].
b. John promised Mary [(*John/*he/*him/*himself) to leave].
c. John persuaded Mary [(*Mary/*she/*her/*herself) to leave].

A second property is that the silent subject of the *to*-infinitive clause must be co-indexed with an argument of the matrix clause, as illustrated in (3).

(3) a. John$_i$ tried [____ $_{i/*j}$ to leave].
b. John$_i$ promised Mary$_j$ [____ $_{i/*j/*k}$ to leave].
c. John$_i$ persuaded Mary$_j$ [____ $_{*i/j/*k}$ to leave].

In (3) the silent element (controllee) in the embedded clauses is necessarily co-indexed with an explicit argument (controller) in the matrix clauses. These are two fundamental features of the subject or object control constructions.

Like the English *persuade*-construction, the corresponding Korean sentence exemplified in (4) has often been assumed to be a control construction (Monahan 2004; Cormack and Smith 2004; Kwon and Polinsky 2006, among others).

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 95-103, Hakodate, Japan, September 13-15, 2019

(4) *John$_i$-i Mary$_j$-lul/-eykey* [____$_{*i/j/*k}$ *ttena- tolok*]
John-Nom Mary-Acc/-Dat leave-Comp
seltukhay-ss-ta.
persuade-Pst-Dec
'John persuaded Mary to leave.'

If *seltukha-* 'persuade' is indeed a control verb like *persuade*, then it is predicted that the *seltukha*-construction shares the two crucial properties of control constructions presented above. However, this paper provides several counterexamples to the premises: (i) the subject of the embedded clauses can explicitly appear, and (ii) the subject of the embedded clauses is not necessarily co-indexed with the matrix object. We present several pieces of critical evidence to support the argument that the *seltukha*-construction is not a control. Instead, the present study presents a *pro*-drop analysis of the *seltukha*-construction (Choe 2006; Park 2013).

2 Previous Analyses

Some prior studies on control construction in Korean are discussed in this section.

2.1 PRO Analysis

Traditionally, PRO is on the subject position of the *to*-infinitive clause as in the following (see Chomsky 1981, Chomsky 1995):

(5) John persuaded [Mary$_i$] [$_{TP}$ PRO$_i$ to leave].

The PRO in (5) is obligatory, which means that the null element should be co-indexed with a matrix argument as indicated by the subscript. Note that the constraint on the co-indexation does not go for the arbitrary PRO and the optional PRO exemplified in (6a-b) respectively.

(6) a. [PRO$_{arb}$/*Anyone to invite Jane] would be good.
b. Robert$_i$ knows that it is important [PRO$_{i/j}$ to read the book].

Despite the different behaviors, all types of PROs commonly are caseless and take place in non-finite clauses.

This PRO analysis is not appropriate for the *seltukha*-construction due to several distributional facts. First, the subject of the *tolok*-clause can appear explicitly as in the following (see examples like (7) in Monahan 2003 and a sentence similar to (8) in Cormack & Smith 2004):

(7) *Jane-i* ____ [*Minswu-ka o-tolok*]
Jane-Nom Minswu-Nom come-Comp
seltukhay-ss-ta.
persuade-Pst-Dec
'Jane persuaded Minswu to come.'

(8) *sensayngnim-i Minswu emeni-lul* [*Minswu-ka*
teacher-Nom Minswu mother-Acc Minswu-Nom
peptay-ey ka-tolok] *seultukhay-ss-ta*.
law school-to go-Comp persuade-Pst-Dec
'The teacher persuaded Minswu's mother that Minswu should go to law school.'

Second, the subject of the *tolok*-clause is nominative, not caseless as shown in (7) and (8). Third, the subject of the *tolok*-clause in (8) is not co-indexed with the matrix object. Thus, there is no *a posteriori* proof for believing that the null element in the *seltukha*-construction involves the essential properties of PROs.

2.2 Movement Analysis

Kwon & Polinsky (2006) and Kwon *et al*. (2010) argue that the two sentences in (9) are not derivationally related, but they are distinct construcitons. This implies that scrambling of (9a) does not result in (9b). They call (9a) ACC1 and (9b) ACC2 respectively.

(9) a. *Jane-i Minswu-lul* [____ *tomangka-tolok*]
Jane-Nom Minswu-Acc run away-Comp
seltukhay-ss-ta.
persuade-Pst-Dec [ACC1]
'Jane persuaded Minswu to run away.'
b. *Jane-i* [____$_k$ *tomangka-tolok*]$_j$ [*Minswu$_k$-lul*
Jane-Nom run away-Comp Minswu-Acc
____ $_j$ *seltukhay-ss-ta*].
seltukhay-ss-ta. [ACC2]
'Jane persuaded Minswu to run away.'

Following the movement analysis of English controls (Hornstein 1999), they argue that in (9a) the subject of the *tolok*-clause moves to the object position in the matrix clause, and the tail of this A-chain is deleted. This construction is called ACC1; i.e., the forward obligatory control (OC). On the other hand, the *tolok*-clause in (9b) moves leftward while the subject of this clause moves to the right. This construction is called ACC2; i.e., the non-obligatory control (NOC).

However, this movement analysis is less tenable for multiple reasons. First of all, there seems to be no syntactic mechanism about case assignment and

case alternation. As shown in (7) and (8) above, the subject of the embedded clause must be nominative. It is not clear how exactly the nominative subject in the embedded clause is switched to the accusative or dative object in the matrix clause. Second, if the matrix object in (9a) really comes from the subject of the *tolok*-clause, then we should say that accusative objects are generally licensed at least in two different ways, base generation as in (10) and movement as in (9a).

(10) *Jane-i Minswu-lul ttayli-ess-ta.*
Jane-Nom Minswu-Acc hit-Pst-Dec
'Jane hit Minswu.'

A naturally occurring question is why we must use the two different ways to license accusative objects in the matrix clauses (cf. Occam's Razor). Third, the active sentences in (9) should have their passive counterparts. Given that the object associated with the accusative case in active sentences is promoted to the subject position in the passive counterpart, the passive sentence in (11) should be derived from the two distinct constructions in (9).

(11) *Minswu-ka Jane-eyuyhay* [___ *tomangka-tolok*]
Minswu-Nom Jane-by run away-Comp
seltuktoy-ess-ta.
persuade.Pass-Pst-Dec
'Minswu was persuaded to run away by Jane.'

Then, the sentence in (11) should be ambiguous between obligatory and non-obligatory control readings since a passive sentence shares the truth-condition with its active counterpart unless a specific operator such as quantifiers and subject-oriented adverbs intervenes. Because they leave how the logical form is made across the active-passive forms out of discussion, there is no clue for such an ambiguity as of yet. Fourth, the silent subject in (9a) is not necessarily co-indexed with the matrix object if a specific context is plausibly given (see §4.3). Fifth, the matrix object and the subject in the embedded clause can refer to different individuals as shown in (8) and (12). The movement analysis cannot derive these sentences.

(12) *sensayngnim-i* [*Minswu-ka peptay-ey ka-*
teacher-Nom Minswu-Nom law school-to go-
tolok] *Minswu emeni-lul seultukhay-ss-ta.*
Comp Minswu mother-Acc persuade-Pst-Dec
'The teacher persuaded Minswu's mother that Minswu should go to law school.'

Sixth, the following sentence devoid of both the object and the subject can be allowed if the omitted NPs are recoverable within the discourse context. It is not clear how the movement analysis can account for sentences like this.

(13) *John-i* ___ [___ *ttena-tolok*] *seltukhay-ss-ta.*
John-Nom leave-Acc persuade-Pst-Dec
'John persuaded someone to leave.'

In sum, the movement analysis causes the latent problems that cannot be fully accounted for.

2.3 Semantic Control

Cormack and Smith (2004) suggest that obligatory semantic control pertains to the control constructions as illustrated in (14).

(14) *Jane-i Minswu$_j$-lul* [*pro$_j$ tomangka-tolok*]
Jane-Nom Minswu-Acc run away-Comp
seltukhay-ss-ta.
persuade-Pst-Dec
'Jane persuaded Minswu to run away.'

As is well known, *pro* can be either a bound variable or a referential pronoun. This means that *pro* in (14) does not have to be co-indexed with the object in the matrix clause. In order to ensure the co-indexation between the matrix object and *pro* in the *tolok*-clause, Cormack & Smith (2004: 66) posited Meaning Postulate as follows:

(15) Meaning Postulate 1:
For all s, x, y, if 'PERSUADE s y x' holds then y is Agent in Event s (s is the Event argument of PERSUADE, y the persuadee, x the persuader, where x and y are individuals).

Due to this Meaning Postulate, the agent of embedded clause in (14) must be identical to the persuadee in the matrix clause. A fundamental assumption in Cormack & Smith (2004: 68) is such that the lexical meaning of *seltukha-* 'persuade' is identical to that of the English *persuade* and the Meaning Postulate is straightforwardly applied to the meaning of the two verbs. However, this does not account for the sentences such as (8) and (12). Cormack & Smith (2004: 68, footnote 23) assume that the sentences like (8) are acceptable due to a

causative coercion of some kind, but they do not dwell on how exactly such a coercion saves the sentence. In addition, according to Monahan (2004), Cormack & Smith's (2004) analysis predicts that the sentence in (16) should allow the second interpretation that is not available for the sentence:

(16) *Minswu-nun* [*ku yepaywu-ka kica-eykey* Minswu-Top the actress-Nom reporter-to *inthepyupat-tolok*] *seltukhay-ss-ta.* interview.Pass-Tolok persuade-Pst-Dec
'Minswu persuaded the actress to get interviewed by the reporter.'
#'Minswu persuaded the reporter to interview the actress.'

Following that Cormack & Smith's (2004: 72) claim that the subject of the embedded clause is agent, the sentence has the correct meaning such as *Minswu persuaded the actress to get interviewed by the reporter*. Indeed, the subject of a passive can be an agent, as shown in the following.

(17) *ku-ka ilpwule saca-eykey mek-hi-ess-ta.* he-Nom intentionally lion-to eat-Pass-Pst-Dec
'He was intentionally eaten by the lion.'

In (17) the adverb *ilpwule* 'intentionally' requires an agent and the subject is the agent. However, this does not mean that the lion is not an agent in the event of eating the person. Likewise, the reporter in (16) is also an agent in the event of interviewing the actress. Then Cormack & Smith's (2004) analysis should license the unwanted interpretation such as *Minswu persuaded the reporter to interview the actress*. We present in the following section some data to support the argument that *seltukha-* 'persuade' is not a control verb even though the *seltukha*-construction may or may not have a control meaning (OC or NOC) depending upon the given contexts. The complement of *seltukha-* 'persuade' is omissible just as with the complement(s) of other transitive verbs in Korean (a *pro*-drop language).

3 Two NPs: Controller and Controllee

In this section we argue that the two NPs (the matrix object and the subject of the *tolok*-clause) can appear simultaneously in *seltukha*-construction, but they tend not to. Along the line of the tendency, the present study suggests the Anti-redundancy Hypothesis as a tendency.

3.1 Co-occurrence of the Two NPs

As shown above, one of the fundamental properties of control constructions is that the controllee must be silent. If *seltukha*-construction is really a control construction, we expect that it behaves like the *persuade*-construction in English; it should never allow the two NPs to appear at the same time. This appears to be verified as follows:

(18) ??*John-i* *Mary$_j$-lul* [*Mary$_j$-ka ttena-tolok*] John-Nom Mary-Acc Mary-Nom leave-Acc *seltukhay-ss-ta*. persuade-Pst-Dec
(lit.) 'John persuaded Mary Mary to leave.'

The sentence in (18) sounds odd. This oddness of the sentence can be accounted for if it is an object control like its English counterpart. In other words, as object control generally requires the subject of the embedded clause to be silent, (18) sounds rather awkward.

Alternatively, we can say that the awkwardness arises because the referential subject *Mary-ka* in the *tolok*-clause violates the Condition C (i.e., an r-expression is free; *John$_i$ adored John$_{j/*i}$*). If the subject in the embedded clause is a pronoun as in (19), the sentence sounds better than (18).

(19) ?*John-i* *Mary$_j$-lul* [*kunye$_j$-ka ttena-tolok*] John-Nom Mary-Acc she-Nom leave-Acc *seltukhay-ss-ta*. persuade-Pst-Dec
(lit.) 'John persuaded Mary she to leave.'

This improvement is an unexpected result if the *seltukha*-construction is an object control in a genuine sense because (object) control does not allow an explicit controlee. Note that the sentence in (19) is not constrained by Condition C in that the subject in the bracketed clause is pronominal. If Condition C (or more broadly, constraints of binding theory) is really responsible for the awkwardness of the sentence in (18), then the sentence in (19) should be fine. Nonetheless, (19) still sounds a bit awkward though it is better than (18). In short, (19) can be a problem for both the control analysis and the binding analysis of the appearance of the two explicit NPs in *seltukha-*

construction. Moreover, if an anaphor comes as the subject in the embedded clause as in (20), the sentence is fine:

(20) *John-i* *Mary$_j$-lul* [(*kunye)* *casin$_j$-i* *ttena-tolok*]
John-Nom Mary-Acc she self-Nom leave-Acc
seltukhay-ss-ta.
persuade-Pst-Dec
'John persuaded Mary herself to leave.'

The sentence in (20) is a strong counterexample to the object control analysis of *seltukha*-construction. Since Korean allows a long-distance binding of anaphor, the sentence in (20) does not violate Condition A (or other conditions) of the binding theory for Korean. Note, however, that the sentences like (18) seem not to be totally unacceptable, and this fact is not likely to be accounted for by Condition C. In the next subsection, we propose an alternative hypothesis to account for the appearance of the two explicit NPs in *seltukha*-construction.

3.2 Anti-redundancy Hypothesis

The present analysis is such that (18) sounds rather awkward for the reason that the two referential NPs referring to the same individual tend not to appear right next to each other in order to avoid redundancy. Based on this observation, we propose the Anti-redundancy Hypothesis formulated in (21).

(21) **Anti-redundancy Hypothesis**:
Two NPs referring to the same entity or having the same form *tend* not to appear right next to each other, since the iteration sounds redundant.

This hypothesis can account for the improvement of the acceptability in (19) compared to (18). The referential matrix object and the pronominal subject in the *tolok*-clause are co-indexed, and they appear right next to each other, so the sentence sounds somewhat redundant. However, (19) is better than (18) since the latter sounds more redundant than the former. In (18) the NPs have almost the same form (*Mary-lul* and *Mary-ka*), but in (19) one is a referential NP (*Mary-lul*) and the other a pronominal NP (*kunye-ka*). The iteration of the same form serves to increase the redundancy. In (20) the anaphor *kunye casin-i* 'herself-Nom' is co-indexed with the matrix referential object, and they appear right next to each other. If the contrastive focus is assigned to the anaphor, the redundancy effect seems to be dramatically alleviated. Likewise, the addition of the adverb *cikcep* 'by herself' reduces the redundancy in the following sentences:

(22) *John-un* *Mary$_j$-lul* [*Mary$_j$-ka* *cikcep*
John-Top Mary-Acc Mary-Nom by herself
ttena-tolok] *seltukhay-ss-ta*.
leave-Acc persuade-Pst-Dec
(lit.) 'John persuaded Mary Mary to leave by herself.'

(23) *John-un* *Mary$_j$-lul* [*kunye$_j$-ka* *cikcep*
John-Top Mary-Acc she-Nom by herself
ttena-tolok] *seltukhay-ss-ta*.
leave-Acc persuade-Pst-Dec
(lit.) 'John persuaded Mary she to leave by herself.'

The adverb *cikcep* 'by herself' imposes the contrastive focus on the subject of the *tolok*-clause. This reduction of redundancy renders the sentences more acceptable. Note that (23) sounds better than (22), as is expected.

Another way to reduce the redundancy is putting an adverbial expression between the matrix object and the *tolok*-clause, as underlined in (24).

(24) *sensayngnim-un* *Jane-ul* *achim-pwuthe*
teacher-Top Jane-Acc morning-from
kankokhakey [*Jane-i/kunye-ka* *hakkyo-ey*
earnestly Jane-Nom/she-Nom school-to
o-tolok] *seltukhay-ss-ta*.
come-Comp persuade-Pst-Dec
'From the morning the teacher has earnestly persuaded Jane to come to school.'

The sentence in (24) sounds much better than the sentences without the adverbial expressions. Similarly, if something like a pause or parenthesis is inserted between the two NPs to lengthen the linear distance between them, the sentence sounds more acceptable.

(25) *sensayngnim-un* *Jane-ul* *pause/um…/kulenikka*
teacher-Top Jane-Acc PAUSE/um…/I mean
[*Jane-i/kunye-ka* *hakkyo-ey* *o-tolok*]
Jane-Nom/ she-Nom school-to come-Comp
seltukhay-ss-ta.
persuade-Pst-Dec
(lit.) 'The teacher persuaded Jane pause/un…/I mean Jane to come to school.'

Neither the control analysis nor Condition C can account for this phenomenon.

Moreover, if we scramble the matrix object as to increase the linear distance between the two NPs as presented in (26) and (27), the sentences sound much better than (18).

(26) *John-un* [*Mary$_j$-ka* *ttena-tolok*] *kankokhakey*
John-Top Mary-Nom leave-Acc earnestly
Mary$_j$-lul *seltukhay-ss-ta*.
Mary-Acc persuade-Pst-Dec
(lit.) 'John earnestly persuaded Mary Mary to leave.'

(27) *John-un* *kankokhakey* *Mary$_j$-lul*
John-Top earnestly Mary-Nom
seltukhay-ss-ta [*Mary$_j$-ka* *ttena-tolok*].
persuade-Pst-Dec Mary-Nom leave-Comp
(lit.) 'John earnestly persuaded Mary Mary to leave.'

The acceptability of these sentences can be accounted for by the Anti-redundancy Hypothesis; reducing the redundancy makes the sentences sound more acceptable.

In order to remove the redundancy completely, one of the two NPs should be omitted. As expected, such sentences are clearly acceptable.

(28) a. *Jane-un* *Minswu-lul* [___ *tomangka-tolok*]
Jane-Top Minswu-Acc run way-Comp
seltukhay-ss-ta.
persuade-Pst-Dec
'Jane persuaded Minswu to run away.'
b. *Jane-un* [___ *tomangka-tolok*] *Minswu-lul*
Jane-Top run way-Comp Minswu-Acc
seltukhay-ss-ta.
persuade-Pst-Dec
'Jane persuaded Minswu to run away.'
c. *Jane-un* [*Minswu-ka* *tomangka-tolok*]
Jane-Top Minswu-Nom run way-Comp
seltukhay-ss-ta.
persuade-Pst-Dec
'Jane persuaded Minswu to run away.'

The three examples in (28) are ACC1, ACC2, and NOM, respectively, under the taxonomy of Kwon & Polinsky (2006) and Kwon *et al*. (2010).

If the missing NPs are sufficiently recoverable with reference to the context, the sentence in (13), repeated in (29), sounds fairly acceptable.

(29) *John-i* ___ [___ *ttena-tolok*] *seltukhay-ss-ta*.
John-Nom leave-Comp persuade-Pst-Dec
'John persuaded someone to leave.'

The acceptability of (29) can be explained by the resolution of the redundancy.

Note finally that some exceptions of Condition C are allowed if a contrastive focus takes place, as shown in the following.

(30) *Sally$_j$-ka* *John-i* *anila* *Sally$_j$-lul*
Sally-Nom John-Nom Neg Sally-Acc
ttayli-ess-ta.
hit-Pst-Dec
'Sally hit Sally, not John.'

If this exception is allowed, then the sentence in (22) may be accounted for by Condition C. However, sentences like (24) and (25) are still acceptable even though the subject of the embedded clause does not receive a contrastive focus. Then Condition C is not sufficient to account for the data. In addition, exceptions of this kind (converting ungrammatical sentences to grammatical sentences) seem not to be theoretically in the right direction and cast a serious doubt on the existence of Condition C itself. Thus we believe that it is better to stick with the Anti-redundancy Hypothesis to account for co-occurrence of the two NPs in *seltukha*-construction.

4 Co-indexation

Co-indexation between the matrix object and the subject of the embedded clause (OC) is required for *persuade*-construction in English. However, it is shown in this section that such co-indexation is not necessary for *seltukha*-construction in Korean.

4.1 Two Explicit NPs

When the matrix object and the subject of the *tolok*-clause appear simultaneously, they are not required to refer to the same individual, as already shown in (8) and (12). They are repeated below:

(31) *sensayngnim-i* *Minswu emeni-lul* [*Minswu-ka*
teacher-Nom Minswu mother-Acc Minswu-Nom
peptay-ey *ka-tolok*] *seultukhay-ss-ta*.
law school-to go-Comp persuade-Pst-Dec
'The teacher persuaded Minswu's mother that Minswu should go to law school.'

(32) *sensayngnim-i* [*Minswu-ka* *peptay-ey* *ka-*
teacher-Nom Minswu-Nom law school-to go-
tolok] *Minswu* *emeni-lul* *seultukhay-ss-ta.*
Comp Minswu mother-Acc persuade-Pst-Dec

'The teacher persuaded Minswu's mother that Minswu should go to law school.'

The acceptability of these sentences indicates that they are not a control construction at all.

Similarly, in the following the matrix object and the subject of the *tolok*-clause refer to different individuals who have the same name.

(33) [Context: There are two people whose name is Minji in the same class. They are close friends. Minji$_k$ does not want to attend school anymore. The teacher tried to persuade Minji$_k$ to come to school again, but failed. So the teacher talked to Minji$_j$ in order to make Minji$_j$ persuade Minji$_k$ to come to school again.]
a. *sensayngnim-un Minji$_j$-lul* [*Minji$_k$-ka*
teacher-Top Minji-Acc Minji-Nom
tasi hakkyo-ey o-tolok] *seltukhay-ss-ta.*
again school-to come-Comp persuade-Pst-Dec
(lit.) 'The teacher persuaded Minji$_j$ Minji$_k$ to come to school.'
b. *sensayngnim-un* [*Minji$_k$-ka tasi hakkyo-ey*
teacher-Top Minji-Nom again school-to
o-tolok] *Minji$_j$-lul seltukhay-ss-ta.*
come-Comp Minji-Acc persuade-Pst-Dec
(lit.) 'The teacher persuaded Minji$_j$ Minji$_k$ to come to school.'

In short, it is not necessary for the two explicit NPs in the *seltukha*-construction to refer to the same individual. This runs counter to the assumption that *seltukha-* 'persuade' in Korean is a control verb.

4.2 One Explicit NP: Subject of *Tolok*-clause

The default reading of the sentence in (34) is that the teacher persuaded Mary to go to law school.

(34) *sensayngnim-un* ____ [*Mary-ka peptay-ey*
teacher-Top Mary-Nom law school-to
ka-tolok] *seltukhay-ss-ta.*
go-Comp persuade-Pst-Dec
'The teacher persuaded Mary to go to law school.'

However, if a certain context is given as in (35), the silent matrix object is not necessarily co-indexed with the subject of the *tolok*-clause.

(35) [Context: The teacher talked to Mary's mother about Mary's career. Mary's mother wanted Mary to go to medical school, but…]
sensayngnim-un ____ [*Mary-ka peptay-ey*
teacher-Top Mary-Nom law school-to
ka-tolok] *seltukhay-ss-ta.*
go-Comp persuade-Pst-Dec
(lit.) 'The teacher persuaded Mary's mother Mary to go to law school.'

In sum, the co-indexation is not required for *seultukha*-constructions when the matrix object is silent, although the co-indexation reading is the most natural reading without a specific context.

4.3 One Explicit NP: Matrix Object

The default reading of the sentence in (36) is that the teacher persuaded Mary's mother to go to law school.

(36) *sensayngnim-i Mary emeni-lul*
teacher-Nom Mary mother-Acc
[____ *peptay-ey ka-tolok*] *seltukhay-ss-ta.*
law school-to go-Tolok persuade-Pst-Dec
'The teacher persuaded Mary's mother to go to law school.'

However, if a context is given as in the following, the matrix object and the understood subject of the *tolok*-clause can refer to different individuals (see the same point in Park 2013: 3, footnote 3).

(37) A: Why did Mary go to law school?
B: *sensayngnim-i Mary emeni-lul*
teacher-Nom Mary mother-Acc
[___ *peptay-ey ka-tolok*] *seltukhay-ss-ketun.*
law school-to go-Comp persuade-Pst-since
(lit.) 'Because the teacher persuaded Mary's mother Mary to go to law school.'

Summarizing, if either the object or the subject is missing, the default reading is the co-indexation reading, but it is not a requirement.

4.4 No explicit NP

In (38) both the matrix object and the subject of the *tolok*-clause are missing.

(38) A: What did the teacher say to Mary's mother? Why did Mary go to law school?
B: *sensayngnim-i* ___ [___ *peptay-ey*
teacher-Top law school-to
ka-tolok] *seltukhay-ss-ketun.*
go-Tolok persuade-Pst-Dec
(lit.) 'Because the teacher persuaded Mary's mother Mary to go to law school.'

The referents of the missing NPs are recoverable from the context: the persuadee is Mary's mother

and the person who went to law school is Mary. The non-control reading is possible for the *seltukha*-construction.

5 A Preliminary Analysis

The data discussed so far lead us to conclude that *seltukha-* 'persuade' is not a control verb although *seltukha*-constructions can be interpreted as OC or NOC in certain contexts. The matrix object is licensed by *seltukha-*, and the subject of the *tolok*-clause is licensed by the lexical verb in the clause. They do not necessarily refer to the same individual whether they appear or not in *seltukha*-constructions. These syntactic and semantic properties of *seltukha*-construction can be roughly represented like the following:

(39) NP-Nom (NP$_i$-Acc) [$_{CP}$ [(NP$_{i/j}$-Nom) … V]-*tolok*] *seltukha-*.

The matrix subject can be also omitted, but here we focus on the two NPs under discussion. When they are omitted since Korean is a *pro*-drop language, their referents are identified according to the linguistic or utterance context.

If the subject of the *tolok*-clause is not necessarily co-indexed with the matrix object, the prediction is that it can be also co-indexed with the matrix subject in a certain context. This seems to be borne out in the following:

(40) *Chelswu$_i$-ka sacang-lul* [*casin$_i$-i ku il-ul math-tolok*] *seltukhay-ss-ta.*
Chelswu-Nom president-Acc self-Nom the task-Acc undertake-Comp persuade-Pst-Dec
(lit.) 'Chelswu$_i$ persuaded the president himself$_i$ to undertake the task.'

(41) *Chelswu$_i$-ka* [*casin$_i$-i ku il-ul math-tolok*] *sacang-lul seltukhay-ss-ta*
Chelswu-Nom self-Nom the task-Acc undertake-Comp president-Acc persuade-Pst-Dec
(lit.) 'Chelswu$_i$ persuaded the president himself$_i$ to undertake the task.'

While admitting that (41) sounds better than (40), we judge both acceptable. The difference in the degree of acceptability seems to be largely due to either the tendency of the accusative object to be closer to the head verb than other complement or the distance between the anaphor and its antecedent (or probably both).

6 Extension

In this paper we have focused on the data with accusative matrix object. However, the persuadee can be realized as a dative NP as in (42).

(42) *John-i Mary-eykey* [____ *ttena-tolok*]
John-Nom Mary-Dat leave-Comp
seltukhay-ss-ta.
persuade-Pst-Dec
'John persuaded Mary to leave.'

The default reading of (42) is the co-indexation reading, but we believe that this co-indexation is not necessary. In (43) the two NPs appear simultaneously, and the sentence sounds quite odd.

(43) ??*John-i Mary-eykey* [*Mary-ka ttena-tolok*]
John-Nom Mary-Dat Mary-Nom leave-Comp
seltukhay-ss-ta.
persuade-Pst-Dec
'John persuaded Mary to leave.'

(43) is not impossible though it sounds redundant. If this redundancy decreases as in (44), the sentence becomes better.

(44) *John-i Mary-eykey cengmal kankokhakey*
John-Nom Mary-Dat really earnestly
[*Mary-ka ttena-tolok*] *seltukhay-ss-ta.*
Mary-Nom leave-Comp persuade-Pst-Dec
'John really earnestly persuaded Mary to leave.'

In addition, the two NPs in the *seltukha*-construction can refer to different individuals, as illustrated in (45).

(45) *sensayngnim-i Minswu emeni-eykey* [*Minswu-ka peptay-ey ka-tolok*] *seultukhay-ss-ta*.
teacher-Nom Minswu mother-Dat Minswu-Nom law school-to go-Comp persuade-Pst-Dec
'The teacher persuaded Minswu's mother that Minswu should go to law school.'

Taken together, we can say that the *seltukha*-constructions with dative object are not a control construction either.

7 Conclusion

We argued in this paper that *seltukha-* 'persuade' in Korean is not a control verb. This opposes quite a few prior syntactic studies in which syntactic

derivation and similarity in meanings are invalidly mixed up. In particular, the two properties of *seltukha*-construction were presented as evidence for non-control analysis of *seltukha*- 'persuade': (i) the matrix object and the subject of the embedded clause can simultaneously appear in *seltukha*-constructions, and (ii) they are not necessarily co-indexed with each other. In addition, we proposed the Anti-redundancy Hypothesis that two NPs referring to the same entity or having the same form tend not to appear right next to each other, since the iteration renders the entire expressions redundant. This accounts for the oddness of some *seltukha*-constructions with the two NPs. Finally, the non-control analysis can be applied to other related constructions in Korean while a more detailed examination awaits further research.

Acknowledgments

We would like to thank the audiences of the Spring Meeting of Korean Association for Corpus Linguistics held in Kyung Hee University in 2019 for their comments, questions, and discussions.

References

Borsley, Robert. 1999. *Syntactic Theory: A Unified Approach*.

Choe, Hyon Sook. 2006. On (backward) object control in Korean. *Harvard Studies in Korean Linguistics* XI, 373–386. Kyunggi: Hanshin.

Chomsky, Noam. 1981. *Lectures on Government and Binding*. Dordrecht: Foris.

Chomsky, Noam. 1995. *The minimalist program*. Cambridge, MA: MIT Press.

Cormack, A., and N. Smith. 2004. Backward control in Korean and Japanese. *University College London Working Papers in Linguistics* 16: 57–83.

Farkas, Donca F. 1988. On Obligatory Control. *Linguistics and Philosophy* 11.1: 27-58.

Hornstein, Norbert. 1999. Movement and control. *Linguistic Inquiry* 30: 69–96.

Kwon Nayoung, Monahan Philip J. and Maria Polinsky. Object Control in Korean: A Backward Control Impostor. In *Movement Theory of Control*. Amsterdam: John Benjamins.

Kwon, Nayoung and Maria Polinsky. 2006. Object control in Korean: Structure and processing. *Japanese/Korean Linguistics* 15: 249–262.

Monahan, Philip J. 2003. Backward object control in Korean. In *Proceedings of the 22nd West Coast Conference on Formal Linguistics*, Gina Garding & Mimu Tsujimura (eds), 356–369. Somerville MA: Cascadilla Press.

Park, Hong-Keun. 2012. Control Constructions in Korean Revisited. *Studies in Generative Grammar* 22.1: 1-22.

Ray Jackendoff and Peter W. Culicover. 2003. The Semantic Basis of Control in English. *Language* 79.3: 517-556.

Sag, Ivan A., Wasow, Thomas and Bender, Emily M. 2003. *Syntactic Theory: A Formal Introduction*, (2nd edition). Stanford, CA: CSLI Publications.

Pretrained language model transfer on neural named entity recognition in Indonesian conversational texts

Rezka Leonandya
Kata.ai
Jakarta, Indonesia
rezka@kata.ai

Fariz Ikhwantri
Kata.ai
Jakarta, Indonesia
fariz@kata.ai

Abstract

Named entity recognition (NER) is an important task in NLP, which is all the more challenging in conversational domain with their noisy facets. Moreover, conversational texts are often available in limited amount, making supervised tasks infeasible. To learn from small data, strong inductive biases are required. Previous work relied on hand-crafted features to encode these biases until transfer learning emerges. Here, we explore a transfer learning method, namely language model pretraining, on NER task in Indonesian conversational texts. We utilize large unlabeled data (generic domain) to be transferred to conversational texts, enabling supervised training on limited in-domain data. We report two transfer learning variants, namely supervised model fine-tuning and unsupervised pretrained LM fine-tuning. Our experiments show that both variants outperform baseline neural models when trained on small data (100 sentences), yielding an absolute improvement of 32 points of test F1 score. Furthermore, we find that the pretrained LM encodes part-of-speech information which is a strong predictor for NER.

1 Introduction

Named entity recognition (NER), the task of assigning a class to a word or phrase of proper names in text, is an essential ability for conversational agents to have. For example, in food delivery application, an agent needs to acquire information about the customer's food detail and address. NER is all the more challenging on conversational texts because of their noisy characteristics, such as typos, informal word variations, and inconsistent naming in named entities. Furthermore, conversational texts are often available in diverse domains and limited amount, making supervised training arduous due to data limitation. To learn from limited data, strong inductive biases are necessary. In this work, we explore transfer learning techniques as a way to help neural models learn and generalize from limited data on NER task in Indonesian conversational texts.

Transfer learning, or sometimes known as domain adaptation, is an important approach in NLP application, especially if one does not have enough data in the target domain. In such scenarios, the goal is to transfer knowledge from source domain with large data to target domain so as to improve the model performance on the target domain and prevent overfitting. Early research in transfer learning, especially with entity recognition in mind, were tackled by feature augmentation (Daumé, 2007), bootstrapping (Wu et al., 2009), and even rule-based approach (Chiticariu et al., 2010).

Recently, neural networks emerge as one of the most potent tools in almost all NLP applications. Although neural models have achieved impressive advancement, they still require an abundant amount of data to reach good performance. With limited data, neural models generalization ability is severely curtailed, especially across different domains where the train and test data distributions are different (Lake and Baroni, 2017). Therefore, transfer learning becomes more critical for neural models to enable them to learn in data-scarce settings.

Transfer learning in NLP is typically done with two techniques, namely parameter initialization (INIT) and multi-task learning (MULT). INIT approach first trains the network on source domain and directly uses the trained parameters to initialize the network on target domain (Lee et al., 2018), whereas

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 104-113, Hakodate, Japan, September 13-15, 2019

MULT simultaneously trains the network with samples from both domains (Aguilar et al., 2017). Recently, INIT approaches were made highlight by the incorporation of pretrained language models (Peters et al., 2018; Ruder and Howard, 2018; Radford, 2018) to neural models, reaching state-of-the-art performance across various NLP tasks.

Indonesian NER itself has attracted many years of research, from as early as using a rule-based approach (Budi et al., 2005) to more recent machine learning techniques, such as conditional random field (CRF) (Luthfi et al., 2014; Leonandya et al., 2015; Taufik et al., 2016), and support vector machine (SVM) (Suwarningsih et al., 2014; Aryoyudanta et al., 2016). The latest research was done by Kurniawan and Louvan (2018) where they investigated neural models performance with word-level and character-level features in Indonesian conversational texts.

In this paper, we apply and evaluate a recent technique of transfer learning, namely language model pretraining. We use the pretrained LM to extract additional word embedding input representations for the neural sequence labeler in Indonesian conversational texts. The work in this paper is organized as follows: We first train a language model on generic domain unlabeled Indonesian texts $\mathbb{U}$ (e.g., Wikipedia). We then use a smaller domain-specific source corpus $\mathbb{S}$ (e.g., task-oriented conversational texts) to either: (a) fine-tune the pretrained LM or (b) train a neural sequence labeler using the pretrained LM's representation as additional input. If we proceed with (a), then the next step is to train a neural sequence labeler on the target domain $\mathbb{T}$ (e.g., small-talk conversational texts) using fine-tuned LMs representation as additional input. If we proceed with (b), then the next step is to fine-tune the neural sequence labeler on the target domain corpus $\mathbb{T}$. We evaluate and compare our approach with other models, namely a neural sequence labeler without LM pretraining and a multi-task approach trained on varying amount of training data from $\mathbb{T}$.

2 Methodology

To allow models to learn from small conversational data, we introduce two variants of three-step training procedure. Both variants use additional input of word embedding representations derived from a bidirectional LSTM language model (biLM). The word embedding representations used in this paper is ELMo (Peters et al., 2018).

2.1 Transfer Learning

As mentioned briefly in Section 1, there are two variants of transfer learning used in this paper. Details of both approaches are shown in Figure 1.

Supervised fine-tuning With ELMo representations, this approach first trains a model using labeled data from source domain $\mathbb{S}$; next, it initializes the target model with the learned parameters; finally, it fine-tunes the target model using labeled data from target domain $\mathbb{T}$. All weights of the model except the pretrained LM are updated.

Unsupervised fine-tuning Unsupervised fine-tuning is inspired by ULMFiT (Ruder and Howard, 2018). Rather than training on the source domain and fine-tuning on the target domain, this approach fine-tunes the pretrained LM using unlabeled data from source domain $\mathbb{S}$; then, it trains the target model using labeled data from target domain $\mathbb{T}$. All weights of the model except the pretrained LM are updated.

2.2 Dataset

In this paper, we use Indonesian language. There are three datasets: unlabeled $\mathbb{U}$, source $\mathbb{S}$, and target $\mathbb{T}$. We evaluate our approach in the settings that $\mathbb{U}$ is of generic domain (newswire or formal) and the $\mathbb{S}$ and $\mathbb{T}$ are of specific domain (conversational).

	N	*DW*	*T*	*AVG*
TO train	10142	16930	129841	12.7
TO dev	1250	4291	16021	12.8
TO test	1205	3868	14610	12.12
ST train	11577	9034	434408	3.74
ST dev	3289	3749	12583	3.82
ST test	1641	2265	6300	3.83
TOL	916838	144028	6652122	7.25

Table 1: Number of sentences (*N*), number of distinct words (*DW*), number of tokens (*T*), and average sentence length (*AVG*) in TO , ST, and TOL dataset.

For the unlabeled data, we use Kompas-Tempo (newswire) (Tala, 2003) and Indonesian Wikipedia. Unless stated otherwise, we refer to the former and the latter as KT and Idwiki, respectively.

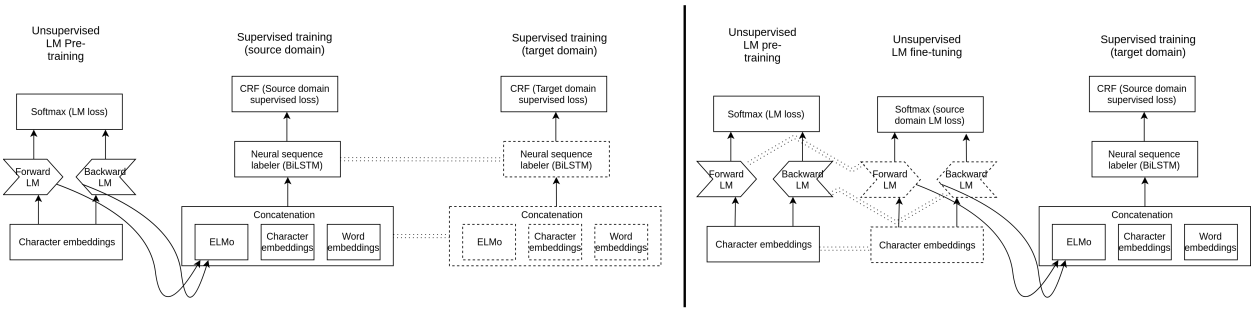


Figure 1: Double dash line represents weight transfer. **Top**: Supervised fine-tuning by fine-tuning the neural sequence labeler on the target domain. The CRF layer is replaced. **Bottom**: Unsupervised fine-tuning by fine-tuning the pretrained LM on the source domain.

Entity	Train	Test	Dev
AREA	78	6	6
CURRENCY	1213	140	139
DATETIME	2982	347	350
DURATION	339	26	38
EMAIL	226	21	23
LENGTH	169	14	20
LOCATION	6322	709	791
NUMBER	3966	404	471
PERSON	4031	477	480
PHONE	466	49	53
TEMPERATURE	70	4	8
VOLUME	58	4	8
WEIGHT	134	11	11

Table 2: Number of labels contained in the TO dataset. There are 13 labels in total.

For source and target domain data, we use our own manually labeled dataset, the same dataset introduced by (redacted for anonymity) , namely SMALL-TALK as the target domain and TASK-ORIENTED as the source domain. The former is a 16K conversational messages from users having small talk with a chatbot, whereas the latter contains 12K task-oriented messages such as movie tickets booking, and food delivery. For the rest of the paper, unless stated otherwise, we refer to the former as ST and the latter as TO.

For the unsupervised fine-tuning approach, since we do not need labeled data for the source domain $\mathbb{S}$, we can easily add more unlabeled dataset to our source domain $\mathbb{S}$. Conveniently, we have a large unlabeled Indonesian conversational texts at our disposal, which is a superset of the TO labeled dataset. We refer to this bigger unlabeled conversational dataset as TOL. Using this dataset, we can perform unsupervised fine-tuning with a larger data size compared to that of supervised fine-tuning.

Due to proprietary and privacy reasons, unfortunately, we cannot publish our Indonesian conversational texts. We present our dataset statistics in Table 1 and the label details in Table 2 and 3.

Entity	Train	Test	Dev
DATETIME	49	20	21
EMAIL	20	7	8
GENDER	241	64	85
LOCATION	2672	813	867
PERSON	2455	749	754
PHONE	44	18	21

Table 3: Number of labels contained in the ST dataset. There are 6 labels in total.

3 Experiments and results

3.1 Experiment setup

We use Rei et al.'s (Rei, 2017) implementation[1] for the multitask model and AllenNLP[2] for the rest of our models. For every combination of training dataset and model, we tune the dropout rate (Srivastava et al., 2014) by grid search on [0.25, 0.35, 0.5, 0.65, 0.75] for both the ELMo and the neural sequence labeler using the same random seed for all configurations. We do not tune other hyperparameters due to computational resource constraints.

For all models, we use identical data preprocessing: words are lowercased, BIO scheme is used, and start end tokens are excluded from the vocabulary. For the multitask model, we use identical settings to that of (Rei, 2017). For the rest of the

[1] https://github.com/marekrei/sequence-labeler

[2] https://github.com/allenai/allennlp

Models	F1 Dev	F1 Test
CNN-BiLSTM-CRF	85.94	85.85
+ELMo Idwiki	87.67	87.72
+ELMo KT	86.88	87.11
+Flair Idwiki	87.12	88.41
+Flair KT	71.16	71.73

Table 4: Accuracy of the baseline model and the baselines with the pretrained embeddings (ELMo and Flair trained on Idwiki and KT data) trained and evaluated on TO data.

models, we use different settings: word and character embedding sizes are set to 50 and 16, respectively. Character embeddings are formed by a CNN with 128 filters followed by highway layers and a ReLU activation layer. Both word and character embeddings are initialized randomly. ELMo embedding size is set to 1024. The LSTMs are set to have 200 hidden units and 2 layers. We apply L2 regularizer of 0.1 to all layers and early stopping is used with a patience of 25. We use Adam (Kingma and Ba, 2014) optimization with learning rate of 0.001 and we clip the gradient at 5.0. We run our experiments with batch size of 32 and epochs of 150. We evaluated all our experiments with CoNLL evaluation: micro-averaged F1 score based on exact span matching.

3.2 Impact of pretrained LM embeddings

To assess the impact of the pretrained LM embeddings, first we compare the performance of the baseline model to the baseline models with pretrained LM embeddings on NER task evaluated on TO dataset. We use CNN-BiLSTM-CRF by (Ma and Hovy, 2016) as our baseline model and ELMo as our pretrained LM embeddings. In addition to ELMo, we also use Flair (Akbik et al., 2018) as it reaches state-of-the-art performance on English NER task[3]. Flair also differs from ELMo because it is trained with character-level softmax. We use the default hyperparameters provided by (Akbik et al., 2018) to train Flair LM[4].

From Table 4 we can see that adding LM embeddings improves the overall performance. ELMo yields small absolute improvement of dev F1 score when trained on both Idwiki and KT than the baseline. Flair, although obtains roughly the same dev F1 score when trained on Idwiki, does not perform well when trained on KT. This result might be attributed to the fact that KT contains far fewer sentences compared to IdWiki (around 9 times fewer). We link this result to an observation made by (Yu et al., 2018) which stated that character language models are unstable when the training data is not big enough. Therefore, we do not proceed with Flair for the next experiment with our two approaches.

3.3 Main experiment

We experiment with three model groups. The first group is the baseline models, consisting of a single-task learning model using CNN-BiLSTM-CRF (Ma and Hovy, 2016) and a multitask approach (Rei, 2017), which uses an LSTM-BiLSTM-CRF with an additional LM loss. The second and third groups involve our first and second approaches, which are the unsupervised language model fine-tuning and the supervised neural sequence labeler fine-tuning, respectively.

We train the models in each group on different percentages of the target domain training data $\mathbb{T}$. We do this to assess the impact of our supervised and unsupervised fine-tuning approach when presented with varying amount of training data. Table 5 shows the number of entities of the ST data with different percentages of training sentences used. Table 6 shows the result of our experiments. All numbers shown in the table come from models with the best hyperparameter on the target domain validation set.

3.3.1 Model name conventions

Here we explain the patterns for naming our models in Table 6 and the rest of our paper: LM indicates an unsupervised step, whereas Sup is for the supervised step. A dash (-) shows a fine-tuning step (supervised or unsupervised), an underscore (_) represents the move from LM training to BiLSTM training, and a square bracket ([]) represents which dataset used in the unsupervised (LM) or the supervised (Sup) step. For example, one can interpret a model named LM[Idwiki]_Sup[TO-ST] as a model which: (1) uses pretrained LM trained on Idwiki

[3] as of 03/01/2019

[4] https://github.com/zalandoresearch/flair

Entity	1%-100	5%-502	10%-1004	25%-2511	50%-5022	75%-7533
DATETIME	2	4	8	18	27	39
EMAIL	2	5	3	7	11	17
GENDER	4	18	26	65	123	174
LOCATION	30	137	265	640	1345	2019
PERSON	20	122	251	606	1216	1829
PHONE	1	2	5	8	22	31

Table 5: Number of labels contained in the ST training dataset. The training data percentage is followed by the total number of training sentences.

Group	Model name	Unlabeled data for pretrained LM	Source domain data	Target domain test F1 score						
				1%	5%	10%	25%	50%	75%	100%
Baselines	CNN-BiLSTM-CRF	-	-	0.59	43.84	50.86	62.19	72.64	72.79	75.81
	Rei (2017)	-	-	41.44	59.43	68.03	74.30	80.47	83.23	85.12
Supervised fine-tuning	LM[Idwiki]_Sup[TO-ST]	Idwiki	TO	**73.17**	**77.25**	77.21	80.31	82.27	**83.54**	84.62
	LM[KT]_Sup[TO-ST]	KT	TO	71.84	75.16	76.42	79.11	82.88	83.18	84.49
Unsupervised fine-tuning	LM[Idwiki-TOL]_Sup[ST]	Idwiki	TOL	67.10	74.05	77.49	79.76	**83.82**	83.17	**85.78**
	LM[KT-TOL]_Sup[ST]	KT	TOL	64.48	75.55	**77.77**	**80.99**	83.09	82.74	85.76

Table 6: Experiment results on the ST test data. Models are trained on ST training data with varying number of training instances. Bold and underline indicates the highest and the second highest test F1 score, respectively.

Group	Model name	Unlabeled data for pretrained LM	Source domain data	Target domain dev F1 score						
				1%	5%	10%	25%	50%	75%	100%
Supervised fine-tuning	LM[Idwiki]_Sup[TO-ST]	Idwiki	TO	**71.12**	75.02	77.28	80.36	81.85	81.82	82.96
	LM[KT]_Sup[TO-ST]	KT	TO	70.42	73.79	75.12	79.95	81.39	82.04	82.33
Unsupervised fine-tuning	LM[Idwiki-TOL]_Sup[ST]	Idwiki	TOL	65.43	74.64	77.71	81.14	82.68	83.71	84.37
	LM[KT-TOL]_Sup[ST]	KT	TOL	63.90	**75.75**	77.84	**82.38**	83.75	83.26	**85.11**
Ablation	Sup[TO-ST]	-	TO	66.08	69.94	71.91	76.68	81.00	80.78	82.04
	LM[Idwiki]_Sup[ST]	Idwiki	-	55.99	69.15	71.94	74.92	79.60	80.31	80.71
	LM[KT]_Sup[ST]	KT	-	53.79	66.70	71.39	74.61	78.68	78.91	81.02
	LM[TOL]_Sup[ST]	TOL	-	61.59	75.05	**78.60**	82.00	**83.91**	**85.04**	84.18

Table 7: Experiment results on the ST dev data. The original models and the ablation models are trained on ST training data with varying number of training instances. Bold and underline indicates the highest and the second highest test F1 score, respectively.

dataset, (2) trains the BiLSTM on the source domain dataset (TO), and (3) fine-tunes the BiLSTM on the target domain dataset (ST).

3.3.2 Results discussion

Baseline and multitask Unsurprisingly, CNN-BiLSTM-CRF fails when the training data is as small as 100 sentences, reaching only 0.59% test F1 score. As the training data gradually increases, CNN-BiLSTM-CRF performs fairly well, reaching 75.81% test F1 score on the whole data. The multi-task model by Rei (2017) reaches 45.10% test F1 score when trained on 100 sentences training data. The absolute improvement from the baseline is huge considering that it only benefits from the LM loss and no additional signals from external labeled/unlabeled data are incorporated. It is also competitive compared to other models when trained on the whole data.

Supervised fine-tuning There is a significant increase of test F1 score using supervised fine-tuning compared to the baseline on every level of training data size. The gain from supervised fine-tuning is largest when the data is small. Both LM[Idwiki]_Sup[TO-ST] and LM[KT]_Sup[TO-ST] achieve the highest and second highest test F1 score when trained on 1% training data. However, when trained on 100% training data, supervised fine-tuning does not perform better than the multitask model. The gain of supervised fine-tuning seems to be diminishing as the training data grows larger. This hints that using supervised fine-tuning might not be necessary if labeled data is already present in adequate amount thus one can opt for simpler models such as the multitask model.

Unsupervised fine-tuning Overall, using unsupervised fine-tuning yields competitive result with the supervised fine-tuning. A notable difference is when the model trained with 1% and 100% training data. LM[Idwiki-TOL]_Sup[ST] and LM[KT-TOL]_Sup[ST] obtains the highest and the second highest F1 test score on 100% training data, respectively. On 1% training data, they fall behind the supervised fine-tuning by about 5 points of test F1 score. The huge unlabeled source domain data for the LM fine-tuning does not seem to be helping much on small training data. This hints that labeled data in small-moderate size is still more effective compared to unlabeled data which comes in massive size. Also, there does not seem to be a significant difference between the unsupervised fine-tuning and the multitask approach on 75% and 100% training data. Multitask approach seems to perform competitively when the training data size is vast enough. Again, this tells us that we might not need to perform unsupervised fine-tuning if labeled data is already present in adequate amount.

3.3.3 Ablation

An ablation study is conducted for both the supervised and unsupervised fine-tuning. We perform three ablations resulting in four models: two for the supervised fine-tuning group and one for the unsupervised fine-tuning. We carry out the ablations on the development set. Table 7 shows the result of the ablated models compared with the models of our two approaches.

The first ablation is removing the LM pretraining step but keeping the supervised fine-tuning using the source domain (Sup[TO-ST]). Notice that there is a sizable drop of test F1 score on every training data size compared to the supervised fine-tuning models, especially on 1% training data. The second ablation is omitting the intermediate supervised fine-tuning step while keeping the pretrained LM intact (LM[Idwiki]_Sup[ST] and LM[KT]_Sup[ST]). Without the supervised fine-tuning, the models' performance is markedly curtailed. Another interesting pattern is that removing the LM pretraining step does less harm than removing the supervised fine-tuning step, especially on small training data. On 1% training data, Sup[TO-ST] outperforms LM[Idwiki]_Sup[ST] and LM[KT]_Sup[ST] by significant margins. It seems that the supervised fine-tuning part is where the model benefits the most. Nonetheless, even without the supervised fine-tuning, the pretrained LM alone is already a huge reinforcement for the models.

The last ablation is conducted by excluding the unsupervised fine-tuning step, which is the model named LM[TOL]_Sup[ST]. We do the LM pretraining directly with the source domain data (`TOL`) without using the generic domain data. LM[TOL]_Sup[ST] obtains highest F1 test score on 10%, 50%, and 75% training data and second highest F1 test score on 5% and 25% training data. This is quite surprising considering that LM[TOL]_Sup[ST] is trained with the same fashion as LM[Idwiki]_Sup[ST] and LM[KT]_Sup[ST]: no fine-tuning and only BiLSTM with LM pretraining. The only difference is in the dataset used for the LM pretraining. We hypothesize that the gain stems from the fact that `TOL` is of conversational domain, whereas `Idwiki` and `KT` is of generic domain. Overall, LM[TOL]_Sup[ST] also performs competitively compared to both the supervised and unsupervised fine-tuning models on every level of training data size. This highlights that one might not need to perform fine-tuning from a generic domain if a large in-domain unlabeled data is already at hand.

4 Analysis

From the observations of the previous section, we conclude that there are three main highlights regarding our fine-tuning approaches: (1) supervised transfer performs best on 1% training data, (2) unsupervised transfer obtains the best F1 score on 100% training data, and (3) pretraining the LM directly on the source domain (`TOL`) without fine-tuning works really well, attaining comparable result with both fine-tuning approaches. Here, we seek to establish more understanding of why (1), (2), and (3) happen.

Recent research have discovered that pretrained LM induce useful knowledge such as syntactic information for downstream tasks (Blevins et al., 2018; Linzen et al., 2016; Gulordava et al., 2018). Motivated by these findings, we formulate a hypothesis that the supervised and unsupervised fine-tuning models also learn advantageous knowledge for predicting named entities, namely part-of-speech (POS)

information. It is also known that POS has been used as features for NER and they help improve the overall performance (Curran and Clark, 2003). To test this hypothesis, we train a diagnostic classifier (Veldhoen et al., 2016; Hupkes et al., 2018) using the models' hidden state representations as features on an Indonesian part-of-speech conversational dataset. In this experiment, the diagnostic classifier is a simple classifier (single layer neural network with a softmax output) to predict the POS tags. During training, the weights of the models are all frozen except the diagnostic classifier's weights. To train the diagnostic classifier, we use our own manually labeled POS dataset because to the best of our knowledge, there is no labeled Indonesian part-of-speech conversational dataset that is publicly available. We also cannot publish our POS dataset due to proprietary and privacy reasons. We provide the dataset statistics in Table 8 and the label details in Table 9.

POS tag schema Here we explain the schema used in our POS tag dataset. The dataset is annotated by an Indonesian linguist. All 21 tags shown in Table 9 are the same tags as the English Penn Treebank POS tags (Marcus et al., 1994) except for: (1) CDI, NEG, PRL, SC, VBI, and VBT, which were taken from the Indonesian POS Tagset 1 (Pisceldo, 2009) and (2) NUM, PNP, and X, which were created based on the Indonesian grammar references. NUM stands for numbers (e.g., *tujuh*-seven). PNP is number pronouns, which is used to refer to person or object identified with numbers (e.g., *keduanya*-the two of them). X is for unrecognized words (e.g., *wkwk*-wkwk).

	N	*DW*	*T*	*AVG*
POS train	4108	7077	34843	8.48
POS dev	1008	2109	9115	9.04
POS test	1283	3209	11111	8.66

Table 8: Number of sentences (*N*), number of distinct words (*DW*), number of tokens (*T*), and average sentence length (*AVG*) in POS dataset.

We train the diagnostic classifier using the hidden representations of: (1) the pretrained LM (for both the supervised and unsupervised fine-tuning approaches), (2) the neural sequence labeler (BiLSTM) trained on the source domain (for supervised fine-tuning), and (3) the fine-tuned language model on the source domain (for unsupervised fine-tuning). We report the accuracy on the development set. The test is carried out to check which step encodes the part-of-speech information better for both the supervised and unsupervised fine-tuning models. Note that here there is no target domain involved since we only want to know the quality of the POS information learned from the fine-tuning step.

Entity	Dev	Test	Train
CC	110	166	412
CDI	17	25	73
DT	295	350	1120
FW	561	505	1679
IN	439	502	1493
JJ	182	238	750
MD	271	365	1126
NEG	110	142	447
NN	1857	2306	7236
NNP	1341	1652	5213
NUM	280	370	1178
PNP	1	2	2
PRL	7	12	27
PRP	247	337	1018
RB	241	307	883
RP	42	61	193
SC	306	339	1126
SYM	1207	1481	4787
UH	382	511	1589
VBI	235	286	892
VBT	807	924	2817
WP	169	216	722
X	8	14	60

Table 9: Number of labels contained in our POS dataset. There are 23 labels in total.

Table 10 shows the result of the diagnostic classifier trained on our POS dataset. In addition to the diagnostic classifier, we provide a simple baseline model that outputs the most frequent tag in the training set for a given word. For OOV word, the simple baseline model outputs the most frequent label in the training data. Given inputs from the pretrained LM, the diagnostic classifier performs considerably well on the development set. LM[TOL], LM[Idwiki], and LM[KT] obtains higher accuracies than the simple baseline model. LM[TOL] reaches higher accuracy than LM[Idwiki] and LM[KT]. This result aligns with our previous result that `TOL` is of conversational domain hence it is more useful for downstream tasks in conversational domain. The diagnostic classifier reaches slightly better accuracy when trained using inputs from the fine-tuned LM, yielding an absolute improvement of 1 point of dev accuracy. Since the differences between the accuracies of the diagnostic classifier are minuscule, this may explain why LM pretraining without fine-

Hidden representations input	Model name	Pretrained LM data	Source domain	Dev acc
-	Most frequent tag	-	-	84.51
Pretrained LM	LM[Idwiki]	Idwiki	-	86.14
	LM[KT]	KT	-	86.68
	LM[TOL]	TOL	-	92.05
Fine-tuned LM	LM[Idwiki-TOL]	Idwiki	TOL	92.52
	LM[KT-TOL]	KT	TOL	93.08
Neural sequence labeler (BiLSTM)	LM[Idwiki]_Sup[TO]	Idwiki	TO	63.64
	LM[KT]_Sup[TO]	KT	TO	63.83
	Sup[TO]	-	TO	58.62

Table 10: Diagnostic classifiers trained on part-of-speech (POS) task using the different models' hidden representations as input. We present the accuracy on the development set.

tuning (LM[TOL]_Sup[ST]) is competitive with its fine-tuning counterpart (LM[Idwiki-TOL]_Sup[ST] and LM[Idwiki-KT]_Sup[ST]). Unsupervised fine-tuning might not be necessary if one already has access to a huge unlabeled in-domain data. With this result, we can also conclude that the LM pretraining (ELMo) induces useful syntactic knowledge, which in this case is part-of-speech information.

Surprisingly, the diagnostic classifier performances badly deteriorate on the development set given inputs from the BiLSTM. All models from this group obtain accuracies below the simple baseline model. This contradicts our previous result where supervised fine-tuning obtains adequate result on small training data. It seems that the BiLSTM does not encode part-of-speech information as good as the pretrained LM, even though it receives additional input from the pretrained LM (LM[Idwiki]_Sup[TO] and LM[KT]_Sup[TO]). We think that the supervised fine-tuning models may learn something other than the part-of-speech information which helps them perform well on the small training data. A plausible explanation would be that BiLSTM is learning NER-specific information during the supervised training on source domain, replacing the part-of-speech information from the pretrained LM.

5 Conclusion

In this paper, we investigate the impact of language model pretraining on named-entity recognition task in Indonesian conversational texts. We use two variants of three step training procedure: supervised fine-tuning (fine-tuning the BiLSTM) and unsupervised fine-tuning (fine-tuning the pretrained LM). Using both approaches, the neural models obtain significant increase from the CNN-BiLSTM-CRF and the multitask baseline on small training data, yielding an absolute improvement of 32 points of test F1 score. However, one might not need to fine-tune if: (1) a large unlabeled in-domain data is already at hand then one can train language model directly without any fine-tuning, and (2) an adequate amount of labeled in-domain data (in our case it's > 5000 sentences) is present then one can opt for simpler models such as the multitask model. Furthermore, we also find that the pretrained LM encodes part-of-speech information, which is a strong predictor for named entity recognition. The neural sequence labeler, on the other hand, seems to encode another information other than part-of-speech to help it perform well on NER task on small training data.

Acknowledgments

We thank the anonymous reviewers for their valuable feedback. We also would like to thank Kemal Kurniawan for reviewing the early version of this work and Pria Purnama for his support.

References

Gustavo Aguilar, Suraj Maharjan, Adrián Pastor López-Monroy, and Thamar Solorio. 2017. A multitask approach for named entity recognition in social media data.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling.

In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Bayu Aryoyudanta, Teguh Bharata Adji, and Indriana Hidayah. 2016. Semi-supervised learning approach for indonesian named entity recognition (ner) using co-training algorithm. *2016 International Seminar on Intelligent Technology and Its Applications (ISITIA)*, pages 7–12.

Terra Blevins, Omer Levy, and Luke S. Zettlemoyer. 2018. Deep rnns encode soft hierarchical syntax. In *ACL*.

Indra Budi, Stéphane Bressan, Gatot Wahyudi, Zainal A. Hasibuan, and Bobby A. A. Nazief. 2005. Named entity recognition for the indonesian language: Combining contextual, morphological and part-of-speech features into a knowledge engineering approach. In Achim Hoffmann, Hiroshi Motoda, and Tobias Scheffer, editors, *Discovery Science*, pages 57–69, Berlin, Heidelberg. Springer Berlin Heidelberg.

Laura Chiticariu, Rajasekar Krishnamurthy, Yunyao Li, Frederick Reiss, and Shivakumar Vaithyanathan. 2010. Domain adaptation of rule-based annotators for named-entity recognition tasks. In *EMNLP*.

James R. Curran and Stephen Clark. 2003. Language independent ner using a maximum entropy tagger. In *CoNLL*.

Hal Daumé. 2007. Frustratingly easy domain adaptation. *CoRR*, abs/0907.1815.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *NAACL-HLT*.

Dieuwke Hupkes, Sara Veldhoen, and Willem H. Zuidema. 2018. Visualisation and 'diagnostic classifiers' reveal how recurrent and recursive neural networks process hierarchical structure. In *IJCAI*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Kemal Kurniawan and Samuel Louvan. 2018. Empirical evaluation of character-based model on neural named-entity recognition in indonesian conversational texts. *CoRR*, abs/1805.12291.

Brenden M. Lake and Marco Baroni. 2017. Still not systematic after all these years: On the compositional skills of sequence-to-sequence recurrent networks. *CoRR*, abs/1711.00350.

Ji Young Lee, Franck Dernoncourt, and Peter Szolovits. 2018. Transfer learning for named-entity recognition with neural networks. *CoRR*, abs/1705.06273.

Rezka Leonandya, Bayu Distiawan, and Nursidik H. Praptono. 2015. A semi-supervised algorithm for indonesian named entity recognition. *2015 3rd International Symposium on Computational and Business Intelligence (ISCBI)*, pages 45–50.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of lstms to learn syntax-sensitive dependencies. *TACL*, 4:521–535.

Andry Luthfi, Bayu Distiawan Trisedya, and Ruli Manurung. 2014. Building an indonesian named entity recognizer using wikipedia and dbpedia. *2014 International Conference on Asian Language Processing (IALP)*, pages 19–22.

Xuezhe Ma and Eduard H. Hovy. 2016. End-to-end sequence labeling via bi-directional lstm-cnns-crf. *CoRR*, abs/1603.01354.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The penn treebank: Annotating predicate argument structure. In *Proceedings of the Workshop on Human Language Technology*, HLT '94, pages 114–119, Stroudsburg, PA, USA. Association for Computational Linguistics.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke S. Zettlemoyer. 2018. Deep contextualized word representations. In *NAACL-HLT*.

Femphy Pisceldo. 2009. Probabilistic part of speech tagging for bahasa indonesia.

Alec Radford. 2018. Improving language understanding by generative pre-training.

Marek Rei. 2017. Semi-supervised multitask learning for sequence labeling. In *ACL*.

Sebastian Ruder and Jeremy Howard. 2018. Universal language model fine-tuning for text classification. In *ACL*.

Nitish Srivastava, Geoffrey E. Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15:1929–1958.

Wiwin Suwarningsih, Iping Supriana, and Ayu Purwarianti. 2014. Imner indonesian medical named entity recognition. *2014 2nd International Conference on Technology, Informatics, Management, Engineering and Environment*, pages 184–188.

Fadillah Z Tala. 2003. A study of stemming effects on information retrieval in bahasa indonesia.

Natanael Taufik, Alfan Farizki Wicaksono, and Mirna Adriani. 2016. Named entity recognition on indonesian microblog messages. *2016 International Conference on Asian Language Processing (IALP)*, pages 358–361.

Sara Veldhoen, Dieuwke Hupkes, and Willem H. Zuidema. 2016. Diagnostic classifiers revealing how neural networks process hierarchical structure. In *CoCo@NIPS*.

Dan Wu, Wee Sun Lee, Nan Ye, and Hai Leong Chieu. 2009. Domain adaptive bootstrapping for named entity recognition. In *EMNLP*.

Xiaodong Yu, Stephen D. Mayhew, Mark Sammons, and Dan Roth. 2018. On the strength of character language models for multilingual named entity recognition. In *EMNLP*.

Long-distance dependencies in continuation grammar

Cara Su-Yi Leong
National University of Singapore
cara@u.nus.edu

Michael Yoshitaka Erlewine
National University of Singapore
mitcho@nus.edu.sg

Abstract

We discuss the treatment of movement and variable binding across finite clause boundaries in the continuation-based grammar of Barker and Shan (2014) and related work. We propose extensions to the theory which make such dependencies compatible with a ban on cross-clausal scope-taking as implemented in Charlow (2014). We demonstrate, however, that this resulting grammar systematically makes incorrect predictions for weak crossover in sentences that combine long-distance movement and variable binding, thus undermining one of the major advantages of continuation-based grammars according to Shan and Barker (2006). We conclude with a critical outlook and a comparison to contemporary LF syntax approaches to scope-taking.

1 Introduction

In a notable application of theoretical computer science principles to natural language grammar, Chris Barker and Chung-Chieh (Ken) Shan have developed categorial grammars which incorporate the notion of *continuations*. In brief, a *continuation* is the computational future of an expression, i.e. the procedures that would then apply to it. (See especially Shan and Barker (2006) §1.2 and citations therein.) Barker and Shan argue that continuations are not only a useful conceptual device for the description of natural language phenomena, but in fact enable a grammatical framework which is in many ways superior to its alternatives. Their continuation-based grammars make positive predictions for phenomena such as superiority, binding, crossover, polarity licensing, donkey anaphora, and reconstruction effects (Barker 2002; Shan 2004, 2007; Shan and Barker 2006; Barker and Shan 2006, 2008, 2014). We refer to these works collectively as B&S.

In this paper, we consider the treatment of examples with embedded clauses in the B&S framework. As has been noted by B&S themselves and Charlow (2014), the B&S framework as unamended overgenerates interpretations for sentences with quantifiers in scope islands — including embedded finite clauses — as it does not inherently impose any restrictions on quantifier scope-taking. We discuss an approach to restricting scope-taking out of scope islands discussed by Charlow (2014), but which in turn complicates examples with long-distance movement and binding. Although these complications can be overcome, the necessary amendments in turn lead to incorrect predictions for crossover effects. We argue that this discussion poses a fundamental challenge to the B&S framework as a model of grammatical behavior when a wider range of data is considered.

We begin in §2 with an introduction to the principles and notation of the B&S grammar as presented in B&S 2014 Part 1. In §3, we discuss scope-taking across clause boundaries and present a restricted theory for long-distance dependencies in the B&S framework. In §4, we discuss the amended theory's predictions for crossover effects. We conclude in §5 with a critical evaluation of the treatment of scope-taking restrictions in the B&S framework, in comparison with LF-based theories for scope-taking and movement.

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 114-122, Hakodate, Japan, September 13-15, 2019

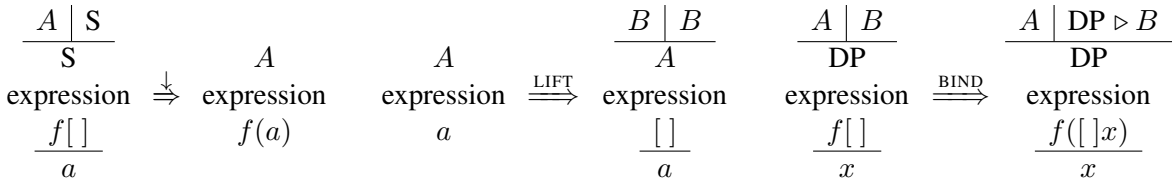


Figure 1: Freely-applying type-shifters in B&S: LOWER (⇓), LIFT, and BIND

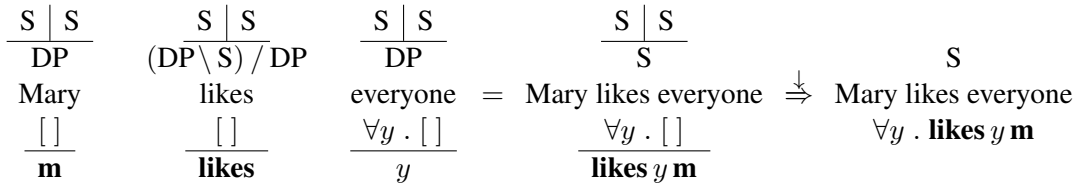


Figure 2: Scope-taking of object *everyone* in *Mary likes everyone*

2 Background: Barker & Shan's continuation-based grammar

We begin by briefly presenting the grammar in Part 1 of B&S (2014) with the notation there.

The B&S grammar is a combinatory categorial grammar which includes continuation-passing expressions. In addition to common $\backslash$ and $/$-type constructors for left and right composition, B&S introduce the $\mathbin{\backslash\!\!\backslash}$ and $\mathbin{/\!\!/}$ constructors for continuation-passing expressions. Informally, $A \mathbin{\backslash\!\!\backslash} B$ "would be a B if we could add an A somewhere (specific) inside of it" whereas $C \mathbin{/\!\!/} D$ "would be a C if we could add a D surrounding it" (B&S 2014: 6).

Syntactic categories are presented above each expression and semantic denotations are presented below. B&S also introduce the notation of "multi-level towers" defined as follows:

(1)
$$\begin{array}{ccc} \dfrac{C \mid B}{A} & & C \mathbin{/\!\!/} (A \mathbin{\backslash\!\!\backslash} B) \\ \text{expression} & := & \text{expression} \\ \dfrac{f[\,]}{a} & & \lambda\kappa \,.\, f[\kappa\, a] \end{array}$$

An expression with the syntactic category $\dfrac{C \mid B}{A}$ behaves internally like an A and takes scope over an expression of category B to produce an expression of category C.

The composition of multi-level towers follows the schema in (2). Composition on the lowest level of the multi-level tower follows the direction of composition of the lowest level of the tower's category (here, the functor f of category $A \backslash B$ taking the left A expression x as its argument), whereas procedures on higher levels compose linearly.

(2)
$$\begin{array}{ccccc} \dfrac{C \mid D}{A} & & \dfrac{D \mid E}{A \backslash B} & & \dfrac{C \mid E}{B} \\ \text{left-exp} & & \text{right-exp} & = & \text{left-exp right-exp} \\ \dfrac{g[\,]}{x} & & \dfrac{h[\,]}{f} & & \dfrac{g(h[\,])}{f(x)} \end{array}$$

In addition, B&S introduce three type-shifters, shown in Figure 1, which apply freely to the relevant expressions. The LOWER type-shifter $\downarrow$ can apply to expressions of type $\dfrac{A \mid \text{S}}{\text{S}}$ for arbitrary A.

2.1 Scope-taking

Continuation-passing through multi-level towers is used to model scope-taking expressions. Scope-taking operations occupy the higher levels of multi-level towers, which are then passed as continuations in composition with their evaluation delayed. This is illustrated in Figure 2. Quantifiers such as *everyone* are two-level towers which introduce a variable on the lower level, together with a corresponding operator on a higher level.

The composition of multi-level towers as in (2) and the definition of LOWER ensures an important result: Content on a higher level of a multi-level tower takes scope over content on the same level to its right and over content on lower levels.

In contrast to quantifiers, non-scope-taking expressions such as *Mary* and *likes* do not inherently have continuation-passing, multi-level denotations.

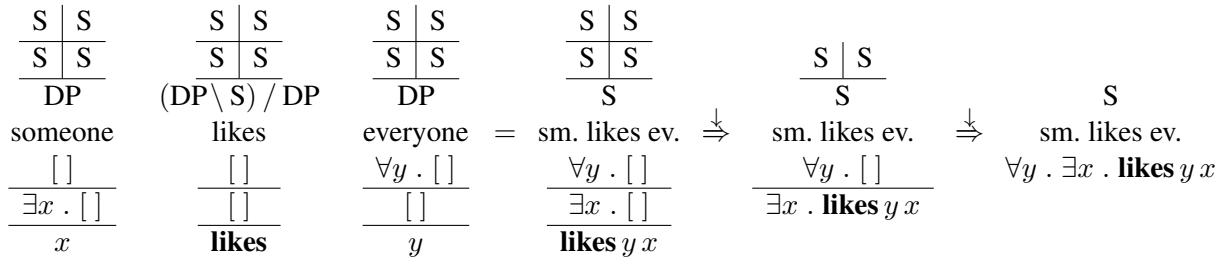


Figure 3: Inverse scope reading of *Someone likes everyone*

The expressions *Mary* and *likes* in Figure 2 are the result of applying the LIFT type-shifter in Figure 1. This was done so that their composition can proceed following the schema in (2), which applies to towers with matching numbers of levels. Notice that LIFT as defined in Figure 1 adds a continuation-passing no-op (identity function) to a top layer of its input. B&S also allow LIFT to apply to sub-parts of denotations, to introduce a no-op intermediate layer. This is called *internal* LIFT.

Now consider the application of these techniques for a sentence with multiple scope-taking expressions. In Figure 3, we model the inverse scope reading of *Someone likes everyone*. The existential *someone* occurs linearly to the left of the universal *everyone*, but the quantificational part of *everyone* has undergone internal LIFT so that it takes scope over an expression that linearly precedes it.

2.2 Pronominal binding

The syntactic category $A \triangleright B$ represents a B that contains an unbound pronoun of category A. A pronoun such as *he* has the denotation in (3): it introduces a variable on the lower level and a corresponding λ binder which will take scope, allowing for binding by an expression to its left.

(3)
$$\frac{\frac{\text{DP} \triangleright \text{S} \mid \text{S}}{\text{DP}}}{\frac{\text{he}}{\frac{\lambda x\ .\ [\,]}{x}}}$$

To bind a pronoun, the BIND type-shifter in Figure 1 is applied to a DP, turning it into an expression that binds a pronoun to its right. The linear nature of binding naturally explains contrasts such as in (4):

(4) a. Every girl$_i$ loves her$_i$ mother.
b. *Her$_i$ mother loves every girl$_i$.

2.3 Movement

Continuation-passing on higher levels of towers also provides an in-situ account of movement dependencies. First, a silent gap of category $A \mathbin{/\!\!/} A$ for some A is placed in the "trace" position of a moved expression. A common choice for a gap will be category $(\text{DP} \mathbin{\backslash\!\!\backslash} \text{S}) \mathbin{/\!\!/} (\text{DP} \mathbin{\backslash\!\!\backslash} \text{S})$ which can be written as a multi-level tower as in (5). Notice that gaps introduce a variable and corresponding λ binder which allows for binding from above, just as pronouns do.

(5)
$$\frac{\frac{\text{DP} \mathbin{\backslash\!\!\backslash} \text{S} \mid \text{S}}{\text{DP}}}{\frac{\overline{\lambda x\ .\ [\,]}}{x}}$$

Second, the FRONT type-shifter (6) is applied specifically to expressions which are in a "moved" position. FRONT ensures that the expression composes with rightward material, i.e. the material that it has "moved over." FRONT also has a secondary effect of requiring the rightward material to be of the form $(A \mathbin{\backslash\!\!\backslash} B)$, which will be important below.

(6) $\dfrac{C \mid B}{A} = C \mathbin{/\!\!/} (A \mathbin{\backslash\!\!\backslash} B) \overset{\text{FRONT}}{\Longrightarrow} C \mathbin{/} (A \mathbin{\backslash\!\!\backslash} B)$

2.4 Crossover effects

A hallmark of the B&S framework is its explanation of "crossover" effects (Postal, 1971), such as in (7). The linear nature of continuation-passing and evaluation, together with the framework's treatment of pronominal binding and movement, leads to a natural explanation for such asymmetries (Shan and Barker, 2006).

(7) a. Which girl$_i$ did you introduce ___ to her$_i$ second cousin?
b.??Which girl$_i$ did you introduce her$_i$ second cousin to ___?

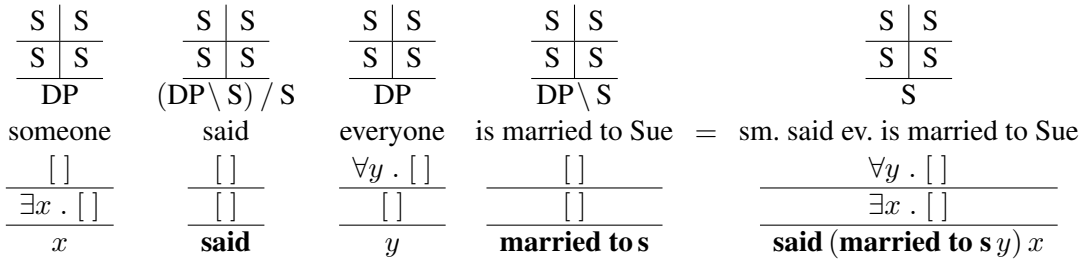


Figure 4: B&S overgenerates the unattested inverse-scope reading of *Someone said everyone is married to Sue* (10)

First, consider the grammatical example in (7a). The gap in (7a) linearly precedes the pronoun *her* and is therefore able to bind it. This computation is sketched in (8). Applying BIND (Figure 1) to the gap in (5) yields (8) below.

(8) $\dfrac{\dfrac{\text{DP} \setminus\!\!\setminus \text{S} \mid \text{S}}{\text{DP}}}{\dfrac{\dfrac{\_\_}{\lambda x\ .\ [\,]}}{x}} \overset{\text{BIND}}{\Longrightarrow} \dfrac{\dfrac{\text{DP} \setminus\!\!\setminus \text{S} \mid \text{DP} \triangleright \text{S}}{\text{DP}}}{\dfrac{\dfrac{\_\_}{\lambda x\ .\ [\,]x}}{x}}$

Using this gap (8) in the gapped clause serves to bind the pronoun to its right, as in (9). A FRONT-ed constituent combines with the resulting gapped clause and simultaneously binds the gap and pronoun.

(9) $\dfrac{\text{DP} \setminus\!\!\setminus \text{S} \mid \text{DP} \triangleright \text{S}}{\text{DP}}$ ... $\dfrac{\dfrac{\_\_}{\lambda x\ .\ [\,]x}}{x}$ $\dfrac{\text{DP} \triangleright \text{S} \mid \text{S}}{\text{DP}}$ ... her $\dfrac{\lambda y\ .\ [\,]}{y}$ $\dfrac{\text{DP} \setminus\!\!\setminus \text{S} \mid \text{S}}{\text{S}}$ ... = ... __ ... her ... $\dfrac{\lambda x\ .\ [\,]}{\ldots x \ldots x \ldots}$

In contrast, the gap in (7b) is to the right of the pronoun and therefore has no opportunity to bind the pronoun. Depending on how the pronoun and gap are LIFT-ed, the gapped clause in (7b) will either have a type with DP▷DP\\S on a single higher level, or have DP ▷ S and DP \\ S on two different higher levels. If the gapped constituent has type DP▷DP\\S on a higher level, or DP ▷ S above DP \\ S, it cannot directly combine with a FRONT-ed constituent since FRONT-ed constituents can only combine with expressions of type $A \setminus\!\!\setminus B$. Thus, the gapped clause must have its pronoun bound first by another DP before its gap can be filled. Finally, if DP \\ S is above DP ▷ S, the gapped clause can combine with the FRONT-ed constituent but cannot simultaneously bind the pronoun, requiring the pronoun to be free or bound from further to the left.

3 Scope-taking across clause boundaries

With this background in place, we now turn to the treatment of complex examples with embedded clauses in the B&S continuation grammar.

An important property of the B&S framework reviewed above is that the scope-taking potential of any expression is unbounded. As is widely known, however, many quantifiers, including universals, are unable to take scope out of finite clauses for most speakers.[1] For example, example (10) modified from Fox (2000, p. 62) is judged by most speakers to be infelicitous, as only its anomalous surface scope reading is available. See the recent discussion in Wurmbrand (2018), as well as references there.

(10) #Someone said [everyone is married to Sue].

The current system incorrectly predicts the inverse scope reading of (10) to be available, as in Figure 4. A modification must be made to the B&S framework to restrict the scope-taking of quantifiers.

In recognition of this problem, and building on discussion in B&S (2008: 27–28), Charlow (2014: 64–66, 90) suggests that all finite clauses — and all scope islands, more generally — *must be evaluated*, i.e. by "collapsing it into a single level" (p. 65).[2] We codify this requirement in (11):

(11) **Scope Island Evaluation:**
If the expression is a scope island, apply LOWER as many times as possible (↓*).[3]

[1] In contrast, indefinite quantifiers are known to be insensitive to a wide range of constraints on quantifier scope; see e.g. Fodor and Sag (1982), Abusch (1994). We concentrate on the scope-taking potential of universal quantifiers here.

[2] The discussion in Charlow (2014) builds on the B&S tradition but presents a distinct grammatical framework based on the notion of monads. In the interest of space, here we only evaluate the B&S framework with the added restriction in (11) and further refinements presented in this section, and leave the full consideration of the Charlow framework for future work.

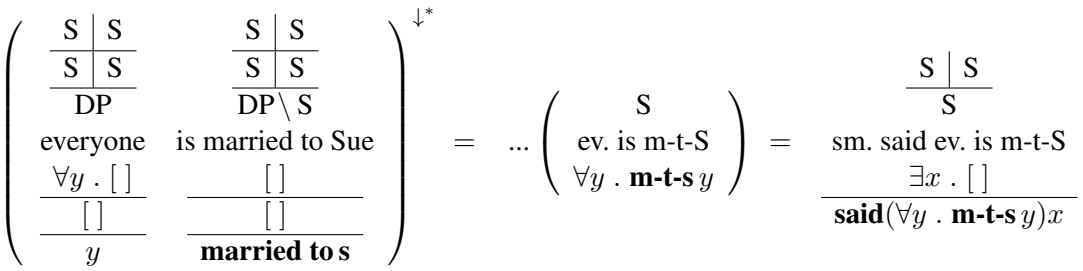


Figure 5: Scope Island Evaluation blocks *everyone* from scoping out of the embedded clause in (10)

Here we concentrate on the behavior of embedded finite clauses as scope islands. The principle of Scope Island Evaluation forces a full LOWER of the embedded clause in (10), successfully blocking *everyone* from scoping above the matrix quantifier, as illustrated in Figure 5.

However, recall that pronouns and movement gaps "take scope" in the B&S framework just as quantificational expressions do, placing a λ binder on the higher level of towers to allow for their binding from the left. In contrast to quantifiers, pronouns and gaps must be able to take scope out of embedded finite clauses, as evidenced by the availability of movement and variable binding across the bracketed clause boundaries in (12):

(12) a. Which girl did you say [Mary saw ___]?
b. Every girl$_i$ said [Mary saw her$_i$].

Scope Island Evaluation (11) thus interrupts the interpretation of embedded pronouns and gaps, incorrectly predicting the ungrammaticality of examples such as (12) without further refinement. We propose such refinements here, first discussing long-distance movement dependencies in §3.1 and then turning to long-distance variable binding in §3.2. This allows us to adopt Scope Island Evaluation as a principle to accurately limit quantifier scope, while also maintaining the availability of long-distance movement and variable binding as in (12).

3.1 Movement with intermediate gaps

Gaps in B&S introduce a λ binder on a higher level and a corresponding variable below (5). Clauses with DP gaps that undergo Scope Island Evaluation will be of category DP$\backslash$S — precisely the category of a clause that is missing a DP to its left. In short-distance movement, the fully LOWER-ed gapped clause is adjacent to the moved item and can immediately combine with it; however, in long-distance movement as in (12a), the moved item cannot immediately compose with the gapped clause.

For example, the embedded clause in (12a) will be as in (13) after Scope Island Evaluation:

(13) Mary saw ___ / λx . [ ] / **saw** x **m** ↓* = DP$\backslash$S: Mary saw ___ / λx . **saw** x **m**

There are two problems with this denotation in (13). First, its type is DP$\backslash$S (see footnote 3), rather than the expected S type for embedded clauses, and therefore will not be able to compose with the standard S-selecting denotation for *say*. Second, since the λ binder for the gap is on the lowest level of the tower, it ceases to propagate as a scope-taking expression. Just as it correctly blocks quantifier scope-taking out of embedded clauses, Scope Island Evaluation incorrectly blocks movement dependencies across embedded clause boundaries.

We propose to resolve this problem with the use of additional, *intermediate gaps*. We first LIFT the fully LOWER-ed embedded clause in (13) into a two-level tower, which can then combine with an immediately preceding gap to yield its original denotation prior to Scope Island Evaluation:

(14) DP ⧹ S | S / DP: ___ / λy . [ ] / y ; S | S / DP$\backslash$S: Mary saw ___ / [ ] / λx . **saw** x **m** = DP ⧹ S | S / S: _ Mary saw _ / λy . [ ] / **saw** y **m**

The structure in (14) can now compose with the embedding verb and further material, passing the λ

[3] In addition, if the category of a one-story tower expression is $A \backslash\!\!\backslash B$, we optionally shift its category to $A \backslash B$ with no change to its denotation.

binder for the inner gap further to the left, to later be successfully saturated by the "moved" element.

Interestingly, decades of work in the derivational syntactic tradition has argued for the presence of intermediate gaps at clause edges in cases of long-distance movement, as a reflex of *successive cyclic movement* (Chomsky, 1977). We review one empirical argument for such intermediate gaps here.

Reflexive pronouns in English must be bound by a local antecedent, leading to the ungrammaticality of (15). However, the reflexive *herself* can be successfully bound in (16), modified from Barss (1986: 25):

(15) *Keely$_i$ thinks [Ted likes a picture of herself$_i$].

(16) Which picture of herself$_i$ does Keely$_i$ think [Ted likes ___]?

In general, *wh*-moved constituents can be bound in their gap position, which B&S accomplish through delayed evaluation of "moved" material. But the grammaticality of (16) teaches us more: *herself* cannot be successfully bound if evaluated in the observable gap position in (16), but can be bound if we postulate an intermediate gap as in (17):

(17) Which picture of herself$_i$ does Keely$_i$ think ___ [Ted likes ___]?

In B&S's representational theory of movement dependencies, then, the adoption of Scope Island Evaluation offers independent motivation for the presence of intermediate gaps as in (17), serving to explain facts such as the grammaticality of (16).

3.2 Long-distance binding with PROLIFT

The Scope Island Evaluation requirement poses a similar problem for embedded pronouns. Pronouns, like gaps, introduce a λ binder on a higher level and a corresponding variable below; enforcing Scope Island Evaluation on the embedded clause in (12b) containing an unbound pronoun fixes the scope of that pronoun by placing it on the bottom level of the tower, as in (18).

(18)
$$\left(\begin{array}{c}\dfrac{\text{DP} \triangleright \text{S} \mid \text{S}}{\text{S}} \\ \text{Mary saw her} \\ \dfrac{\lambda x\ .\ [\,]}{\mathbf{saw}\ x\ \mathbf{m}}\end{array}\right)^{\downarrow *} = \begin{array}{c}\text{DP} \triangleright \text{S} \\ \text{Mary saw her} \\ \lambda x\ .\ \mathbf{saw}\ x\ \mathbf{m}\end{array}$$

The resulting denotation in (18) could by itself be an utterance with a free variable, i.e. an open proposition, but it is not an appropriate denotation for an embedded clause. Just as with embedded clauses containing gaps in §3.1, the resulting object of category DP ▷ S cannot combine with an S-selecting verb such as *say*. But unlike in §3.1, we are unable to resolve this problem by adding an intermediate gap. There is no gap that will compose with (18) and serve to extend its scope.

More generally, there are *no* existing expressions in the B&S framework that can combine with an expression of category DP▷S. To see this, we first note that except for pronouns which introduce them, lexical items never make reference to ▷-categories. The only way that a ▷-containing category is bound is by a DP which has undergone BIND, which have a category of the form $\dfrac{A \mid \text{DP} \triangleright B}{\text{DP}}$. The ▷-containing category (in this case, DP ▷ S) is thus required to be on a higher level of the tower, rather than the bottom.

We thus propose a new type-shifter PROLIFT (19) that lifts the pronoun out of the bottom layer of a tower. Semantically, the variable in g is locally saturated and then abstracted over on a higher level.

(19)
$$\begin{array}{c}\dfrac{B \mid C}{\text{DP} \triangleright A} \\ \text{expression} \\ \dfrac{f[\,]}{\lambda x\ .\ g(x)}\end{array} \xrightarrow{\text{PROLIFT}} \begin{array}{c}\dfrac{\text{DP} \triangleright B \mid C}{A} \\ \text{expression} \\ \dfrac{\lambda x\ .\ f[\,]}{g(x)}\end{array}$$

Using PROLIFT returns the pronoun's λ binder and its corresponding DP▷-category to a higher level, from which position it can propagate to the left until it is bound. The adoption of a principle such as Scope Island Evaluation, motivated by observed limitations on quantifier scope, together with the grammatical possibility of binding embedded pronouns as in (12b), makes the PROLIFT a necessary addition to the grammar.

4 Crossover in long-distance configurations

We have seen that the adoption of Scope Island Evaluation (11), intermediate gaps, and PROLIFT together resolve a number of limitations of B&S's original framework with regards to the behavior of embedded clauses. However, we now demonstrate that the combination of these three improvements together lead to incorrect predictions for crossover effects with gaps and pronouns in embedded clauses:

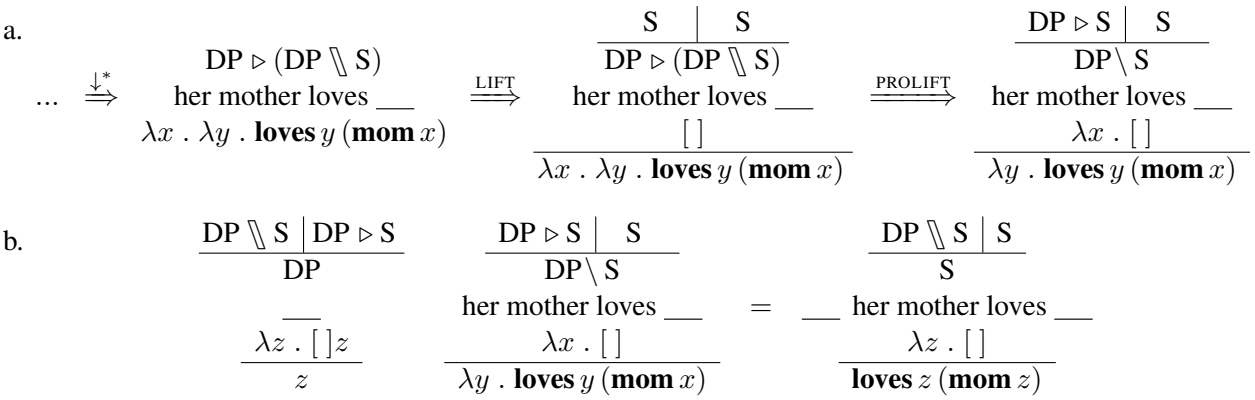


Figure 6: The revised B&S grammar overgenerates, undoing the weak crossover violation in (20b).

(20) a. Which girl$_i$ do you think
[___ loves her$_i$ mother]?
b.?? Which girl$_i$ do you think
[her$_i$ mother loves ___]?

In particular, our revised B&S framework predicts the crossover violation configuration in (20b) to be grammatical.

Let us see how our revised B&S framework from §3 makes such a prediction. One option for the interpretation of the embedded clause *her mother loves* ___ in (20b) yields an expression of category DP ▷ (DP \ S) after Scope Island Evaluation. In Figure 6a, we apply LIFT and PROLIFT to this structure.[4] Then in Figure 6b, we introduce the intermediate gap which has undergone BIND (8), also used in the grammatical (9) above. This BIND-shifted gap (8) serves to introduce a λ binder on the higher level for the gap and also simultaneously bind the pronoun. This is possible despite the fact that the pronoun precedes the gap position in the embedded clause. This demonstration shows that the B&S framework as amended here in §3 now overgenerates, predicting structures of the form in (20b) to be grammatical, contrary to fact.

Before applying PROLIFT and combining with the intermediate gap, the pronoun in the gapped clause needed to be bound before the gap could be filled. However, PROLIFT gives its raised λ operator widest scope as shown in Figure 6a, which generates a configuration in which the pronoun and gap are located on different levels. When the BIND-shifted intermediate gap is introduced as in Figure 6b, then, the pronoun can be bound at the same time as the gap is filled, since these operations occur on different levels. The upshot is that applying a combination of Scope Island Evaluation, intermediate gaps and PROLIFT on a long-distance crossover configuration predicts that (20b) is acceptable.

We appear to be at an impasse. The ingredients that together lead to the overgeneration of (20b) were each independently motivated. Scope Island Evaluation (11) is necessary to account for the observed limitations on quantifier scope-taking (B&S, 2008; Charlow, 2014). Without it, quantifiers would be able to take scope out of scope islands, including embedded finite clauses. Meanwhile, intermediate gaps and PROLIFT are minimal alternations to the theory to make long-distance movement and binding dependencies compatible with Scope Island Evaluation. The intermediate gap introduced in Figure 6b is the same simple gap shifted with BIND (8) used by B&S to account for grammatical configurations of variable binding by movement (Shan and Barker, 2006). Given that each of these steps cannot be omitted from the theory, it appears that the B&S framework systematically overgenerates in a way that undermines this crucial argument for the approach: the explanation of crossover effects.

5 Discussion

The Barker and Shan grammar fragment is notable both technically, for being built on the theoretical notion of continuations as a model for scope-taking, and empirically, for its ambitious consideration and

[4] In the final step, we also shift the lower level type from DP \\ S to DP \ S, following footnote 3, as the lowest levels of towers are not naturally continuation-passing.

treatment of a wide range of challenging syntactic and semantic phenomena. One of the first and most prominent arguments presented for the framework is its account for binding and crossover effects (Shan and Barker, 2006). The apparently linear nature of crossover effects receives a natural explanation in the B&S framework, where scope-taking expressions compose linearly.

As has been recognized before, the B&S framework as presented does not by itself adequately restrict the scope-taking potential of quantifiers. Here we codified a suggestion made in Charlow (2014) following B&S (2008), that scope islands such as embedded finite clauses are recursively lowered so that scope-taking operations cannot take scope further: Scope Island Evaluation (11). We then presented minimal improvements to the theory to maintain the availability of long-distance movement and binding configurations, but ultimately showed in §4 that these tools lead to a fatal overgeneration.

Perhaps the greatest limitation of the B&S framework that this work highlights is its uniform treatment of pronouns, gaps, and quantifiers as scope-takers.[5] The ultimately problematic amendments we proposed in §3 were needed because the scope-taking of quantifiers, but not movement or bound pronoun relationships, is blocked by intervening finite clause boundaries. In fact, these dependencies can be further distinguished: *islands* such as relative clauses block movement dependencies across them (21a) (Ross, 1967), but do not block pronominal binding (21b):

(21) a. *Who$_i$ did you say [(that) Sarah ate [$_{\text{island}}$ the food that ___$_i$ made]]?

b. Who$_i$ did you say [___$_i$ ate [$_{\text{island}}$ the food that she$_i$ (herself) made]]?

In contrast, many common LF-based approaches to semantic interpretation — such as that introduced in Heim and Kratzer (1998) — utilize a fundamentally different mechanism for pronominal binding that is insensitive to intervening syntactic boundaries.[6] These theories do however relate quantificational scope-taking and movement, hypothesizing "covert" movement to account for quantifier scope — also known as *Quantifier Raising* (May, 1977).

At first glance, this may suggest that LF-based accounts will face similar difficulties in distinguishing between configurations of licit movement dependencies and licit quantifier scope-taking. However, the LF theorist will note that different forms of "overt" movement are already sensitive to different locality restrictions: for example, between A-movement, $\overline{\text{A}}$-movement, and head-movement. (See Rizzi (2013), Belletti (2018), for recent overviews and approaches.) The clause-boundedness of quantifier scope-taking then reduces to an existing and independently necessary task of developing syntactic theories to explain the different locality profiles of different types of movement. We hope that such work will, in the future, lead to principled accounts, beyond simply stipulating that QR cannot take place out of scope islands. See Wurmbrand (2018) for some discussion of recent approaches.

Of the Scope Island Evaluation requirement critiqued here, Charlow (2014) writes: "The requirement that scope islands must be evaluated is intended as nothing more and nothing less than the denotational correlate of prohibiting QR out of scope islands" (p. 90). We disagree with this characterization. The effects of adopting Scope Island Evaluation are far greater than simply limiting QR out of scope islands, leading to serious overgeneration if we are to maintain the B&S framework's existing explanations for binding and crossover facts.

Finally, we note that our discussion has followed B&S in taking crossover effects to be an issue of grammatical competence. As a reviewer notes, it is possible that crossover effects, as well as other locality effects such as restrictions on scope-taking and island effects on movement, may instead be due to considerations of on-line processing.

Acknowledgments

We thank Chris Barker for helpful correspondence and encouraging discussion.

[5] Dowty (2007, sec. 2.8) refers to this approach to pronominal binding as a "combinatory" approach, primarily discussing Jacobson (1999), in contrast to "free variable binding" theories such as LF-based theories, below. As an anonymous reviewer notes, there are other frameworks in the CCG tradition which do not assume such a unification, such as Steedman (2011).

[6] In turn, LF-based theories have been challenged for requiring c-command configurations for binding, rather than linear precedence; see Barker (2012).

References

Dorit Abusch. 1994. The scope of indefinites. *Natural Language Semantics*, 2:83–136.

Chris Barker and Chung-Chieh Shan. 2006. Types as Graphs: Continuations in Type Logical Grammar. *Journal of Logic, Language and Information*, 15(4):331–370.

Chris Barker and Chung-Chieh Shan. 2008. Donkey anaphora is in-scope binding. *Semantics and Pragmatics*, 1(1):1–46.

Chris Barker and Chung-Chieh Shan. 2014. *Continuations and Natural Language*. Oxford University Press.

Chris Barker. 2002. Continuations and the Nature of Quantification. *Natural Language Semantics*, 10:211–242.

Chris Barker. 2012. Quantificational Binding Does Not Require C-Command. *Linguistic Inquiry*, 43(4):614–633, October.

Andrew Barss. 1986. *Chains and anaphoric dependence*. Ph.D. thesis, Massachusetts Institute of Technology.

Adriana Belletti. 2018. Locality in syntax. In *Oxford Research Encyclopedia of Linguistics*. Oxford University Press.

Simon Charlow. 2014. *On the Semantics of Exceptional Scope*. Doctoral Dissertation, New York University.

Noam Chomsky. 1977. On *wh*-movement. In Peter Culicover, Thomas Wasow, and Adrian Akmajian, editors, *Formal Syntax*, pages 71–132. New York: Academic Press.

David Dowty. 2007. Compositionality as an empirical problem. In Chris Barker and Pauline Jacobson, editors, *Direct Compositionality*, volume 14 of *Oxford Studies in Theoretical Linguistics*, pages 23–101. Oxford University Press, Oxford.

Janet Dean Fodor and Ivan A. Sag. 1982. Referential and quantificational indefinites. *Linguistics and Philosophy*, 5(3):355–398.

Danny Fox. 2000. *Economy and semantic interpretation*. MIT Press.

Irene Heim and Angelika Kratzer. 1998. *Semantics in Generative Grammar*. Blackwell Textbooks in Linguistics. Wiley.

Pauline Jacobson. 1999. Towards a Variable-Free Semantics. *Linguistics and Philosophy*, 22(2):117–184.

Robert Carlen May. 1977. *The Grammar of Quantification*. Doctoral Dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Paul M. Postal. 1971. *Cross-over Phenomena*. Holt, Rinehart and Winston.

Luigi Rizzi. 2013. Locality. *Lingua*, 130:169–186.

John Robert Ross. 1967. *Constraints on Variables in Syntax*. Doctoral Dissertation, Massachusetts Institute of Technology, Cambridge, MA.

Chung-Chieh Shan and Chris Barker. 2006. Explaining Crossover and Superiority as Left-to-right Evaluation. *Linguistics and Philosophy*, 29(1):91–134.

Chung-Chieh Shan. 2004. Delimited continuations in natural language. In *Proceedings of the Fourth ACM SIGPLAN Continuations Workshop*.

Chung-Chieh Shan. 2007. Linguistic side effects. In Chris Barker and Pauline Jacobson, editors, *Direct Compositionality*, pages 132–163. Oxford University Press, Oxford.

Mark Steedman. 2011. *Taking scope: The natural semantics of quantifiers*. MIT Press.

Susanne Wurmbrand. 2018. The cost of raising quantifiers. *Glossa*, 3(19):1–40.

On Null Clausal Complements in Taiwan Southern Min

Huei-Ling Lin
Department of Foreign Languages and Literature
National Chung Cheng University
168, Sec. 1, University Rd., Minhsiung, Chiayi 62102, Taiwan
folhll@ccu.edu.tw

Abstract

This paper aims to investigate the less discussed null argument—null clausal complement in Taiwan Southern Min (TSM). The discussion issues include the derivation, status, and replacement of null clausal complements in TSM. This paper proposes that being either syntactically or pragmatically controlled, the null clausal complement in TSM is a type of deep anaphora, which is not derived through deletion. Moreover, possessing features such as not being A-bound, and possibly being but not required to be Ā-bound, the null clausal complement in TSM is argued to have the status of an epithet. While null clausal complements are not allowed with all kinds of verbs, in some cases where null clausal complements are prohibited, an obligatory pro-S *an-ne* 'so' is then required.

1 Introduction

Null arguments are common among languages. In the literature discussing null arguments, null objects are often the topic of discussion. As demonstrated in (1-3), Mandarin Chinese, Japanese, and Korean all allow null objects.

(1) Zhangsan bu xihuan [guanyü ziji-de yaoyan];
Zhangsan not like [about self-Gen rumor
Mali ye bu xihuan [$_{NP}$ e]. (Mandarin Chinese)
Mary also not like (Kim 1999)
'Zhangsan doesn't like rumors about himself, and Mary doesn't, either.'
a. Mary does not like rumors about herself, either
b. Mary does not like rumors about Zhangsan, either

(2) John-wa [zibun-no tegami]-o sute-ta;
John-Top [self-Gen letter-Acc discard-Perf;
Mary-mo [$_{NP}$ e] sute-ta. (Japanese)
Mary-also discard-Perf (Kim 1999)
'John threw out his letters, and Mary did too.'
a. Mary threw out her (= Mary's) letters, too
b. Mary threw out his (= John's) letters, too

(3) a. Jerry-nun [caki-uy ai]-lul
Jerry-Top self-Gen child-Acc
phal-ul ttayli-ess-ta. (Korean)
arm-Acc hit-Past-Ind (Kim 1999)
'Jerry hit his child on the arm.'
b. Kulena Sally-nun [$_{NP}$ e] tali-lul ttayli-ess-ta.
But Sally-Top leg-Acc hit-Past-Ind
i) But Sally hit her (= Sally's) child on the leg
ii) But Sally hit his (= Jerry's) child on the leg

In addition to NP objects, clauses are often subcategorized for by verbs. However, null clausal complements are much less discussed in the literature. This paper aims to discuss the derivation, status, and replacement of null clausal complements in Taiwan Southern Min (TSM), a Chinese language spoken by more than 80% of people in Taiwan (Cheng 1985). To illustrate, as shown in (4), the verb *tsai-iann* 'know' is subcategorized for a clausal complement, which is spelled out as *sin-long pai-kha* 'the bridegroom is crippled' in the first half, but in the second half of the sentence the clausal complement is null, marked as [$_{CP}$ e]. This paper discusses how the null clausal complement is derived, what its status is, and whether it can be replaced by other syntactic elements.

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 123-128, Hakodate, Japan, September 13-15, 2019

(4) sin-niu tsai-iann sin-long pai-kha,
bride know groom crippled
mue-lang-po ma tsai-iann [$_{CP}$ e].[1]
matchmaker also know
'The bride knows that the bridegroom is crippled, and the matchmaker also knows.'

2 Literature Review

Hankamer and Sag (1976) have specified two anaphoric processes: surface anaphora, which results from "deletion under identity with antecedent forms", and deep anaphora, which is not derived via deletion and allows pragmatic control. They have argued that null complement anaphora demonstrates no sign of syntactic deletion and thus should be taken as deep anaphora, which can be either syntactically controlled or pragmatically controlled. To illustrate, the omitted clausal complements in (5-6) should be taken as deep anaphora. In (5) the antecedent is syntactically controlled, while that in (6) is pragmatically determined.

(5) We needed somebody to carry the oats down to the bin, but nobody volunteered.
(Hankamer and Sag 1976)
(6) [Indulgent father feeds baby chocolate bar for dinner]
Mother: I don't approve. (Hankamer and Sag 1976)

More recently, some scholars such as Huang (1991) and Saito (2007) take sloppy reading as evidence of deletion. To illustrate, a Mandarin Chinese example such as (7) seems to involve a missing object. However, Huang (1991) argues that as its English counterpart (8) shows, (7) actually involves VP-ellipsis for the reason that both (7) and (8) are ambiguous with strict and sloppy readings. That is, (7) and (8) have both the strict reading that John saw John's mother and Mary also saw John's mother and the sloppy reading that John saw John's mother and Mary saw Mary's mother. Since strict/sloppy ambiguity is typical of constructions involving VP-ellipsis, both (7) and (8) are argued to involve VP-ellipsis.

(7) John kanjian-le tade mama, Mary ye kanjian-le.
John see-PERF his mother Mary also see-PERF
'John saw his mother, and Mary did, too.'
(Huang 1991)
(8) John saw his mother, and Mary did [$_{VP}$ e], too.
(Huang 1991)

However, Hoji (1998, 2003) and Kasai (2014) argue that deep anaphora may demonstrate sloppy reading as well. For instance, the null object as in (9) is allowed when no linguistic antecedent is available. This is often a case of deep anaphora (Hankamer and Sag 1976).

(9) Bill-ga *e* tataita. (Kasai 2014)
Bill-NOM hit
'Bill hit *e*.'

In (10) the null argument as an empty pronoun without a linguistic antecedent allows sloppy reading. That is, (10) could be interpreted as Hanako hits his arm or Hanako hits her arm.

(10) [Watching a boy hitting his arm] (Kasai 2014)
Taroo: Hanako-mo *e* yoku tataiteru yo.
Hanako-also often hit PARTICLE
'Hanako also often hits *e*.'

Likewise, null clausal complement is allowed when no linguistic antecedent is available as in (11), where the null clausal complement refers to Mary's flirting with someone else. The null clausal complement thus should be taken as deep anaphora.

(11) [John suspects that Mary, who is his girlfriend, flirts with someone else. John and his friend happen to watch Mary's flirting with someone else.
John: Zituwa *pro* mae-kara *e*
in-fact before-from
omottetanda yonaa. (Kasai 2014)
thought-be PARTICLE
'In fact, I have long thought *e*.'

Even for example (12), which involves sloppy reading, the null clausal complement is also argued to be a pro.

[1] The romanization used in this paper for Taiwan Southern Min examples is according to the Taiwan Southern Min Romanization Proposal（臺灣閩南語羅馬字拼音符號方案）, which was promulgated by the Ministry of Education in Taiwan in 2006.

(12) Hanako-wa [$_{CP}$ zibun-no teian-ga
TOP self-GEN proposal-NOM
saiyoosareru to] omotte iru ga,
accepted-be that think though
Taroo-wa _____ omotte inai
TOP think not (Saito 2007)
'Hanako thinks that her proposal will be accepted, but Taroo does not think that her/his proposal will be accepted.'

3 The Proposal

3.1 Derivation

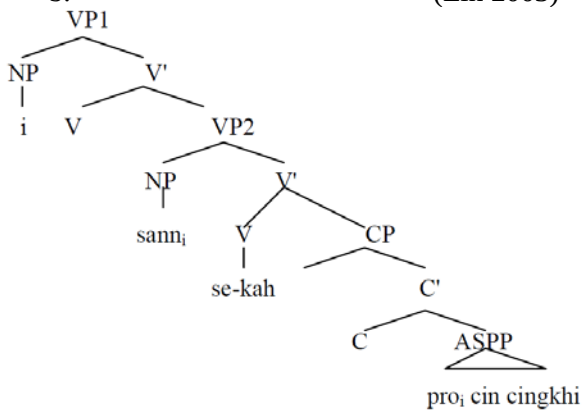
Null clausal complements in TSM can be either syntactically or pragmatically controlled. Therefore, following Hankamer and Sag's (1976) proposal, this paper argues that null clausal complements in TSM are deep anaphora, which are either syntactically or pragmatically controlled. To illustrate, in (4) the antecedent of the null clausal complement can be identified to be *sin-long pai-kha* 'the bridegroom is crippled' in the previous clause; that is, the null clausal complement is syntactically controlled in (4).

Clauses can serve as complements after predicative verbs such as *tsai-iann* 'know' in (4); in addition, they can be complements subcategorized for by V-*kah* in TSM as in (13).

(13) Ong-e huan-lo-kah long be tsiah,
Ong-e worry-kah all not eat
Li-e m huan-lo-kah [$_{CP}$ e].
Li-e also worry-kah
'Ong was so worried that he couldn't eat anything, and Li was also so worried.'
(syntactically controlled or pragmatically controlled)

Before the null clausal complement in (13) is discussed, a few words on *kah*-constructions in TSM are in order. As discussed in Lin (2003), V-*kah* can take three types of secondary predicate, expressing result, state, or extent. Example (14a) involves a resultative *kah*-construction, where the clause after *kah*, *cin cingkhi* 'very clean', expresses the result of the event *i se sann* 'he washed clothes'. The resultative *kah*-clause as argued for by Lin (2003) is a clausal complement subcategorized for by V-*kah* with a structure as in (14b).

(14) a. i ciong sann se-kah cin cingkhi.
he CIONG clothes wash-KAH very clean
'He washed his clothes clean.' (Lin 2003)
b. (Lin 2003)



In the second half of (13), the missing element after *kah* can be understood to be syntactically controlled as the case in (4); that is, the null clausal complement is understood to refer to *long be tsiah* 'cannot eat' in the first half of (13). It can also be construed as being pragmatically controlled, and the missing element is understood to be something similar to the second half in (15), *be khun tsit* 'cannot sleep'.

(15) Ong-e huan-lo-kah long be tsiah,
Ong-e worry-kah all not eat
Li-e m huan-lo-kah be khun tsit.
Li-e also worry-kah not sleep can
'Ong was so worried that he couldn't eat anything, and Li was also so worried that he couldn't sleep.'

As to example (16), the null clausal complement can only be pragmatically controlled, and its antecedent is understood through the context. A possible antecedent could be something like the second half of (17), *be kui-a king tshu* 'by several houses'.

(16) Ong-e tso-sing-li than-tsinn than-kah [$_{CP}$ e].
Ong-e do-business make-money make-kah
'Ong made so much money out of doing business.' (pragmatically controlled)
(17) Ong-e tso-sing-li than-tsinn than-kah
Ong-e do-business make-money make-kah
be kui-a king tshu.
buy several CL house
'Ong made so much money out of doing

business that he bought several houses.'

It should be noted that resultative *kah*-constructions in TSM can be further classified into subject-oriented and object oriented. In a subject-oriented resultative *kah*-construction, the pro in the resultative clause is predicated of the subject of the main clause. To illustrate, in (13) the resultative clause *long be tsiah* 'cannot eat' is predicated of the subject Ong-e. On the other hand, in an object-oriented resultative *kah*-construction, the resultative clause is predicated of the object of the main clause. For instance, in (14), the resultative clause *cin cingkhi* 'very clean' is predicated of the object *sann* 'clothes'. The two types of resultative *kah*-construction still differ in that only the subject-oriented ones allow null clausal complements as in (13) and (16); object-oriented ones do not as the ungrammaticality of (18) shows.

(18) *i ciong sann se-kah. (cf. (14a))
he CIONG clothes wash-KAH
'He washed his clothes as a result...'

Discussing anaphora types, Kasai (2014) argues that deep anaphora, which does not involve deletion, may involve sloppy reading. Sloppy reading can also be identified in TSM examples such as (19), where the second half has the sloppy reading in which Li knows that Li's plan is good as well as the strict reading in which Li knows that Ong's plan is good. Sloppy reading thus does not argue against the deep anaphora analysis proposed in this paper.

(19) Ong-e tsai-iann kati-e ke-ue tsin ho,
Ong-e know self-GEN plan very good
Li-e m tsai-iann [$_{CP}$ e].
Li-e also know
i. 'Ong knows that his (= Ong's) plan is good, and Li also knows that his (= Ong's) plan is good.' (Strict Reading)
ii. 'Ong knows that his (= Ong's) plan is good, and Li also knows that his (= Li's) plan is good.' (Sloppy Reading)

3.2 Status

This paper argues that the null clausal complement in TSM is a null epithet as it has the four properties of an epithet as mentioned in Huang (1991): (a) it may not be A-bound, (b) it may be Ā-bound, (c) it need not be Ā-bound, and (d) it may be coindexed with an argument as long as the argument does not c-command it (pp. 61-62). To illustrate, the null clausal complement may not be A-bound as shown in (20); the antecedent of the null clausal complement cannot be in an argument position, such as the subject position.

(20) *tse sin-long pai-kha ma tsai-iann [$_{CP}$ e].
this groom crippled also know
intended meaning: '*That this bridegroom is crippled also knows (that this bridegroom is crippled). (cf. (4))

As shown in (21), the null clausal complement may be Ā-bound, that is, referring to the topic.

(21) tse sin-long pai-kha mue-lang-po
this groom crippled matchmaker
tsai-iann [$_{CP}$ e] a. (cf. (4))
know PRT
'As to the fact that this bridegroom is crippled, the matchmaker knows.'

However, it is not necessary for the null clausal complement to be Ā-bound as in (22),

(22) mue-lang-po tsai-iann [$_{CP}$ e]. (cf. (4))
matchmaker know
'The matchmaker knows.'

Furthermore, it may be coindexed with an argument as long as the argument does not c-command it as in (13), where the antecedent occurs in the first half and does not c-command the null clausal complement.

3.3 Replacement

As pointed out by Kennedy and Merchant (2000), not all verbs allow null clausal complements as illustrated in (23-24).

(23) The missile test had failed, but only the brass knew. (Kennedy and Merchant 2000)
(24) *The missile test had failed, but only Prof. Hicks {said / thought / expected / predicted / admitted / wanted}. (Kennedy and Merchant 2000)

Likewise, in TSM some verbs such as *tsai-iann* 'know' in (19) allow null clausal complements, while others such as *lin-ui* 'think' in (25) don't.

(25) *Ong-e lin-ui kati-e ke-ue tsin ho,
Ong-e think self-GEN plan very good
Li-e m lin-ui [$_{CP}$ e].
Li-e also think
'Ong thinks that his plan is good, and Li also thinks [that his plan is good].'

What is also intriguing about null clausal complements in TSM is that some ungrammatical sentences with null clausal complements such as (25) turn grammatical after the word *an-ne* 'so' is added as in (26). The clausal complement cannot be omitted in (25), and the addition of *an-ne* turns the ungrammatical sentence (25) into a grammatical one (26).

(26) Ong-e lin-ui kati-e ke-ue tsin ho,
Ong-e think he-GEN plan very good
Li-e m lin-ui an-ne. (cf. (25))
Li-e also think so
'Ong thinks that his plan is good, and Li also thinks so.'

Among the various usages of *so* in English, *so* in (27) functions similarly as *an-ne*. Ross (1972) argues that this *so* is a pro-S. However, against Ross's proposal, Hankamer and Sag (1976) propose that unlike regular clauses, *so* cannot take the subject position, and thus *so* should be a surface anaphora.

(27) Is the moon out? -I believe so.
(Hankamer and Sag 1976)

In fact, unlike *so* in English, TSM *an-ne* can take the subject position as in (28). Cheng (1989) has proposed that in addition to being an adverb as in (29), *an-ne* can function as a proform to refer to a certain action or method. Being a pro-S, *an-ne* is obligatory, and thus a sentence that does not allow null clausal complement such as (25) would be ungrammatical without it.

(28) an-ne ho-m-ho? (Cheng1989)
so good-not-good
'Is it ok to do it this way?'

(29) tsit kiann tai-tsi to an-ne pan looh.
this CL matter then this-way handle PRT
'This matter then can be handled this way.'
(Cheng 1989)

An adverb *an-ne*, however, is optional as in (30). Moreover, *an-ne* can co-occur with the recovered missing element as in (31), which proves that *an-ne* in (31) is indeed an adverb, not a pro-S.

(30) Ong-e huan-lo-kah long be tsiah,
Ong-e worry-kah all not eat
Li-e m huan-lo-kah (an-ne). (cf. (13))
Li-e also worry-kah so
'Ong was so worried that he couldn't eat anything, and Li was also so worried.'
(31) Ong-e huan-lo-kah long be tsiah,
Ong-e worry-kah all not eat
Li-e m huan-lo-kah (an-ne) long tsiah be loh.
Li-e also worry-kah so all eat not down
'Ong was so worried that he couldn't eat anything, and Li was also so worried that he couldn't eat anything.' (cf. (30))

4 Concluding Words

This paper has looked into the null argument that has drawn much less attention in the literature—null clausal complement in Taiwan Southern Min (TSM). Not only clausal complements subcategorized for by predicative verbs but also clausal complements peculiar to TSM—those subcategorized by V-*kah* are discussed.

This paper argues that the null clausal complement in TSM is a type of deep anaphora because it does not require a linguistic antecedent. As to the status of the null clausal complement in TSM, it is argued to be an epithet as it possesses the features of an epithet. Lastly, in some cases where null clausal complements are not allowed, the addition of the pro-S *an-ne* 'so' turns the ungrammatical sentence into a grammatical one.

Acknowledgments

This research was supported by a grant from the Ministry of Science and Technology in Taiwan (MOST 107-2410-H-194-071 -). I would like to thank the anonymous reviewers for their valuable comments and take sole responsibility for any possible errors.

References

Cheng, Robert L. 1985. A comparison of Taiwanese, Taiwan Mandarin, and Peking Mandarin. *Language* 61:352-377.

Cheng, Robert L. et al. (eds.) 1989. *Mandarin function words and their Taiwanese equivalents*. Taipei: The Crane Publishing Co., Ltd.

Hankamer, Jorge, and Ivan Sag. 1976. Deep and surface anaphora. *Linguistic Inquiry* 7:391-426.

Hoji, Hajime. 1998. Null object and sloppy identity in Japanese. *Linguistic Inquiry* 29:127-152.

Hoji, Hajime. 2003. Surface and deep anaphora, sloppy identity, and experiments in syntax. In *Anaphora: A reference guide*, ed. by A. Barss, 172-236. Oxford: Blackwell.

Huang, C.-T. James. 1991. Remarks on the status of the null object. In *Principles and parameters in comparative grammar*, ed. by R. Freidin, 56-76. Cambridge: MIT Press.

Kasai, Hironobu. 2014. On the nature of null clausal complements in Japanese. *Syntax* 17:168-188.

Kennedy, Christopher, and Jason Merchant. 2000. The case of the missing CP and the secret case. In *The Jorge Hankamer WebFest*, ed. by C. Sandra, J. McCloskey, and N. Sanders. Santa Cruz: University of California, Santa Cruz. Available at: http://ling.ucsc.edu/Jorge/index.html

Kim, Soowon 1999. Sloppy/strict identity, empty objects, and NP ellipsis. *Journal of East Asian Linguistics* 8:255-284.

Lin, Huei-Ling. 2003. Postverbal secondary predicates in Taiwanese. *Taiwan Journal of Linguistics* 1:65-94.

Ross, John Robert. 1972. Act. In *Semantics of natural language*, ed. by D. Davidson and G. Harman, 70-126. Dordrecht: Reidel.

Saito, Mamoru. 2007. Notes on East Asian argument ellipsis. *Language Research* 43:203-227.

A Community Detection Method Towards Analysis of Xi Feng Parties in the Northern Song Dynasty

Qianying Liu[*1], **Qiyao Wang**[*2], **Wending Chen**[3] **and Daisuke Kawahara**[1]
[1] Graduate School of Informatics, Kyoto University
[2] National School of Development, Peking University
[3] Department of History, Peking University
ying@nlp.ist.i.kyoto-u.ac.jp; qywang2018@nsd.pku.edu.cn;
paulorchen@pku.edu.cn; dk@i.kyoto-u.ac.jp

Abstract

Factional conflict has played an important role in historical research, espcially studying the features of political scholars. Previous studies mainly analyze factional conflict based on hand-work, which makes it difficult to consider minor political scholars. In this paper, we use the Chinese Historical Biography Database to build a graph based on the relations among scholars in the Xi Feng factional conflict, and then use social network algorithms to model and analyze their impact. We use community detection algorithms to unsupervisedly extract the clusters of the two major parties. By analyzing the data obtained from the algorithm, we confirms several conclusions about the political influence of Cai Jing, Wang Anshi, and the new and old parties in the study of North Song history.

1 Introduction

Factional conflict plays an important role in studying historical events. It is a key figure of parties and it is representative of the conflicts among the ruling class during this dynasty. Factional conflicts in different dynasties often present various characteristics, which are a result of their unique political environment and culture, especially the culture of the political scholars. For example, the conflict between the Niu and the Li in the Tang Dynasty focused on whether to open the path of imperial examination to the poor people, while the factional conflict in the Song Dynasty emphasized the connection between the administration and the law of the ancestors. In general, party disputes may be concentrated within the literati group, or may be composed of foreigners, eunuchs, and scholar-officials, and are not limited to one party. Therefore, studying factional conflict has an important role in studying the characteristics of the dynasty and the literati group. In the dispute between the New and Old parties in the Xifeng period of the Northern Song Dynasty, the New and the Old made a division in the administration. They were a challenge to the ancestral concept of the Song Dynasty, and they were commendable to the spirit of the reform of the previous generation.

The relation network between political scholars during the Xi feng (熙丰) factional conflict is complex. The centralization of power was greatly strengthen in the early years of the Northern Song Dynasty by Song Taizu (太祖), which led to the result that the political environment was potentially full of crises. The ruling class has several new attempts such as "Qingli New Deal (庆历新政)" and the Wang Anshi Reformation (王安石变法), that tried to solve this problem. While previous studies focused on major political scholars such as Wang Anshi and Sima Guang, and also the some common features of the parties (Deng, 2006; Rongke, 1999; Luo, 2011; Liu, 1999). However, due to the limitation of manpower, the studies on minor scholars and indirect relations were limited. Also, the evaluation of different parties and their characters is strongly based on the subjectivity of the historical researchers, and may lose sight of facts that do not match with his expectation. The advantage of digital humanity methods is that it can automatically

[*] This denotes equal contribution.

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 129-137, Hakodate, Japan, September 13-15, 2019

process large-scale data, and the data itself cannot be subjectively influenced by the researchers.

However, one problem for using digital humanity methods on this theme is that labeling data for such kind of party classification would require expert knowledge, and is hard to adapt across different domains (e.g. different dynasties). Thus here we introduce an unsupervised method based on the relation network graph among scholars, which is easy to obtain through databases. Community detection algorithms can be considered as unsupervised clustering algorithms in graph. If nodes are members of the same 'community', they are more likely to connect; if they do not share the 'community', they are less likely to link. The community can correspond to parties or groups in the factional conflict.

We use the data from Chinese Historical Biography Database (CBDB) to construct our relation network. The relationship between the scholars can be seen as a directed multi-graph $G = (V, E)$, where each node V represents a political scholar and each edge E represents a relationship between them, which might be kin relationship or social relationship. The relationships are extracted from CBDB with Breadth-First Search. In this view, many characteristics of the graph can be related to important issues in the history. Therefore, based on this database and graph, we introduce algorithms in graph to analyze our relationship-based graph.

In this paper, we use the Fluid Communities algorithm (Parés et al., 2017) and the improved Kernighan-Lin algorithm kernighan1970efficient, for community detection, and use the results to further analyze some of the properties of the two parties. We use a pipeline-based framework here. First, we use the Fluid Communities algorithm, and get a preliminary result under the condition of considering only the positive relationship. Then we use the results of this step to initialize the improved Kernighan-Lin algorithm so that the new community detection method can consider both positive and negative relationships. Our experiments show that such a pipeline algorithm can achieve better results than using the two alone and benefit our analysis.

We mainly consider the first old and new party conflict in Xifeng years. We combine our method with Liu at el. (2018) to further analysis not only the relations but also the impact of political scholars. Previously, although Liu at el. (2018) used link analysis to study the impact of the scholars, no one has focused on the parties before. Our approach has naturally combined graph models with historical research.

2 Historical Background

In this section we give a brief overview of the details of the factional conflict between the new and old parties. Historians widely accepted Guangming Deng's work "Wang Anshi", which has been polished and modified for four times during the second half of the last century.

In 1067, Emperor Shenzong took the throne, young and ambitious. He highly recommended Wang Anshi's "Reformation Speech Book", in which conveying a core topic, "we should not conform to the politics of the first king, or we will never be achieving success." In the second year of Xining (which is Shenzong's era title, 1069), Shenzong appointed Wang Anshi to be the leader of the government to prepare for the reformation. Wang established the "Regulations on Three Divisions" to carry out the implementation of the new law. Its ideas are mainly divided into two categories: financial one and military one. This challenges the old party. During the reformation, the new factions supporting the new law are obviously opposed to the factions opposing the new law. The ruling group headed by Sima Guang and Wen Yanbo disapproved the new law, and believed that the ancestral law could not be abolished. Wang Anshi had twice stopped, which caused the break of the new law. It is generally believed that the organizational capacity of the New Party is stronger than that of the Old Party.

After the death of Emperor Shenzong, Emperor Zhezong was too young to control the government, giving whose mother Empress Gao a chance to listened to the government and reused Sima Guang and the Old Party. Within a year, the new law was abolished. Coincidentally, Wang Anshi and Sima Guang died in the first year of Yuanyou Era (1086), the new law basically ended in failure in Wang Anshi's time. The second round of factional conflict is after their death. In the 8th year of Yuanyou Era (1093), the emperor became older and also being young and ambitious as his father. He reenabled the charac-

ters of the New Party and protected the reformation again, with changing his era to Shaosheng. However, at this time, the New Party was completely different from the Wang Anshi period, being divided into many little factions. The differences in personality between Zhang Wei, Zeng Bu, Cai Wei and Cai Jing also led to their final destinies. The New Party was also completely divided. This time the conflict was harmful to the dynasty. The parties gradually turned to political factions, not just the contradicting opinions on the ancestral laws, showing their maturity of methods of struggle. In the first month of Yuanfu era (1100), Emperor Zhezong died without a child. Zeng Bu and Zhang Wei and the heir to the throne completely overturned. The second round ends (Li, 2000; Yang, 2010; Liu, 1999; Deng, 2005; Luo, 2011; Tian, ; Zhou, 1996; Rongke, 1999; Deng, 2006).

3 Model

3.1 Page Rank Algorithm

The Page Rank algorithm (Page et al., 1999) is a link analysis algorithm that assigns influence rate to all nodes in the graph, which measures its relative importance within the set. The algorithm can be applied to the graph, and the influence rate r of node E is called the pagerank of E. We follow Liu at el.(2018) and use this algorithm to study the impact of each political scholar in the graph. We follow Liu at el.(2018) and adjust the weights of each edge and use the Page Rank algorithm to calculate the three impact rankings which are the person relationship influence rate, the political influence rate and the positive political influence rate. For the character relationship influence table, we simplified the multi-directional graph to a directed graph with a weight of 1. For the political influence table, we reduce the multi-directed graph to a directed graph whose weight is the number of edges in the multi-graph. For the positive political influence table, we manually label all the relationships and classify them as positive and negative and non-prone, and only use the positive relationships to run the algorithm. The calculated impact rate can represent the influence of the scholar, the total political influence of the scholar and the supporting that the scholar receives.

3.2 Community Detection Algorithm

3.2.1 Fluid Communities Algorithm

In the study of complex networks and graphs, if the nodes of the network can be easily grouped into (possibly overlapping) groups of nodes so that each group of nodes is densely linked internally, the network is said to have a community structure. In the special case of non-overlapping community detection, this means that the network is naturally divided into internal dense linked groups of nodes that sparsely connected between different groups. It can be seen that the community detection algorithm can be used to divide the parties in an unsupervised manner.

Here we use the Fluid Communities algorithm (Parés et al., 2017) for community detection. The Fluid Communities algorithm is a high-speed community detection algorithm based on the idea of introducing several fluids in a non-uniform environment, affected by the environmental topology, and the fluids expand and push each other until they reach a steady state. Given a graph $G = (V, E)$ that consists of a set of vertices V and a set of edges E. The algorithm initializes k fluid communities $C = \{c_1, ..., c_k\}$, where $0 < k \leq |V|$. Each community $c \in C$ is initialized in a different random vertex $v \in V$. The density d of each initialization community is in the range $(0, 1)$. The density of the community is the reciprocal of the number of vertices that make up the community:

$$d(c) = \frac{1}{v \in c} \tag{1}$$

The algorithm operates with supersteps. On each superstep, the algorithm traverses all vertices of v in a random order, updating the community to which each vertex belongs using an update rule. When the vertex's assignment to the community does not change in two consecutive supersteps, the algorithm has converged and ended. The update rule for a particular vertex v returns the community with the largest aggregate density in v's neighbour network. The update rules are defined in the equation below:

$$S = argmax_{c \in S} \sum_{w \in (v, \Gamma(v))} d(c) * \delta(c(w), c) \tag{2}$$

$$\delta(c) = \begin{cases} 1 & c(w) = c; \\ 0 & c(w) \neq c; \end{cases} \tag{3}$$

where v is the updated vertex, $\Gamma(v)$ is the neighbor of v, $d(c)$ is the density of community c, $c(w)$ is the community to which vertex w belongs, S is the set of candidates, and community vertex w belongs to and $\delta(c(w), c)$ is the Kronecker delta.

We only use the positive edges in the graph in this stage. In our problem, since Wang Anshi and Sima Guang have quite frequent positive relationship with Ouyang Xiu and Fan Zhongyan when they are young, it is ineffective to directly divide the map into two communities. In order to better capture the relationship between the parties, we chose to detect five smaller communities and then manually use the prior knowledge to merge the communities.

3.2.2 Improved Kernighan-Lin Algorithm

In the above section, we describe a community detection algorithm based on the positive edges only, but its shortcoming is very obvious. It can not use the large number of negative connections in the data (e.g. attack, impeachment, opposition, etc.) to improve the party classification. In the case of factional conflict, negative relationships tend to more prominently represent the party's affiliation, and the two parties tend to attack each other more frequently. Therefore, we propose a method that adds negative relationships to the community detection algorithm to overcome this shortage.

The Kernighan-Lin algorithm (Kernighan and Lin, 1970) is a greedy heuristic algorithm that divides the network into two communities. The goal of the algorithm is to maximize the objective function Q, which is defined as the difference between the number of connected edges in the two communities and the number of edges between the two communities. The algorithm can be divided into four steps:

1. Randomly initialize or use other algorithms to pre-initialize two sets of nodes;
2. Calculate the number of internal and external edges of each node in the community;
3. Consider all possible node pair exchanges, calculate ΔQ, and exchange greedily;
4. Repeat step 3 until the preset maximum number of iterations is reached, or the maximum ΔQ is negative.

Here we simplify the directed graph into a weighted undirected symbolic network. A symbolic network is a network with both positive and negative weights. This simplification is reasonable because the simple addition of the frequency of positive and negative relationships that occurred between two characters can be representative of the relationship between them.

Here we modify the objective function Q so that we can consider both positive and negative weights. We assign the weights w to edges when calculating Q considering the internal and external edges I_s and E_s. Then the negative weight is naturally included in the objective function.

$$Q = \sum_{s \in I_s} w_s - \sum_{s \in E_s} w_s \tag{4}$$

One major drawback of the Kernighan-Lin algorithm is that the size of the two sets must be given at the beginning. Here we initialize the algorithm using the Fluid Communities algorithm explained above, and then use our improved Kerighan-Lin algorithm. This makes the model both capable to automatically decide the size of the two communities and also consider negative edge information.

3.2.3 Total Party Impact

Here we will directly assess the political impact calculated by the Page Rank algorithm to the two communities to estimate the total influence of the two parties.

$$r_{party} = \sum_{e \in E_{party}} r_e \tag{5}$$

where r_{party} represents the overall impact rate of the party, and E_party represents the nodes in one party. Although the marginal scholars in the two communities are not necessarily members of the two parties, we are actually considering the overall political impact of the two parties. Therefore, a direct weighted sum is a reasonable estimate.

3.3 Breadth-First Search

Breadth-first search (BFS) (Moore, 1959) is an algorithm for searching a graph. The algorithm can start from any node of the graph, which is called a search key, and then first explores all the neighbor nodes of the start key, and then moves on to the next-level neighbor node. The algorithm stops when all reachable nodes of the search key are obtained. In this paper, we follow Liu at el.(2018) and choose Wang Anshi as the search key. In the end we get multiple directed graphs based on the relationships of the scholar.

4 Experiments

4.1 Database

The Chinese Historical Biography Database (CBDB) is a freely accessible relational database with biographical information about approximately 427,000 individuals, primarily from the 7th through 19th centuries. The data is meant to be useful for statistical, social network, and spatial analysis as well as serving as a kind of biographical reference. It involves not only the personal info of historical characters but also the social relationship between them. We show one example of the CBDB database in Table 1.

id 1762: Wang Anshi	
PersonInfo	Basic Info:Eng Name, etc. PersonSources:Source PersonAliases:Alias
...	
PersonSocialStatus	PersonSocialStatus PersonKinshipInfo PersonSocialAssociation

Table 1: Example of CBDB.

4.2 Restrictions

In the search algorithm, we add some restrictions based on historical prior knowledge. To ensure that the person relationship network is established on the characters of the old and new party, we will simplify the network and remove some isolated points that are invalid on the algorithm so that the data extracted from in the original database is clean.

First, there are some dirty data problems in the original database. For example, the relationship with the id 350 is called "temporary reservation, to be deleted". We think that this kind of relationship can be considered as invalid. There are also some relationships that are linked to person ids that do not exist in the database. We remove all such kinds of relationships. Meanwhile, some special relationships such as "Entering Yuanyou Party" are very valuable from a historical point of view, but CBDB does not handle this label very well. Not all people who are involved in this event are marked with this relation. We will not consider these multivariate special relationships here and may consider adding labels manually in further exploration.

For the characters, we require them to have at least one social relationship, and the sum of relatives and social relationships should be at least three. At the same time we require that the index year of the character must exist between 1048 and 1110. The index year is a concept proposed by CBDB and is an artificial value which is used to locate a character in a certain year. The specific calculation is very complicated, here we briefly introduce the rationality of the restriction. If a person's age of death is determined and is less than 60 years old, the index year is the year of death; if a person clearly knows that his or her life is greater than 60 years and knows the year of birth of the person, the index year is the year corresponding to 60 years; If a person's data is missing, use his relatives, the scholar's year, etc., to calculate his life and the year of 60 years old. If the life is less than 60 years old, the estimated death year is taken. If it is more than 60, the estimated 60 years old is used.

We design this limitation based on prior knowledge of the Xifeng parties. Ouyang Xiu's index year is 1068, and Wang Anshi's index year is 1080. In the calculation of the index year, the CBDB used the assumption that the average age that a scholar begins his political career is 30 years old. We use this hypothesis, and set a limitation that the scholars who should their career when Wang Anshi is 60-year-old. The upper limitation bound is calculated by a imaginary character that started his career in the same year of Ouyang Xiu and died 10 years later. The characters in this time period can cover the main body of the old and new party struggles and also consider the

influence of the second old and new party factional conflict and the Qingli New Deal. For a person does not have an index year, it means that with the 20 rules that CBDB designed, we still cannot calculate its index year, then the database lacks information of this person and we will not consider this node.

Here we specifically point out that although the database has already extracted the relationship between the father-in-law pairs and the brother-in-law pairs, and women are often not actual political scholars in ancient China, we still retain the women characters in our graph to further mine implicit information. Most women's social relationships are extracted from their epitaph, which we consider to be important indirect knowledge that is not fully considered by previous studies. Relationship between women also implicitly represent their relatives' political view.

4.3 Community Detection

We show the results of different community detection algorithms in Table 2. We list the scholars that have the highest political influences of the new and old parties. The KL_{noinit} method here is initialized as both parties have the same size. Our algorithms are implemented based on NetworkX (Hagberg et al., 2008).

The underlined names denote classification errors, which is manually labeled by experts based on the historical mainstream views. We can see that the results of the top 20 scholars of the KL-FC initialization algorithm matches with the historical mainstream view, while the uninitialized KL algorithm and FC algorithm present errors. Meanwhile, it is especially difficult to classify scholars of the new party for FC and KL_{noinit}, this is because negative relation is important for detecting the new party members, which makes the two algorithm fail. Meanwhile the new party has less members than the old party, so simply using half of the scholars for initialization of KL algorithm has drawbacks. The results can show the effectiveness of our method.

Meanwhile, the final results show that the old party had a greater influence, with a total of 541 people, it gets an impact score of 0.66491; the influence of the new party was smaller, with a total of 380 people, it gets an impact score of 0.33340.

5 Analysis

We mainly focus on Wang Anshi and Sima Guang, who are the two representative scholars in the Xifeng factional conflict. The factional conflict can be seen for the picture of their political influence.

First, the top four figures of interpersonal influence, political influence, and positive political influence table are consistent. They are Wang Anshi, Ouyang Xiu, Su Shi and Sima Guang. This shows that they are all very powerful and positive in terms of not only from the interpersonal point of view but also from a political point of view. Second, on the issue of the identification of the new and the old parties, we can also see that the results of the KL-FC initialization are basically in line with expectations. The specific analysis will be carried out below.

5.1 Sima Guang and the non-political nature of the old party

During this period, the political culture of the Northern Song Dynasty has taken shape, i.e. the traditional concept of destiny and of orthodoxy. Scholar's political role had been clear, manifested the emperor's centralization of power, to rule the country with scholars but not the military. Scholar-officials gradually had a concept of "to worry about the world before the world, and to feel joy about the world after the world." Therefore, no matter who they are, they must not exceed their own roles, and they must maintain the policy culture of the country. By concentrating on the maintenance of the "ancestral laws", culturally, they recognize political systems related to the nationality formed by ancestral laws. In addition, they also formed a strong sense of political participation, but at the same time, they saw the disintegration of the past Tang Dynasty due to the dispute between Niu and Li, the Song people were very vigilant about the factional struggle formed by the scholars. Therefore, in the Qingli Era New Deal, Fan Zhongyan fell into the party trap because of the old scholars' opposition, but with weak factional effects. Later in Wang Anshi's reform, the factions supporting the new law are obviously opposed to the factions opposing the new law. The group headed by Sima Guang and Wen Yanbo opposed the new law and believed that the ancestral law could not be abolished.

FC		KL_{noinit}		KL_{init}	
Old	New	Old	New	Old	New
Ouyang Xiu	Wang Anshi	Su Shi	Wang Anshi	Ouyang Xiu	Wang Anshi
Su Shi	Wang Gui	Sima Guang	Ouyang Xiu	Su shi	Wang Gui
Sima Guang	Cai Jing	Huang Tingjian	Wang Gui	Sima Guang	Cai Jing
Huang Tingjian	Zhang Dun	Su Shi	Cai Jing	Huang Tingjian	Zhang Dun
Su Zhe	Fan Chunren	Lv Tao	Zhang Fangping	Su Zhe	Lv Huiqing
Zhang Fangping	Fan Zhongyan	Fan Zuyu	Zhang Dun	Zhang Fangping	Yang Jie
Zheng Xie	Bi Zhongyou	Fan Chunren	Zheng Xie	Zheng Xie	Zeng Bu
Zeng Gong	Han Qi	Bi Zhongyou	Zeng Gong	Zeng Gong	Liu Yan
Lv Tao	Su Song	Su Song	Liu Chang	Lv Tao	Wang Anli
Fan Zuyu	Lv Huiqing	Wen Yanbo	Fan Zhongyan	Fan Zuyu	Cai Que
Liu Chang	Yang Jie	Fu Bi	Liu Ban	Fan Chunren	Hua Zhen
Liu Ban	Zeng Bu	Yang Jie	Han Qi	Liu Ban	Zheng Xia
Wen Yanbo	Liu Yan	Lv Gongzhu	Lv Huiqing	Fan Zhongyan	Lu Dian
Fu Bi	Wang Anli	Fan Zhen	Zeng Bu	Liu Ban	Tang Jie
Lv Gongzhu	Song Qi	Zhao Bian	Liu Yan	Bi Zhongyou	Mao Pang
Fan Zhen	Liu Zhi	Hu Yuan	Wang Anli	Han Qi	Huang Lv
Zhao Bian	Li Gou	Liu Zhi	Han Wei	Su Song	Peng Ruli
Han Wei	Cai Que	Chao Puzhi	Song Qi	Wen Yanbo	Jiang Zhiqi
Hu Yuan	Hua Zhen	Qin Guan	Cai Xiang	Fu Bi	Lin Xi
Chao Buzhi	Lv Gongbi	Han Jiang	Shen Gou	Lv Gongzhu	Chao Zhongshen

Table 2: Results of different community detection methods. FC refers to Fluid Communities Algorithm. KL_{noinit} refers to only using the improved Kernighan-Lin Algorithm. KL_{init} refers to using the pipeline model. The list is in the personal impact order.

Sima Guang is best known for his writing of "A General Reflection for Political Administration", who is also very decent personally. He opposed Wang Anshi publicly, yet there is no excessive accusation in personality. Abandoning personal prejudice is a gentleman's behavior and is also a manifestation of the political culture of the Northern Song Dynasty scholars. In fact, many people in the Old Party have similar situations, such as Su Shi, and even some of them in the New Party. They are all typical "spiritual aristocrats", saying, "All the careers are in low status, only learning is the best." In the second round of their conflict, the Old Party and Cai Jing are pure political struggle, that caused substantial personnel changes. Therefore, the old party's "combat power" during the Xifeng period was insufficient, or even can not be called a "party". This is merely an "opposition" for Wang Anshi's ideas. Because the Old Party only appeared for nine years during the period we calculated, this part of the opposition reflects a much lesser relationship with the political mission.

Literary writers and artists have long-term missions. Ouyang Xiu, Su Shi, Zeng Gong, Mi Fu, and Huang Tingjian have left outstanding works of literature and art for later generations, although both Ouyang Xiu and Su Shi have little effect in the struggle between the old and new parties. The children and their sons preserved their work, such as the communication between friends, and even the historical materials of traveling. The literati identity of the Northern Song Dynasty scholars, especially the status during the Qingli and Xifeng years, may be more apparent in the old party. At the same time, their and control in speech made their positive influences stronger. These non-political identities led them to rise in the positive political influence table.

5.2 Party Organizational Capacity

Figure 1 is a personal relationship network table representing the interpersonal influence, corresponding to the left side of Table 1. It can be seen that the new

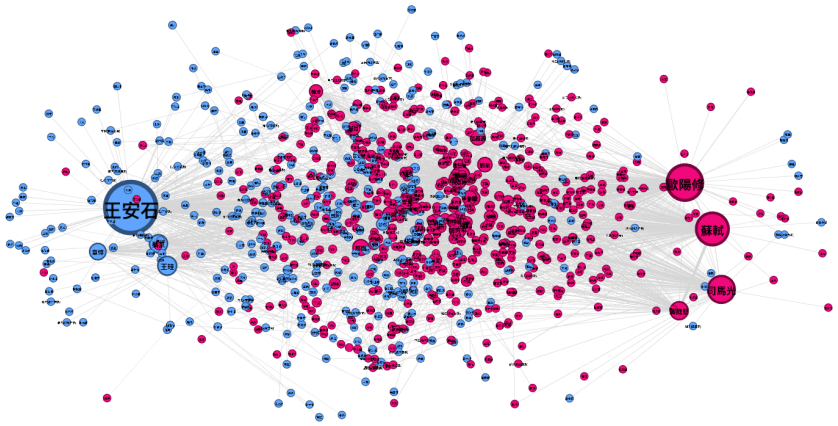


Figure 1: A figure of the result of the pipeline model. The blue dots denotes the new party and the red dot denotes the old party. The size of the dot denotes the personal relation impact of the scholar.

party is dyed in blue and the old party is dyed in red. A circle represents a person. The larger the circles, the greater the impact is and vice versa. Connections indicate an association between the two. The new party is dominated by Wang Anshi's big circle, surrounded by small circles such as Cai Jing, Zhang Wei and Wang Hao. The old party is composed of several middle circles: Ouyang Xiu, Su Shi and Sima Guang. The red circle of the old party is more than the blue circle. According to the above description, it can be verified that Wang Anshi's influence on the New Party is also greater. The old party is more scattered, and Ouyang Xiu, Su Shi and Huang Tingjian have more influence on the influence of the old party on literature than on the political influence. The old party does not have an obvious big circle figure, which matches the fact that their party policy is weaker than the new party.

5.3 Regionality

In the new party, Wang Anshi is a southerner, Lu Huiqing, Zeng Bu, Zhang Wei and later Cai Jing are also southerners; Sima Guang is a northerner, Fuyu, Cheng Hao and Liu Wei are all northerners. The cultural conflicts that existed and evolved into political disputes between the North and the South, and evolved into disputes between the old and new parties. Mr. Qian Mu concluded: "The new party has a large rate of southerners, and the opposition is a big man." "Wang Anshi is in power. It seems that some places represent a new and radical smell of intellectuals in the south, and Sima Guang seems to be Some places represent a traditional and steady attitude in the wisdom of the North at the time." It can be counted that Wang Anshi was mostly removed from the North. After the change of Yuanyou, most of the people who were arrested were southern scholars, such as Cai Zheng and Zhang Wei. After Zhezong's pro-governance, he re-raised the southerners. When Cai Jing was in the country, the southerners regained momentum. "And the northern scholars sighed again and again."

6 Conclusion

This paper constructs a graph based on the relationships of the Northern Song Dynasty Xifeng parties from the CBDB database, on which a community detection method is used. We first use BFS and constraints to construct the graph, and then use the Page Rank algorithm to estimate the impact of different views. We propose a two-step pipeline method that uses the results of the Fluid Communities algorithm to initialize an improved Kernighan-Lin algorithm, so that the new community detection method can consider different types of relations at the same time and automatically learn the size of the communities. Our experiments show that such serial pipeline algorithm can get better results than the two methods seperately, and the experimental results of the two-step algorithm confirms many historical views about parties and factional conflict.

References

Guangming Deng. 2005. The Articles of Deng Guangming, 1907-1998. Hebei Education Press.

Xiaonan Deng. 2006. The Law of the Ancestral Temple. Beijing Sanlian Bookstore.

Aric Hagberg, Pieter Swart, and Daniel S Chult. 2008. Exploring network structure, dynamics, and function using networkx. Technical report, Los Alamos National Lab.(LANL), Los Alamos, NM (United States).

Brian W Kernighan and Shen Lin. 1970. An efficient heuristic procedure for partitioning graphs. Bell system technical journal, 49(2):291–307.

Changxian Li. 2000. Sima guang's biography. Journal of Nanjing University (Philosophy, Humanities and Social Sciences), (1):120.

Qian-ying LIU, Qi-yao WANG, and Wen-ding CHEN. 2018. An analysis of xi feng parties in the northern song dynasty through complex network algorithms. DEStech Transactions on Computer Science and Engineering, (ammms).

Fusheng Liu. 1999. The evolution of the "factional conflict" in the northern song dynasty and the confucianism revival movement[j].

Changfan Luo. 2011. Research on Factional conflicts and Party Inscriptions in the Northern Song Dynasty [D]. Huazhong Normal University.

Edward F Moore. 1959. The shortest path through a maze. In Proc. Int. Symp. Switching Theory, 1959, pages 285–292.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford InfoLab.

Ferran Parés, Dario Garcia Gasulla, Armand Vilalta, Jonatan Moreno, Eduard Ayguadé, Jesús Labarta, Ulises Cortés, and Toyotaro Suzumura. 2017. Fluid communities: a competitive, scalable and diverse community detection algorithm. In International Conference on Complex Networks and their Applications, pages 229–240. Springer.

Wang Rongke. 1999. Political culture of the northern song dynasty and wang anshi's reformation. Journal of Nantong University (Philosophy and Social Sciences), (6):114–119.

Gengyu Tian. The theory of the controversy between the chinese and western niu and li dynasties and the new and old factional conflicts in the northern song dynasty[j]. Journal of Southwest University for Nationalities, Humanities and Social Sciences, 2003, 24(1): 145-151.

Shili Yang. 2010. Officials drop and political culture in the Northern Song Dynasty [M]. Zhongzhou Ancient Books Publishing House.

Yushao Zhou. 1996. On huang tingjian and the factional conflict in the northern song dynasty. Social Science Research, (2):55–59.

Analysis of Reply-Tweets for Buzz Tweet Detection

Kazuyuki Matsumoto[1], **Yuta Hada**[2], **Minoru Yoshida**[1], **Kenji Kita**[1]
[1]Tokushima University, Graduate School of Technology, Industrial and Social Sciences
Minamijosanjima-cho 2-1, Tokushima, Japan
[2]Tech Information Corp., Technical support department,
Inubushihigashidani 6-23, Itano, Tokushiima, Japan
{matumoto;mino;kita}@is.tokushima-u.ac.jp

Abstract

In this study, we propose a method for predicting whether a tweet will create a buzz on the Internet by examining tweeted replies posted by others. We also investigate the distinguishing characteristics of replies to buzz tweets by analyzing feature amounts. Our proposed method first converts each reply tweet into a vector expression using a word distributed representation or some other vectorization method. We then apply a machine learning method for binary classification to determine whether the reply is to a buzz tweet or a non-buzz tweet. We classify the target tweet into "buzz" or "non-buzz" categories by comparing the total "buzz" and "non-buzz" scores produced by the classifier. The proposed method using StarSpace achieved 93.1% F1-score, while an approach that used number of retweets and number of favors ("likes") achieved 77.8% F1-score. We also found that there are a number of words that are characteristic of buzz tweet replies and a number of words that are characteristic of non-buzz tweet replies.

1 Introduction

In recent years, portable information and communication devices have become a common means by which individuals interact with one another via the Internet. Social networking sites such as Twitter, Facebook and Instagram are immensely popular. In Japan, the term "*Bazuru*" (or "buzz") is frequently used to describe a situation in which a topic expands dramatically in a short period of time, attracting the attention of many. The term typically refers to the rapid spread of a topic on the Internet through social media, etc. On Twitter, such a phenomenon is caused by followers or other Twitter users retweeting (RT), registering a "like" (or favor) or replying to a tweet.

In this paper, we describe a method for detecting a "buzz tweet" (i.e., a tweet that induces a "*Bazuru*" phenomenon) using reply tweets as features. This technique makes it possible to discover important tweets and topics from various viewpoints that were difficult to detect with conventional methods.

2 Related Works

Numerous studies have analyzed trend keywords on the web or social media (Cataldi et al., 2010),(Lau et al., 2012),(Cheong and Lee, 2009),(Yu et al., 2011),(Naaman et al., 2011),(Kaushik et al., 2015). Twitter, in particular, is one of the more popular targets of such studies, as it has superior real-time posting. Many of these studies have attempted to detect keywords or topics that sharply increase usage rates and analyze the stream of times considering the keywords as trend keywords. Their main focus has been on analyzing the relation between keywords in the posted tweets and trends in the real world and assessing the factor of the trend.

To clarify the mechanism by which a "buzz" is created, we believe that it is important to determine the distinctive features of buzz tweets by identifying the various types of buzz tweets that exist and analyzing the responses they produce. As a method to grasp the scale of the response, numerical information such as the number of retweets or favors

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 138-146, Hakodate, Japan, September 13-15, 2019

("likes"), as well as the number follows and various follower characteristics, can be useful. However, to fully assess the response to a target buzz tweet, it is necessary to analyze the contents of the replies to the posted tweet.

Counting the number of the retweets is not a particularly reliable indicator of "buzz," as retweeting is a mechanical and easy way to produce information diffusion. This renders simply counting the number of retweets of a given posted tweet a rather limited way to distinguish a true buzz phenomenon from an artificially created one. Various methods (Zaman et al., 2010),(Suh et al., 2010),(Morchid et al., 2014),(Firdaus et al., 2018) that predict the scale of information diffusion based on a change in relationships of users such as the number of follows, or followers, or retweets or favors have been proposed. However, none are capable of determining whether the information diffusion resulting from a particular tweet is due to a true popular phenomena.

There have also been studies that estimate the probability of retweeting by using the correspondence relation between the contents of a tweet and the interest of users (Imamori and Tajima, 2016). However, having interest is not equal to having a favorable opinion of a tweet, which makes it difficult to distinguish a buzz tweet from a "flame."

Another study (Deusser et al., 2018) used the metadata of articles posted on Facebook as features to predict the popularity of an article. The authors of this study used binary classification to determine whether an article is popular by employing a machine learning method. Here, the contents of the article are not used as a feature, as the authors believed that the interaction between an article 's author and his/her friends (viewers) was more related to buzz. In the case of Facebook, the probability that an article will be read by complete strangers is lower than on Twitter. There are thus fewer uncertain elements and the amount of noise in feature values is thought to be smaller. For this reason, such a method may not be particularly effective for evaluating postings on Twitter.

3 Buzz Tweet Classification Method

3.1 Definition of problem

In this study, buzz tweets are tweet contents that have a large number of RTs, replies, likes, give strong impact to readers, and are sympathetic to many people, or that are accompanied by photos or videos. To obtain such tweets, in this paper we define buzz tweets as those listed on the websites (curation sites) that collects buzz tweets filtered by numerical indicators such as the number of RTs.

3.2 Target data

To detect a buzz tweet, it is necessary to collect buzz tweets as training data. For that purpose, a definition of "buzz tweet" is necessary. Because the definition of a buzz phenomenon is ambiguous and it is difficult to establish a clear basis for determining buzz, we collected buzz tweets from various buzz tweet roundup websites.

Many of the roundup sites determine the buzz/non-buzz status of a tweet by using the number of retweets, i.e., the tweet 's degree of diffusion. However, if buzz tweets are identified solely by counting the number of retweets, tweets by celebrities or flame tweets are more likely to be identified as buzz tweets.

In this study, we considered "flame" and "buzz" as different phenomena. Generally, "flame" indicates that a tweet has attracted negative attention, while "buzz" is associated with largely positive responses.

Tweets posted by famous persons such as entertainers, politicians, athletes and YouTubers tend to be diffused at a higher rate than those of general users, which means that their tweets often become buzz tweets. It is natural that the tweets of authors with more fans or friends will produce more retweets or favors, which increases the probability of buzz. When we surveyed the buzz tweet roundup websites, we found that there were several tweets posted by famous persons, but they were small in number. Therefore, we decided not to remove such tweets from the buzz tweets used in our study. Nevertheless, we recognize that diffusion in the case of tweets by famous persons reflects the attributes of the person rather than the contents of the tweet,

Type	# of replies	Avg. # of RTs	Avg. # of Likes
Buzz	13218	30253.6	80845.8
Non-Buzz	16012	2425.0	12925.0
Total	29230	20977.4	58205.5

Table 1: Number of replies and the average # of RTs/Likes.

which makes it difficult to establish them as representing a true buzz phenomenon.

Additionally, as pictures or videos can be posted on Twitter, tweets using such non-verbal media tend to attract more attention and, as a result, are more likely to be identified as buzz tweets.

The following are the websites from which we collected the buzz tweets used in the study. The period of collection was from December 2018 to August 2019:

Collection source of roundup websites:

- iitwi: `https://service.webgoto.net/iitwi/`
- Matome site: `https://matome.naver.jp/odai/2150908164548234501`

In this paper, we validate whether it is possible to classify buzz/non-buzz tweets by using reply tweets. Non-buzz tweets are posted by famous persons with a larger number of retweets or favors. The tweets of famous persons were selected from the tweets of users who ranked in the top 500 in number of followers during the period from 2016 through April 2019.

The number of unique user accounts was 150 for buzz tweets and 151 for non-buzz tweets. For each of the targeted buzz/non-buzz tweets, we collected replies. Table 1 shows the number of replies and the average number of Retweets/Likes for the buzz/non-buzz tweets used in the study.

Because Twitter 's API is unable to collect replies to specific tweets, we manually collected the replies that could be viewed. We defined seven categories, from A to G, for the buzz tweets based on their factors of buzz. A breakdown of the 150 buzz tweets is shown in Table 2.

As shown in the table, categories C (buzz tweets due to images or videos) and F (buzz tweets due to jokes or funny behavior/utterance) make up the vast majority of the buzz tweet factors. Table 3 shows the example of buzz tweet and its replies.

Category	Example (Factor)	Count
A	knowledge	13
B	surprise news	3
C	image, video	61
D	celebrity news, information	7
E	moral, social remarks	11
F	joke, funny behavior/utterance, etc.	51
G	common thing	4

Table 2: Number of tweets for each buzz category.

3.3 Flow of the buzz tweet classification

Our proposed method constructs a binary classifier (buzz/non-buzz) that uses the reply texts posted to buzz/non-buzz tweets as features and uses the buzz/non-buzz tweet to which a reply text is posted as a label.

To judge whether an unknown tweet is a buzz or non-buzz tweet, a buzz/non-buzz classification score is produced by the classifier for each posted reply to the tweet. These scores are then aggregated to produce a total classification score for each of the two classifications (buzz/non-buzz). The larger of the two scores determines whether the unknown tweet should be judged a buzz tweet or a non-buzz tweet.

Eq.1 and 2 show the buzz score calculation and label judgement criteria. $Prob_{i,x}$ indicates the probability of label x estimated by the classifier for reply i posted to tweet t; $label_t$ shows the result of the label judgment for tweet t, determined by comparing the magnitude of the two total label scores.

$$Score_x = \sum_{i=1}^{N} Prob_{i,x} \tag{1}$$

$$label_t = \begin{cases} buzz & (Score_{buzz} > Score_{nonbuzz}) \\ nonbuzz & (Score_{buzz} \leq Score_{nonbuzz}) \end{cases} \tag{2}$$

The flow of the proposed method is shown in Fig.1.

3.4 Conversion of reply tweet into vector

It is not easy to identify features from the reply tweets posted to a buzz tweet without formatting. Therefore, we embedded each reply into the feature

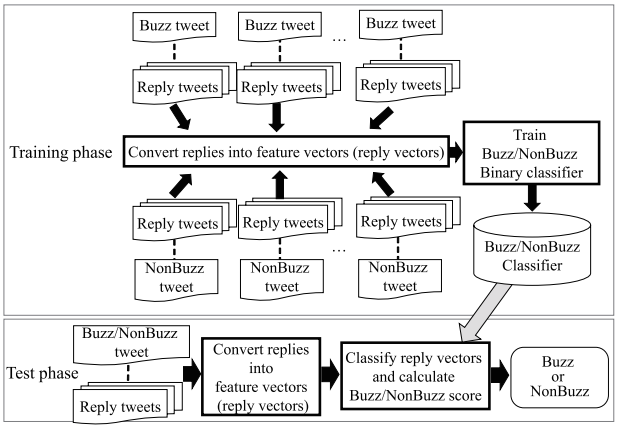


Figure 1: Flow of the proposed method.

space. Recently, techniques such as word2vec that express words or sentences with fixed length vectors are being used not only in studies but also in various actual services.

Buzz tweet (# of RT: 13034, # of Favorite: 23504)
If the insistence that "an artist is arrested and all the work of that person can not be used" is strictly used, perhaps the most significant impact on Japan at the time of becoming a suspect is probably "Illust-ya's creator".
Example of replies
Ex.1) Conspiracy of people who are trying to capture the share of "Illust-ya" starts to move. I understand. **Ex.2)** I think Asei Kobayashi is quite good (lol). ← I checked again by chance.

Table 3: Example of buzz tweet and its replies.

In this study, we employ a method to convert reply data into vectors by using unsupervised pre-training based on a large-sized corpus. By pre-training the reply vectors, we are able to create a buzz/non-buzz judgement model that is robust to unknown words based on a small-sized corpus.

As a baseline, we apply the dimensional compression method that applies TF-IDF-based keyword weighting to bag of words vectors. This method is often used for document retrieval or document classification.

In this study, we used the following vectorizing methods and machine learning models:

- Averaged word vector (AWV)
- CNN, bi-LSTM, bi-GRU
- character-AutoEncoder-Decoder trained by CNN, LSTM, and GRU
- StarSpace
- BERT (Bidirectional Encoder Representations from Transformers)
- Baseline: bag of words vector (tfidf-weight)

The following subsections explain each method in sequence.

3.4.1 Averaged word vector(AWV)

This method employs the averaged vector(AWV) of a word distributed representation trained by the fastText (Joulin et al., 2016) algorithm using a Japanese tokenized corpus. Because the fastText algorithm can consider the sub word information of words, this method is more robust to unknown words than word2vec. The buzz/non-buzz binary classifier was created by training feed forward neural networks (FFNN) using averaged vectors as features.

In our experiment, we use 300 dimension distributed representations that were trained based on Japanese Wikipedia articles.

3.4.2 CNN, bi-LSTM, bi-GRU

Here we created a buzz/non-buzz binary classifier by training Convolutional Neural Networks (CNN), Bidirectional Long-short Term Memory (bi-LSTM), Bidirectional Gated Recurrent Unit (bi-GRU) using pre-trained word distributed representations as features, which is the same as the features used in the averaged word vector method.

Because the length of the various reply texts differs, we applied padding to the reply data as preprocessing. From the average number of words, we set the maximum word number as 30.

We also created a classifier by neural network using a simple attention mechanism (Luong et al., 2015). The structure of the self-attention bi-LSTM network is shown in Fig.2.

3.4.3 Character-AutoEncoder-Decoder trained by CNN, LSTM, and GRU

Because the reply texts include several character strings that are difficult to divide morphologically, we applied training per character. We first created

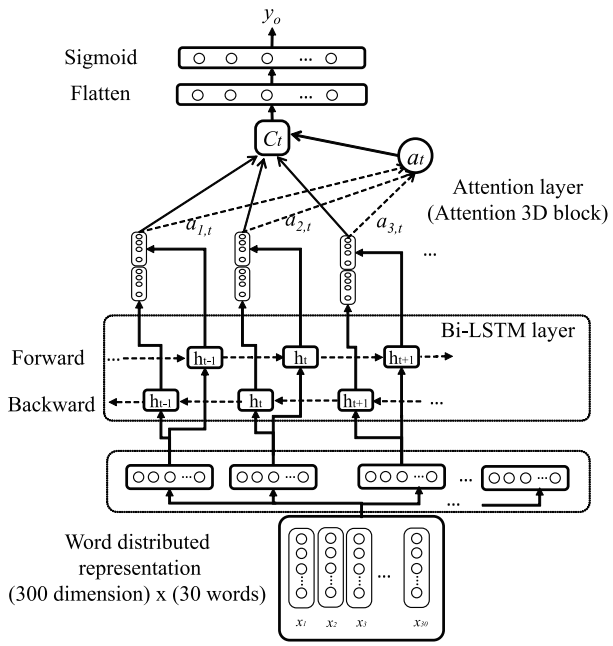


Figure 2: Bi-LSTM attention network.

	Layer type	In/out	Tensor shape
Encoder	Input	in:	(None, 140, 500)
		out:	(None, 140, 500)
	Conv1D	in:	(None, 140, 500)
		out:	(None, 140, 256)
	MaxPooling1D	in:	(None, 140, 256)
		out:	(None, 70, 256)
	Conv1D	in:	(None, 70, 256)
		out:	(None, 70, 128)
	MaxPooling1D	in:	(None, 70, 128)
		out:	(None, 35, 128)
	Conv1D	in:	(None, 35, 128)
		out:	(None, 35, 64)
Decoder	Conv1D	in:	(None, 35, 64)
		out:	(None, 35, 64)
	UpSampling1D	in:	(None, 35, 64)
		out:	(None, 70, 64)
	Conv1D	in:	(None, 70, 64)
		out:	(None, 70, 128)
	UpSampling1D	in:	(None, 70, 128)
		out:	(None, 140, 128)
	Conv1D	in:	(None, 140, 128)
		out:	(None, 140, 256)
	Conv1D	in:	(None, 140, 256)
		out:	(None, 140, 500)

Figure 3: CNN auto encoder-decoder note: ('None' is the batch dimension).

character-based one-hot-vectors (maximum character length: 140), then trained an encoder-decoder that reproduced the original texts by using CNN, LSTM and GRU. By using the output of the encoder of the trained model, we converted the reply texts into fixed length vectors.

The buzz/non-buzz binary classifier was created by training the FFNN using the obtained vector as input. Fig. 3 shows the character-based AutoEncoder-Decoder by CNN. The AutoEncoder-Decoder based on LSTM and GRU, respectively, consists of four layers for the encoders and three layers for the decoders. There were 15 training epochs for CNN, 50 for LSTM, and 8 for GRU.

3.4.4 StarSpace

The "StarSpace" (Wu et al., 2018) algorithm converts text into distributed representations. Because StarSpace can learn effective distributed representations for the text classification task, we were able to create a model to classify replies accurately without pre-learning.

We trained a one-to-one classifier (which estimates one label for the one inputted text) by StarSpace, and used it to classify replies as buzz/non-buzz. We set the parameters n of the word n-gram at 3 and the number of word distributed representation dimensions at 100. We used Sentencepiece (Sentencepiece,) as a tokenizer with vocabulary size is 10000.

3.4.5 BERT(Bidirectional Encoder Representation from Transformers)

BERT, developed by Google (Devlin et al., 2018), is a model that can produce versatile distributed representations. To apply the BERT model to Japanese reply texts, we generated 768 dimensional distributed representation vectors by using the pre-trained BERT model trained with Japanese Wikipedia articles (BERT,). Using the vectors as feature vectors, we created a buzz/non-buzz classifier using a perceptron without hidden layers.

3.4.6 Baseline: bag of words vector(TF-IDF)

As feature words, we selected words with high importance values (applying a threshold) based on TF-IDF values, set vector dimensions, respectively, for the important words, and created vectors with TF-IDF values as dimension values. We constructed a buzz/non-buzz classifier by FFNN using the TF-IDF vectors as features. In this paper, we removed

words with TF-IDF values under the threshold and decided on 197 dimensions.

4 Experiment

4.1 Preliminary experiment

As features other than the reply tweets to the buzz/non-buzz tweets, numerical measures such as the number of users following the poster, the poster ' s number of followers, and the number of retweets, favors, etc., as well as features obtained from the tweets themselves or from user profiles, were available. We conducted a preliminary experiment to evaluate variations of a classifying method based on combining these features. A support vector machine (SVM) was used to train the classifier. We divided the data by 10 and conducted a cross validation to evaluate the performance of the method. We used Recall, Precision and F1-score as evaluation scores (see Eq.3, 4, 5). TP_x means true positive of label x, FP_x means false positive of label x, and FN_x means false negative of label x.

$$R(Recall)_x = \frac{TP_x}{TP_x + FN_x} \times 100 \quad (3)$$

$$P(Precision)_x = \frac{TP_x}{TP_x + FP_x} \times 100 \quad (4)$$

$$F1(F1 - Score)_x = 2 \times \frac{R_x \times P_x}{R_x + P_x} \quad (5)$$

Table 4 shows the features used in the experiment. Results from the preliminary experiment are shown in Table 5. As indicated, the M2 feature combination produced the highest F1-score level (77.8%) for buzz tweets. The lowest F1-Score for buzz tweets of combinations M6 shows that the number of Follows/Followers and total number of favors ("likes") were more important than the features obtained from the tweets themselves (feature ID: 14). With the exception of combination M7, the various feature combinations showed higher F1-Score in the non-buzz tweet classification than in the buzz tweet classification.

These results suggest that famous persons who posted target non-buzz tweets might have distinctive characteristics with respect to the number of retweets or favors.

Because the buzz tweet classification F1-Scores were all under 80% in the preliminary experiment, we concluded that features from the buzz tweets themselves and user account information were not suitable for identifying buzz tweets.

ID	Feature type
1	# of replies
2	# of Retweets
3	# of Favors
4	# of Follows by the account
5	# of Followers of the account
6	Total # of Favors
7	Total # of List
8	Total # of Moment
9	Total # of tweets
10	Elapsed days from the date when the account registered
11	Whether the account is locked
12	Whether an image is attached
13	The # of characters of the tweet
14	The averaged word vector of the tweet (300 dimension)
15	The averaged word vector of the profile text (300 dimension)

Table 4: Feature type.

Method	Buzz			Non-Buzz		
(Feature IDs)	R	P	F1	R	P	F1
M1 (1-13)	84.8	71.1	77.4	75.4	87.4	81.0
M2 (1-12)	84.9	71.8	**77.8**	**75.9**	87.4	**81.2**
M3 (1-3, 12)	62.0	53.7	57.6	59.6	67.5	63.4
M4 (1-3)	**89.7**	17.4	29.2	54.6	**98.0**	70.1
M5 (4-11)	85.0	68.5	75.8	73.9	88.1	80.4
M6 (13, 14)	55.3	17.4	26.5	51.4	86.1	64.4
M7 (15)	50.0	**84.8**	62.9	46.3	13.4	20.8

Table 5: Result of preliminary experiment.

4.2 Evaluation experiment

Table 6 shows the experimental results when features of the replies were used. We conducted our cross validation test by using the same data divided into 10 groupings as in the preliminary experiment. As indicated, 94.7% classification precision for buzz tweets was achieved when BERT was used. In the case of AWV, TF-IDF, bi-LSTM+Attention, and LSTM-AE, the classification precisions for buzz tweets were over 85%.

We think the reason of the highest F1-score of StarSpace is mainly due to supervised learning of word embedding. On the other hand, BERT and the

other methods used unsupervised pre-trained word embedding.

The baseline method using TF-IDF produced 85.0% precision for buzz tweets, which was not particularly low. In fact, the classification recall for buzz tweets was over 94%.

Method	Buzz			Non-Buzz		
	R	P	F1	R	P	F1
AWV	89.3	86.5	87.9	86.1	89.0	87.5
TF-IDF	94.7	85.0	89.6	83.4	94.0	88.4
CNN	93.3	83.8	88.3	82.1	92.5	87.0
bi-LSTM	95.3	84.1	89.4	82.1	94.7	87.9
bi-GRU	**96.7**	83.3	89.5	80.8	**96.1**	87.8
bi-LSTM + Attention	95.3	86.1	90.5	84.8	94.8	89.5
CNN-AE	80.0	82.2	81.1	82.8	80.6	81.7
LSTM-AE	88.0	86.3	87.1	86.1	87.8	87.0
GRU-AE	90.0	84.9	87.4	84.1	89.4	86.7
StarSpace	94.7	91.6	**93.1**	91.4	94.5	**92.9**
BERT	88.2	**94.7**	91.3	**94.3**	87.4	90.7

Table 6: R, P, F1 of each method.

5 Analysis and Discussion

One of the primary aims of our study was to analyze the features of buzz tweets. Accordingly, in this section, from the training results of the classifier, we present our analysis of the distinguishing characteristics of replies to buzz tweets and replies to non-buzz tweets. We first randomly extracted 10000 reply vectors based on BERT. We compressed the vectors into two dimensions using the t-SNE algorithm (Maaten and Hinton, 2008) and plotted them in two-dimensional space. As can be seen in Fig.4, the replies to buzz tweets and non-buzz tweets are not clearly divided into buzz and non-buzz groupings.

To analyze this plot, the two-dimensional reply vectors compressed by t-SNE were clustered into eight clusters by k-means algorithm. Among these clusters, there were two clusters (cluster1, cluster4: these clusters are "magenta" and "lime" in Fig.5) where the numbers of non-buzz replies were twice as large as the numbers of buzz replies. As for these two clusters, we investigated the frequently appearing expressions in non-buzz replies by calculating the frequency of word appearance. The feature expressions frequently appeared in the non-buzz replies are shown in Table 7.

This result showed that among the non-buzz

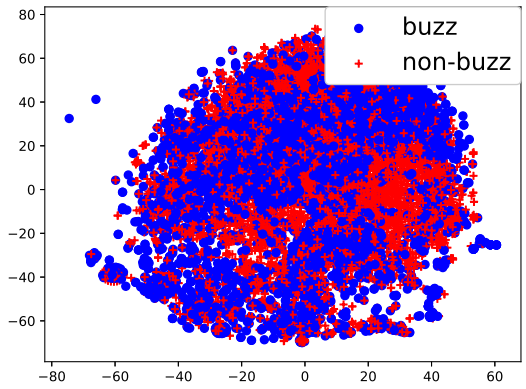


Figure 4: t-SNE plotting of BERT reply vectors.

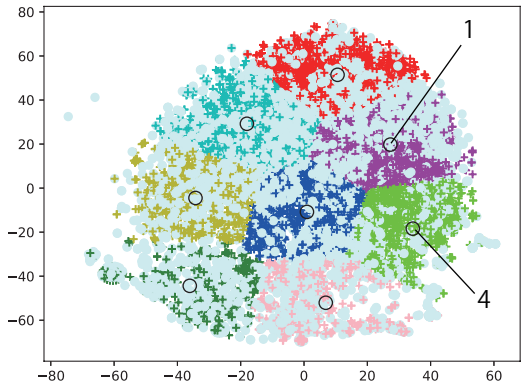


Figure 5: 8-clusters by k-means.

replies, there were comments on the events such as the concert or on their performance on TV programs, expression of thanks, support messages from the fans to their admiring famous persons. Next, we analyzed the differences between buzz and non-buzz tweets according to the word distributed representations obtained by training the reply texts, using features in the replies to buzz and non-buzz tweets based on the distributed representations trained by StarSpace.

StarSpace classifies texts by calculating the inner product of the label distributed representations and the vector summation of the distributed representations of the words in the texts. Therefore, we believed that the feature words for buzz/non-buzz could be obtained by calculating the similarity be-

Cluster	Frequently expressions in non-buzz replies
1	*Arigato* (Thank you) *Omedeto* (Congratulations) *ouen* (support)
4	*Tanoshimidesu* (I 'm looking forward it) *Yoroshiku* (Glad to see you) *Saikoudeshita* (It was the best)

Table 7: Frequently expressions in non-buzz replies.

tween the word distributed representations and the label distributed representations of buzz/non-buzz.

A partial list of the feature words is provided in Table 8. The numerical values indicate the cosine similarity with the given label. The table shows that in the replies posted to buzz tweets, distinctive expressions such as "*buzztteru*," "RT," and "FF" (an abbreviation of Follow/Follower) appeared, all of which are related to buzz phenomena on Twitter.

In contrast, in the replies posted to non-buzz tweets, there were many proper nouns (names/nicknames of famous persons, affiliations, etc.), as well as emojis (emoticons) or greeting expressions. This is thought to be because the fans (primarily, followers) of famous persons often post relatively polite replies that include greeting expressions or emotional expressions with many emojis.

Because many of the buzz tweet authors are not famous persons, they typically have a relatively small number of followers. Therefore, the attributes of the reply users are not limited to followers and cover a wider range than the reply users of famous figures. This may be one important reason why replies would be effective features for buzz/non-buzz classification.

On the other hand, proper nouns did not appear with exceptional frequency in the replies to buzz tweets. However, a large number of admiration expressions; such as "*warota*"(means have laughed), "genius" were found among the expressions that appeared in the buzz tweet replies.

6 Conclusions

In this paper, we proposed a method to classify buzz tweets by using reply features, which contain much

Buzz			
RT (Retweet)	0.90	***buzztteru***[1](now buzzing)	0.81
warota[1] (laughed)	0.89	***maji***[1](really)	0.81
kusa[1] (laugh)	0.87	*Garigarigarikuson*[2]	0.81
dekusa[1] (laugh)	0.84	Snape teacher	0.79
FF (Follow-Follower)	0.83	Excel	0.78
operation	0.83	station	0.77
Wakuwakusan[2]	0.83	*waratta* (laughed)	0.74
foreign citizen	0.82	*tensai* (genius)	0.73
Mac	0.82	(; ∀;)[3]	0.69
Non-Buzz			
Yoshimoto	0.94	politician	0.87
election	0.93	**emoji** (cherry blossom)	0.86
I'm looking forward to it	0.92	It was the best	0.86
pleasure	0.90	I support you	0.86
Sugichan[2]	0.89	***Murosan***[2]	0.85
Korea	0.87	participation	0.82
TV	0.87	***KeisukeHonda***[2]	0.85
jerky	0.87	**emoji** (dazzle)	0.83

Table 8: Example of terms that are important in each category.

richer information than numerical information such as the number of replies or favors. The proposed method converts the replies posted to buzz tweets into feature vectors and constructs a buzz tweet classification model by training the vectors with a machine learning method.

Based on results from an evaluation experiment and treating tweets by famous persons as non-buzz tweets, the method using StarSpace with SentencePiece tokenizer classified buzz tweets with 93.1 F1-score. This score is significantly higher than that of other methods that use such features as the number of retweets, number of follows, etc., all of which achieved less than 80 F1-score.

In the future, we plan to analyze whether the level of accuracy would change if we consider the posting times of the replies. According to our analysis of the differences in feature words between buzz and non-buzz tweet replies, we perceived a certain bias in the appearance of expressions related to differences in various attributes of the replying users.

As part of our extended investigation, we intend to determine whether the proposed method using reply features is capable of distinguishing buzz tweets

[1]Japanese slang
[2]Name of famous person
[3]Japanese emoticon

from flame tweets through additional experiments. We also plan to include additional features such as whether the replying user has a relationship to the author of the original tweet (e.g., is a follow or follower of the author), as this would seem to be an important feature for classification.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP18K11549.

References

Mario Cataldi, Luigi D. Caro, and Claudio Schifanella. 2010. Emerging topic detection on Twitter based on temporal and social terms evaluation *Proceedings of the Tenth International Workshop on Multimedia Data Mining*, Article 4.

JeyHan Lau, Nigel Collier, and Timothy Baldwin. 2012. On-line Trend Analysis with Topic Models: #twitter trends detection topic model online, *Proceedings of COLING 2012: Technical Papers* 15191534.

Marc Cheong and Vincent Lee. 2009. Integrating web-based intelligence retrieval and decision-making from the twitter trends knowledge base, *Proceedings of the 2nd ACM workshop on Social web search and mining*, 1-8.

Louis Yu, Sitaram Asur and Bernardo A. Huberman. 2011. What Trends in Chinese Social Media, *Proceedings of The 5th SNA-KDD Workshop '11 (SNA-KDD '11).*

Mor Naaman, Hila Becker and Luis Gravano. 2011. Hip and Trendy: Characterizing Emerging Trends on Twitter, *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*, 62(5):902918.

R. Kaushik, S. Apoorva Chandra, Dilip Mallya, J. N. V. K. Chaitanya and S. Sowmya Kamath. 2015. Sociopedia: An Interactive System for Event Detection and Trend Analysis for Twitter Data, *Proceedings of 3rd International Conference on Advanced Computing, Networking and Informatics*, 63-70.

Tauhid R. Zaman, Ralf Herbrich, Jurgen V. Gael and David Stern. 2010. Predicting Information Spreading in Twitter, *Computational Social Science and the Wisdom of Crowds Workshop (colocated with NIPS 2010).*

Bongwon Suh, Lichan Hong, Peter Pirolli, and Ed H. Chi. 2010. Want to be Retweeted? Large Scale Analytics on Factors Impacting Retweet in Twitter Network, *2010 IEEE Second International Conference on Social Computing*.

Mohamed Morchid, Georges Linares, and Richard Dufour. 2014. Characterizing and Predicting Bursty Events: The Buzz Case Study on Twitter, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, 27662771.

Syeda N. Firdaus, Chen Ding and Alireza Sadeghian. 2018. Retweet: A popular information diffusion mechanism A survey paper, *Online Social Networks and Media*, 6(2018), 26-40.

Daichi Imamori and Keishi Tajima. 2016. Predicting Popularity of Twitter Accounts through the Discovery of Link-Propagating Early Adopters, *Proceedings of Conference of Information and Knowledge Management (CKIM2016)*.

Clemens Deusser, Nora Jansen, Jan Reubold, Benjamin Schiller, Oliver Hinz and Thorsten Strufe. 2018. Buzz in Social Media: Detection of Short-lived Viral Phenomena, *WWW '18 Companion Proceedings of the The Web Conference 2018*, 1443-1449.

Armand Joulin, Edouard Grave, Piotr Bojanowski and Tomas Mikolov. 2016. Bag of Tricks for Efficient Text Classification, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Volume 2, Short Papers, 427431.

Thang Luong, Hieu Pham and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP2015)*, 1412-1421.

Ledell Wu, Adam Fisch, Sumit Chopra, Keith Adams, Antoine Bordes, and Jason Weston. 2018. StarSpace: Embed All The Things!, *Proceedings of the thirty-second AAAI conference on artificial intelligence (AAAI-18)* , 5569-5577.

Sentencepiece: https://github.com/google/sentencepiece

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, arXiv:1810.04805.

BERT Japanese Pretrained Model: http://nlp.ist.i.kyoto-u.ac.jp/index.php?BERT

Laurens V. D. Maaten and Geoffrey Hinton. 2008. Visualizing Data using t-SNE, *Journal of Machine Learning Research*, 9(2008), 2579-2605.

Evaluating the suitability of human-oriented text simplification for machine translation

Rei Miyata
Nagoya University
miyata@nuee.nagoya-u.ac.jp

Midori Tatsumi
Rikkyo University
midori.tatsumi@rikkyo.ac.jp

Abstract

We present the results of an experiment to evaluate the suitability of simplified text as a source for machine translation (MT). Focusing on Japanese as the source language, we first proposed a simplest possible rule set to write text that can be easily understood by language learners and children. Following this rule set, we manually rewrote expository sentences concerning Japanese cultural assets in simplified Japanese, through two steps: (1) splitting long sentences into short complete sentences, and (2) further simplifying these. We then conducted a human evaluation to assess the quality of the English MT outputs of the original, split, and simplified sentences. The results indicated the potential of simplified text as an effective source for MT, demonstrating that nearly 80% of the raw MT outputs achieved usable quality. The qualitative analyses also revealed occasional side effects of simplification and fundamental difficulties for MT.

1 Introduction

Text simplification is the process of reducing the complexity of the sentence structure and difficulties of the words in a given text. The applications of text simplification vary from reading aids for human readers to preprocessing for natural language components, such as machine translation (MT). While automatic text simplification techniques have been proposed, with the effectiveness demonstrated on certain evaluation tasks, many practical attempts, such as Simple English Wikipedia, rely mostly on manual text simplification with some writing guidelines. In this context, we have developed a simplified Japanese rule set for non-professional writers, which requires the rules to be simple for such writers to follow. Our rule set is intended for writing expository text on Japanese cultural assets. This is challenging, as such texts contain many culture-specific technical terms that are difficult to simplify, even for human writers.

Although the primary purpose of our simplified Japanese is to enhance the text readability for those with limited Japanese proficiency, such as language learners and children, we are also interested in investigating the machine translatability of a simplified source text, especially considering the recent developments of neural MT (NMT) technology. To date, little effort has been invested in examining the compatibility between text simplification approaches for human readers and MT in detail. Here, three major questions arise:

1. To what extent can manual text simplification improve MT outputs?
2. What types of simplification operations are effective or ineffective for improving the MT quality?
3. What types of translation difficulties remain even after the source text is simplified?

Therefore, in this study, focusing on Japanese and English as the source and target languages, respectively, we address these questions by proposing simplified Japanese for human readability and evaluating the suitability of the simplified text as a source for MT. To investigate the effect of the simplification process in detail, we decompose it into two op-

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 147-155, Hakodate, Japan, September 13-15, 2019

erations: (1) splitting long sentences into short complete sentences, and (2) further simplifying these. To test the suitability of this simplification for MT, we evaluate the MT output quality and diagnose the MT errors.

We discuss related work in Section 2, and introduce our guidelines for simplifying Japanese in Section 3. Section 4 describes the process and product of the manual simplification of text. We explain our experimental setup in Section 5, and present our results with in-depth analyses in Section 6. Finally, Section 7 concludes the paper and proposes future research directions.

2 Related work

Automatic text simplification has been tackled in the natural language processing research field for various purposes and languages (Siddharthan, 2014; Shardlow, 2014). However, full automation remains difficult, particularly for human-oriented text simplification tasks, which require the produced text to be of high quality. In many practical applications, human writers conduct text simplification tasks by means of authoring guidelines and technological aids. For instance, Wikipedia provides guidelines and introduces several existing support tools for writing simplified English versions of regular Wikipedia pages.[1] The guidelines specify vocabulary lists such as Basic English 850/1500 and simple sentence structures. They also define the preferred use of voice (active voice) and tenses (past, present or future only).

One of the most widely-acknowledged simplified Japanese rule sets is *Yasashii Nihongo*, or 'Easy Japanese', proposed by the Sociolinguistics Laboratory at Hirosaki University.[2] This consists of 12 writing rules, which restrict the vocabulary and regulate certain types of complex structures, such as long sentences and double negation. Because the original purpose of Easy Japanese was to provide foreign residents in Japan with emergency information, the vocabulary restrictions are rather strict, with about 1400 basic words, which corresponds to the Japanese-Language Proficiency Test (JLPT) Grade 3 and Grade 4.[3]

Inspired by this rule set, several human-oriented simplified Japanese guidelines have been developed, such as those for disseminating local community information (Iori, 2016) and writing news report scripts (Tanaka et al., 2013). While the details of these simplified languages differ depending on the purpose and audience, the shared core idea is to prescribe a vocabulary list and restrict complex sentence structures, which basically corresponds to two major subtasks of (automatic) text simplification: lexical and syntactic simplification (Shardlow, 2014; Saggion, 2017).

One of the most important aspects of a practical implementation of simplification lies in the simplicity of the guidelines. Some simplified languages that are mainly utilised by professional writers specify detailed usages of lexicons, grammars and styles. For example, ASD-STE100 (ASD, 2017), also recognised as a controlled language, defines 53 writing rules and a dictionary of approved and not-approved words for writing technical documentation. However, for non-professional writers, the guidelines themselves should be sufficiently simple for utilisation.

MT-oriented text simplification has also been undertaken (Hung et al., 2012; Štajner and Popovic, 2016; Štajner and Popovic, 2018). For example, Štajner and Popovic (2016) employed two automatic text simplification systems to produce lexically and syntactically simplified versions of source text for English-to-Serbian statistical MT, and evaluated the MT outputs in terms of the fluency, adequacy and post-editing effort. While these studies demonstrate the efficacy of automatic text simplification techniques for MT applications, two major issues remain: (1) human readability is not explicitly taken into account, and (2) the potential gain in MT quality when manual text simplification is fully performed is not measured.

In the research field of controlled language, several evaluation experiments have examined the compatibility or commonality between human-oriented

[1] Simple English Wikipedia, https://simple.wikipedia.org/wiki/Wikipedia:How_to_write_Simple_English_pages

[2] http://human.cc.hirosaki-u.ac.jp/kokugo/EJ9tsukurikata.ujie.htm

[3] JLPT Grade 3 and Grade 4 correspond to current versions of N4 (the ability to understand basic Japanese) and N5 (the ability to understand some basic Japanese). https://www.jlpt.jp/e/about/levelsummary.html

and MT-oriented controlled language rules (O'Brien and Roturier, 2007; Aikawa et al., 2007; Hartley et al., 2012; Miyata et al., 2015). However, these studies tend to focus on structural and stylistic aspects of technical documents. The effect of vocabulary restriction, which is a major task of text simplification, has not been significantly investigated. Moreover, NMT systems has not yet been examined.

As Koehn and Knowles (2017) demonstrated, despite its recent advancements NMT still faces difficulties in dealing with low-frequency words and long sentences, among others. This naturally motivates us to assume that text simplification that restricts vocabulary and sentence complexity can be helpful to enhance MT quality, even if it is intended for human readability. However, as noted by Hartley et al. (2012) and Miyata et al. (2015), there are incompatibilities between human readability and machine translatability. Therefore, an in-depth analysis of the suitability of human-oriented text simplification for MT is required to understand its potential and limitations.

3 Simplified Japanese rule set

There is no single standard rule set for simplified Japanese. Variations exist to adjust the level of Japanese depending on the type of information to be conveyed and the target audience, as mentioned in Section 2. For writing about cultural assets, at least an upper-intermediate vocabulary level will be required. On the other hand, the sentence structure could be limited to a basic level.

In general, simplified Japanese is written by Japanese language teachers, or those who are trained to author in simplified Japanese. However, our aim is to create a rule set that is sufficiently simple to be understood and followed by lay people, namely those that are neither professional linguists nor Japanese language teachers. Therefore, we avoid using grammatical terms or complicated linguistic concepts when setting the rules. Essentially, our rule set consists of just the following three rules.

Rule 1: Present no more than one idea per sentence.

Rule 2: Specify the subject as far as possible, and if the subject is implied then use the passive tense.

Rule 3: Use only the vocabulary and Kanji (Chinese characters) of up to JLPT Grade 2.

JLPT no longer has an official list of vocabulary and Kanji for each level. Thus, we employed the equivalent list available on the website of the Faculty of Humanities and Social Sciences at Hirosaki University.[4] This list contains 3,708 words for Grade 2, 688 words for Grade 3 and 740 words for Grade 4, where smaller grades indicate a higher level.

It should be noted that there are cases in which we could not rewrite a sentence to strictly conform to these rules. For example, there are sentences that are left without a subject, as specifying a subject for every predicate can make some Japanese sentences sound unnatural. We also left proper nouns as they are, even if they are not found in the list of vocabulary and Kanji up to Grade 2.

4 Simplification

4.1 Dataset

We collected 1,274 Japanese sentences from leaflets on historical buildings and houses that have officially been designated as Japanese cultural assets. These leaflets are available either as printed matters at the physical sites, or in downloadable electronic format (PDF) on their official websites.

In the collected text, we identified the following nine topics: *Style and features*, *History and episodes*, *Owner and resident*, *Architect*, *Environment*, *Artefacts and objects*, *Access information*, *Captions and titles*, and *Other*. We categorised each sentence according to the topics, because the topic is an important determining factor for the grammatical construction of a sentence. For example, many of the sentences in the *Style and features* category are descriptive, and can be written in the form of 'X is/are ...' and 'There is/are ...', while the majority of the sentences in the *History and episodes* category are anecdotal and expressed in past tense. For the present study, as a starting point we focus on *Style and features*, which is the most dominant topic in the collected data.

Some of the sentences were comprised of a mixture of different categories. We eliminated such cases, and obtained 206 sentences for the original Japanese source text (**ST-org**).

[4] http://human.cc.hirosaki-u.ac.jp/kokugo/CATtwo/youziyougoziten/youziyougoziten_96_165.pdf

4.2 Simplification process

Rewriting was performed by one of the authors of this paper, who is not a teacher of Japanese language nor trained in writing simplified Japanese. This is preferable, as our presumed writers do not necessarily have such qualifications. ST-org were rewritten according to our simplified Japanese rules in two steps: sentence splitting and further simplification.[5] In principle, sentence splitting covers Rules 1 and 2 and further simplification covers Rules 2 and 3.

Sentence splitting To fulfil Rule 1 of our simplified Japanese, we split the original sentences as required, such that each sentence presents only one idea. For example, 敷鴨居などには白檀が使われ、暖房時には芳香が漂いました。('Sandal wood was used for "Shikikamoi", and it smelled good when heating the room.") was split into two shorter sentences at the location of 'and'. Some splitting operations required the supplementation of linguistic elements, such as subjects and objects, to follow Rule 2. In addition, we tried to utilise the simplest sentence patterns as far as possible. For example, complex predicates such as ～が設置されています ('... is installed') and ～が施されています ('... is in place') have been changed to ～があります ('there is ...'). For the 206 ST-org sentences, we obtained 509 corresponding split sentences (**ST-split**).

Further simplification Based on Rule 2, we further specified a subject for each predicate, and when this was not possible we changed the active voice to the passive voice. For example, 床の間には掛け軸を飾っていました。 has no subject, and a literal translation would be 'Used to display a painting in the alcove.' This was changed to 床の間には掛け軸（絵）が飾られていました。, meaning 'A painting used to be displayed in the alcove.' At this stage, according to Rule 3, we changed the words and expressions such that the sentences consisted as far as possible of only vocabulary and Kanji up to JLPT Grade 2. For example, we changed 採光性に優れています (literally meaning 'excellent in daylighting') to 光がたくさん入ります ('a lot of light enters'). We call this final version of source text **ST-simple**, which consists of 511 sentences. The reason that the number of sentences in ST-simple is slightly larger than that in ST-split is that in rare cases there was a need to further split sentences to simplify them.

[5]Recent studies on building Japanese simplification resources, such as Maruyama and Yamamoto (2018), tend to focus on lexical simplification, a subset of the whole process.

	ST-org		ST-split		ST-simple	
	#	%	#	%	#	%
OOV	1064	21.62	1120	18.76	453	7.58
Grade 2	712	14.47	867	14.52	1220	20.41
Grade 3	371	7.54	529	8.86	620	10.37
Grade 4	1500	30.48	1999	33.48	2110	35.31
F/S	1275	25.90	1456	24.38	1573	26.32
Total	4922	100	5971	100	5976	100

Table 1: Statistics of vocabulary level (OOV > Grade 2 > Grade 3 > Grade 4 > F/S: Functional words/Symbols)

4.3 Vocabulary level of simplified Japanese

Table 1 presents the statistics for the vocabulary levels of words in each of the source versions (ST-org, ST-split and ST-simple). The number of total words increased from ST-org to ST-split, because we supplemented necessary words and did not omit information as far as possible when splitting sentences.

Out-of-vocabulary (OOV) can be regarded as difficult words above the Grade 2 level of JLPT. The ratio of OOV was reduced considerably from ST-split (18.76%) to ST-simple (7.58%), which demonstrates the effect of lexical simplification, although it was not possible to completely eliminate OOV even after manual simplification.

We also observe that the ratio of Grade 2 words considerably increased from ST-split (14.52%) to ST-simple (20.41%). This means that most of the OOV were changed to Grade 2.

5 Experimental setup

We translated the three versions of the Japanese source text using Google Translate,[6] to obtain three versions of English target text: **MT-org**, **MT-split** and **MT-simple**. The resulting English translations were then evaluated by a professional linguist, whose native language is Japanese and who has 10 years of experience in professional Japanese to English translation. The reason we chose a native Japanese speaker was that the Japanese original source sentences are loaded with culture-specific terms that need to be understood without facing a cultural barrier. Furthermore, it was not necessary or desirable to review the translation in terms of

[6]https://translate.google.com/

Good	The information of the source text has been completely translated and there are no grammatical errors in the translation. There may be some unnatural word choices and/or phrasings, but these would not hinder understanding of the meaning.
Fair	There are some minor errors in the translations of less significant parts of the source text, but the meaning of the source text can easily be understood.
Acceptable	Some of the source text is omitted or incorrectly translated, but the core meaning can still be understood with some effort.
Incorrect	Even the core meaning of the source text is not conveyed.
ST unclear	It is impossible to assess the quality of the MT output because of incomprehensible/ambiguous words and/or expressions in the source text.

Table 2: Evaluation criteria

	MT-org		MT-split		MT-simple	
	#	%	#	%	#	%
Good	53	25.73	240	47.15	317	62.04
Fair	8	3.88	26	5.11	37	7.24
Acceptable	17	8.25	35	6.88	49	9.59
Incorrect	65	31.55	114	22.40	76	14.87
ST unclear	63	30.58	94	18.47	32	6.26
Total	206	100	509	100	511	100

Table 3: MT quality

the naturalness or stylistic appropriateness from the viewpoint of a native English speaker.

The 1,226 sentences comprising the three versions were put in a random order to prevent the evaluator from deducing their meanings from the surrounding sentences. We asked the evaluator to rate the quality of the English translations using the five grades shown in Table 2, which are versions of the acceptability evaluation grades used by Goto et al. (2013) modified for the purpose of the present study.

The grade *ST unclear* was added to isolate cases in which the source text contains highly technical terms that lay people, even adult native Japanese speakers, would not understand. In such cases, we may not be able to expect a meaningful evaluation.

The evaluator was also asked to highlight sections in the source and target texts that were incomprehensible, enabling us to qualitatively diagnose the translation difficulties.

6 Results and analyses

6.1 Overall results for MT quality

Table 3 summarises the results of the quality evaluation of the English translation. Approximately 30% of the MT-org sentences are rated as *ST unclear*. The majority of the elements in Japanese source sentences reported as incomprehensible are technical terms relating to architecture or Japanese culture (technical terms related to a tea ceremony, for example). After splitting the sentences, the proportion of *ST unclear* is reduced to less than 20% in MT-split. This is because one or some of the split sentences still contain the same terms, while others become free of them. For MT-simple, only 6.26% are rated as *ST unclear*, because most of the technical terms have been replaced with simpler words or explanatory expressions using simple words.

Simply splitting a sentence to allow each sentence to contain only one idea can double the rate of producing a *Good* translation (25.73% to 47.15%), and employing simple words and expressions can further increase the ratio to 62.04%. Similarly, the percentage for *Incorrect* decreases from 31.55% to 22.40% by splitting the sentences. Further simplification can reduce the percentage of *Incorrect* to 14.87%.

We consider translations with the grades *Good*, *Fair* and *Acceptable* as 'usable', as at least the core meaning of the source text is conveyed. This means that while less than 40% of the ST-org sentences can produce usable translations, approximately 60% of those in ST-split and almost 80% of those in ST-simple can. This result illustrates the high suitability of human-oriented text simplification for MT.

6.2 Analysis of simplification operations

6.2.1 Sentence splitting

Among the 63 MT-org sentences rated as *ST unclear*, there were no cases in which all corresponding MT-split sentences obtained *Good* or *Fair* ratings. This is expected, because as mentioned in Section 6.1 the reasons for incomprehensibility mostly relate to technical terms, which remain even after splitting a sentence.

There are 65 cases in which MT-org sentences received *Incorrect* ratings, and in 12 cases all corre-

ST-org	十字の形でその四方に窓があり日差しを多く取り込めるデザインになっています。
MT-org	It has a window in the form of a cross and has a design that can capture a lot of sunlight.
ST-split	この部屋は十字の形です。/ この部屋の四方には窓があります。/ 日差しを多く取り込めるデザインになっています。
MT-split	**This room** is in the form of a cross. / There are windows on all sides of **this room**. / It is designed to capture a lot of sunshine.

Table 4: Example of the positive effect of sentence splitting

ST-org	天井板には美しい木目を活かした木板が組み合わされています。
MT-org	The ceiling board is a combination of **wood boards that take advantage of beautiful wood grain**.
ST-split	天井板には木板が組み合わされています。/ 美しい木目が活かされています。
MT-split	**A wood board** is combined with the ceiling board. / **Beautiful wood is used**.

Table 5: Example of the negative effect of sentence splitting

ST-org	1 階居間の暖炉には、アールヌーボー風のタイルを使い、ケヤキ材の前飾りがついている。
MT-org	**The fireplace** in the living room on the first floor is made of Art Nouveau-style tiles and decorated with a zelkova wood front decoration.
ST-simple	1 階の居間の暖房には、アールヌーボーのタイルが使われています。/ 飾りは「ケヤキ」という木でできています。
MT-simple	Art Nouveau tiles are used to **heat** the living room on the first floor. / The decoration is made of a tree called “keyaki”.

Table 6: Example of the negative effect of lexical simplification

sponding MT-split sentences obtained *Good* or *Fair* ratings. The main reason for this is that the ill-formedness of sentences is corrected by splitting them into shorter ones. Table 4 presents an example; the complex dependency relations were resolved, and the missing subject この部屋 (‘this room’) was supplemented, as a result of applying Rules 1 and 2, respectively, in the sentence splitting step.

However, there are cases in which splitting the source sentence degrades the quality of the MT outputs. Table 5 presents an example. Here, the sentence was split to prevent the noun 木板 (‘wood boards’) from having the long adjective clause 美しい木目を生かした (‘that take advantage of beautiful wood grain’), which was actually translated correctly in MT-org. In this example, it appears that separating the latter part caused a mistranslation of the relationship between the ‘ceiling board’ and ‘wood boards’. Excessive splitting of a sentence may reduce contextual information within the sentence, leading to the degradation of the MT output.

6.2.2 Further simplification

Among the 208 *Incorrect*/*ST unclear* cases in MT-split, 132 became *Good*/*Fair*/*Acceptable* in MT-simple. The reasons for the majority of the improvements in the MT outputs lie in the rephrasing of technical terms using their hypernyms or explanatory expressions. For example, 袖塀, the name of a special type of wall, has been replaced with 壁, which simply means ‘wall’. In addition, 板透し彫, the name of a special type of decoration, has been replaced with the explanatory expression 木で作った模様, meaning ‘decorations made of wood’. This shows that Rule 3 (Use only the vocabulary and Kanji of up to JLPT Grade 2) is not only beneficial for human readers, but also for MT.

However, there are 32 cases in which further simplification degraded grades from *Good*/*Fair*/*Acceptable* to *Incorrect*/*ST unclear*. Table 6 presents an example of the harmful effect of replacing the term 暖炉 (‘fireplace’) with the presumably simpler term 暖房 (‘heating’). This mistranslation was caused by the equivocality of the word 暖房, which can mean both ‘heating equipment’ and ‘the act of heating’.

Current MT systems have significantly larger vocabularies than those used in human-oriented text simplification. In other words, most general words, even if they are difficult, can be covered by MT systems. In summary, the simplification of rare technical terms is effective for both human and MT applications, but simplifying general words may result in ambiguous words, having an adverse effect on MT.

6.3 Analysis of factors for MT quality

6.3.1 Relation between source sentence characteristics and MT quality

Our motivation for investigating the suitability of text simplification for MT is based on the assumption that long sentences and difficult words can be major factors in degradation in MT quality. Here, we further explore the relation between source sentence characteristics and MT quality.

Table 7 presents correlation scores (Spearman's ρ and Kendall's τ), demonstrating the weak correlations between the MT quality and the numbers of words, characters and OOV in a sentence. The number of OOV is a slightly better indicator than the sentence length for estimating the MT quality.

Figures 1 and 2 present box plots for the sentence length and number of OOV for each MT quality grade. The bold vertical line in each box indicates the median. The majority of *Good*/*Fair*/*Acceptable* MT outputs are produced from source sentences that are no more than 15 words in length and contain no more than two OOV words.

However, some rather long sentences resulted in *Good* quality translations. Table 8 presents an example. In ST-org, the subject この建物 ('the building') only appears once, while there are two predicates 'is ...'. While Japanese sentences often omit the subject, and even change the subject in the middle of a sentence without clearly indicating this change, in this example the subject この建物 is present at the beginning, and is the subject for both predicates. The MT system successfully supplements 'it' to continue the sentence, although it failed to add 'and' before the pronoun. These examples indicate that the source sentences do not necessarily have to be short, so long as they employ grammatically correct subject–predicate combinations.

6.3.2 Remaining difficulties for MT

Finally, we focus on the 76 cases in which MT-simple sentences received *Incorrect* ratings. Referring to the highlighted sections of text that were judged as incomprehensible by the evaluator (see Section 5), we identified a total of 87 critical MT errors, ignoring minor grammatical, orthographic and stylistic errors that do not impair the core meaning of the source text. Based on the MT error taxonomies presented in Costa et al. (2015) and Popovic (2018), we classified the errors as shown in Table 9.

	Spearman's ρ	Kendall's τ
# of words	0.278	0.217
# of characters	0.220	0.170
# of OOV	0.328	0.271

Table 7: Correlations with MT quality

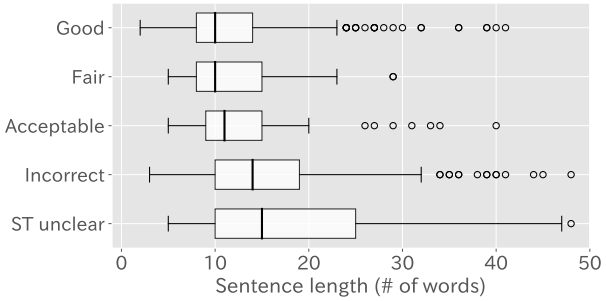


Figure 1: Relation between MT quality and # of words

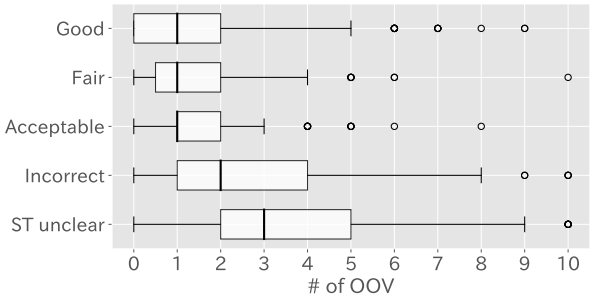


Figure 2: Relation between MT quality and # of OOV

The most frequent error type is the mistranslation of technical terms, including proper nouns. By nature, it is difficult for NMT to correctly handle rare words (Li et al., 2016; Koehn and Knowles, 2017). Although we reduced the technical terms as far as possible through the simplification process, it was impossible to write text on cultural assets without any technical terms. Nevertheless, we can predict possible MT errors if we are aware of the existence of such words, which enables the strategic deployment of post-editing.

The second most frequent error type is the confusion of senses. For example, in many cases 木 was translated as 'tree', although the correct translation was 'wood'. Human translators can easily disambiguate the senses using subtle clues in the text and common knowledge. As detailed contextual information tends to be avoided in simplified text, word

ST-org	この**建物**は、木造総 2 階建て住宅で、細部にはカーペンターゴシック様式の意匠が見られる、19 世紀後半のアメリカ郊外住宅の特色を写した質素な外国人住宅です。
MT-org	**The building** is a two-story wooden house with a carpenter gothic design in every detail, **it** is a frugal foreign house that features the characteristics of an American suburb in the late 19th century.

Table 8: A long ST-org that produced a *Good* MT output

Level 1	Level 2	Level 3	#
Lexis	Mistranslation	Common words	5
		Technical terms	39
	Omission		3
	Untranslated		1
Semantic	Confusion of senses		26
	Mistranslation	Subjects	4
		Others	9

Table 9: Classification of remaining MT errors

sense disambiguation remains a major issue for MT. One solution is domain-adaptation. In a general domain, 'tree' is the most probable translation, while in this particular domain of cultural assets, 'wood' would be the most probable. Thus, retraining MT using in-domain data would be effective if sufficient data is available. Another solution is the use of concrete words. For example, 木材 is likely to be translated as 'wood', as this word has a smaller range of meaning than 木. Although 木材 is more difficult for the target audience than 木, it is still in the vocabulary list for our simplified Japanese.

Although not frequent, the mistranslation (or misidentification) of subjects is noteworthy. For example, 山小屋のような感じがします。 is translated as 'I feel like a mountain hut.' The correct translation is 'It feels like a mountain hut.' In this case, the lack of a subject caused the insertion of the incorrect subject 'I' by the MT system. Although it is possible for human writers to supplement a subject such as これ ('it') or このデザイン ('this design') in the source, repeated use of the same subject may be regarded as unnatural in Japanese. To cope with the incompatibility between source naturalness and machine translatability, we need to incorporate an additional process to further modify the human-oriented simplified source text such that it can contains the necessary subjects to produce a better MT result.

7 Conclusion

In this study, we have proposed a simple rule set for simplified Japanese for human readability, and examined the suitability of simplified text as a source for machine translation (MT). Focusing on expository sentences on Japanese cultural assets, we manually conducted a simplification task in two steps: (1) splitting long sentences into short complete sentences, and (2) further simplifying them. The Japanese-to-English neural MT outputs of the original, split and simplified sentences were manually evaluated in terms of the MT quality.

The experimental results demonstrated the strong potential of human-oriented text simplification for MT purposes, showing that almost 80% of the raw MT outputs achieved a usable quality, among which approximately 80% were of *Good* quality, i.e., the information of the source text was completely translated without grammatical errors. Although the fact that structural and lexical simplification helps to improve the MT quality is not surprising per se, this result reveals the detailed gains we can expect to obtain from simplification.

We also conducted in-depth analyses of the results. The findings can be summarised as follows:

- Splitting sentences is effective when this can resolve ill-formed structures, while excessive splitting may degrade the MT outputs.
- Avoiding rare technical terms is generally effective, while lexical simplification sometimes makes the source text simple but ambiguous.
- Technical terms, word sense ambiguity and a lack of subjects are critical difficulties for MT, which remain even after the text is simplified.

In future work, we intend to tackle the identified difficulties, specifically technical terms and lacking subjects. For technical terms, we plan to develop a tool to generate alternative expressions, such as hypernyms and explanatory phrases. For lacking subjects, we will introduce a semi-automatic process to add subjects necessary only for MT.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP17K00466 and JP19K20628.

References

Takako Aikawa, Lee Schwartz, Ronit King, Monica Corston-Oliver, and Carmen Lozano. 2007. Impact of controlled language on translation quality and post-editing in a statistical machine translation environment. In *Proceedings of the Machine Translation Summit XI*, pages 1–7, Copenhagen, Denmark.

ASD. 2017. ASD Simplified Technical English. Specification ASD-STE100, Issue 7. `http://www.asd-ste100.org`.

Ângela Costa, Wang Ling, Tiago Luís, Rui Correia, and Luísa Coheur. 2015. A linguistically motivated taxonomy for machine translation error analysis. *Machine Translation*, 29(2):127–161.

Isao Goto, Ka Po Chow, Bin Lu, Eiichiro Sumita, and Benjamin K. Tsou. 2013. Overview of the Patent Machine Translation Task at the NTCIR-10 Workshop. In *Proceedings of the 10th NTCIR Conference*, pages 260–286, Tokyo, Japan.

Anthony Hartley, Midori Tatsumi, Hitoshi Isahara, Kyo Kageura, and Rei Miyata. 2012. Readability and translatability judgments for 'Controlled Japanese'. In *Proceedings of the 16th Annual Conference of the European Association for Machine Translation (EAMT)*, pages 237–244, Trento, Italy.

Bui Thanh Hung, Nguyen Le Minh, and Akira Shimazu. 2012. Sentence splitting for Vietnamese-English machine translation. In *Proceedings of the 4th International Conference on Knowledge and Systems Engineering (KSE)*, pages 156–160, Danang, Vietnam.

Isao Iori. 2016. The enterprise of Yasashii Nihongo: For a sustainable multicultural society in Japan. *Jinbun-Shizen Kenkyu*, (10):4–19.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the 1st Workshop on Neural Machine Translation (NMT)*, pages 28–39, Vancouver, Canada.

Xiaoqing Li, Jiajun Zhang, and Chengqing Zong. 2016. Towards zero unknown word in neural machine translation. In *Proceedings of the 25th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2852–2858, New York, USA.

Takumi Maruyama and Kazuhide Yamamoto. 2018. Simplified corpus with core vocabulary. In *Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC)*, pages 1153–1160, Miyazaki, Japan.

Rei Miyata, Anthony Hartley, Cécile Paris, Midori Tatsumi, and Kyo Kageura. 2015. Japanese controlled language rules to improve machine translatability of municipal documents. In *Proceedings of the Machine Translation Summit XV*, pages 90–103, Miami, Florida, USA.

Sharon O'Brien and Johann Roturier. 2007. How portable are controlled language rules? In *Proceedings of the Machine Translation Summit XI*, pages 345–352, Copenhagen, Denmark.

Maja Popovic. 2018. Error classification and analysis for machine translation quality assessment. In Joss Moorkens, Sheila Castilho, Federico Gaspari, and Stephen Doherty, editors, *Translation Quality Assessment: From Principles to Practice*. Springer.

Horacio Saggion. 2017. *Automatic Text Simplification*. Morgan & Claypool.

Matthew Shardlow. 2014. A survey of automated text simplification. *International Journal of Advanced Computer Science and Applications: Special Issue on Natural Language Processing*, 4(1):58–70.

Advaith Siddharthan. 2014. A survey of research on text simplification. *International Journal of Applied Linguistics: Recent Advances in Automatic Readability Assessment and Text Simplification*, 165(2):259–298.

Hideki Tanaka, Hideya Mino, Shinji Ochi, and Motoya Shibata. 2013. News services in simplified Japanese and its production support systems. In *Proceedings of the International Broadcasting Convention 2013 (IBC)*, Amsterdam, The Netherlands.

Sanja Štajner and Maja Popovic. 2016. Can text simplification help machine translation? *Baltic Journal of Modern Computing*, 4(2):230–242.

Sanja Štajner and Maja Popovic. 2018. Improving machine translation of English relative clauses with automatic text simplification. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 39–48, Tilburg, The Netherlands.

Building Cendana: a Treebank for Informal Indonesian

David Moeljadi
Department of Asian Studies, Faculty of Arts
Palacký University Olomouc
Czechia
davidmoeljadi@gmail.com

Aditya Kurniawan, Debaditya Goswami
NLP, Vision, Speech (NVS)
Traveloka Services Pte. Ltd.
Singapore
{akurniawan,
debaditya.goswami}@traveloka.com

Abstract

This paper introduces Cendana, a treebank for informal Indonesian. The corpus is from a subset of online chat data between customer service staff and customers at Traveloka (traveloka.com), an online travel agency (OTA) from Indonesia that provides airline ticketing and hotel booking services. Lines of conversation text are parsed using the Indonesian Resource Grammar (INDRA) (Moeljadi et al., 2015), a computational grammar for Indonesian in the Head-Driven Phrase Structure Grammar (HPSG) framework (Pollard and Sag, 1994; Sag et al., 2003) and Minimal Recursion Semantics (MRS) (Copestake et al., 2005). The annotation was done using Full Forest TreeBanker (FFTB) (Packard, 2015). Our purpose is to create a treebank, as well as to develop INDRA for informal Indonesian. Testing on 2,000 lexically dense sentences, the coverage is 64.1% and 715 items or 35.8% was treebanked, with correct syntactic parses and semantics. INDRA has been developed by adding 6,741 new lexical items and 22 new rules, especially the ones for informal Indonesian. The treebank data was employed to build a Feature Forest-based Maximum Entropy Model Trainer. Testing against the annotated data, the precision was around 90%. Moreover, we leveraged the treebank data to develop a POS tagger and present benchmark results evaluating the same.

1 Introduction

This work is an attempt to build a new open resource for colloquial/informal Indonesian annotated corpus or a treebank, i.e. a linguistically annotated corpus/text data that includes some grammatical analyses, such as parts-of-speech, phrases, relations between entities, and meaning representations. The existing treebanks for Indonesian are mainly for formal Indonesian, e.g. manually tagged Indonesian corpus (Dinakaramani et al., 2014) and JATI (Moeljadi, 2017). Thus, building a treebank for informal Indonesian can be considered as a pioneer. This treebank is named Cendana, the Indonesian word for "sandalwood", built using tools developed in the Deep Linguistic Processing with HPSG (DELPH-IN) community.[1] This paper describes the construction of this new language resource and gives new analyses and implementations on phenomena in informal Indonesian morphology and syntax.

2 Sociolinguistic situation in Indonesia

Indonesian (ISO 639-3: ind), called bahasa Indonesia (lit. "the language of Indonesia") by its speakers, is spoken mainly in the Republic of Indonesia by around 43 million people as their first language and by more than 156 million people as their second language (2010 census data). The lexical similarity is over 80% with Standard Malay (Lewis, 2009). It is written in Latin script. Morphologically, Indonesian is a mildly agglutinative language. It has a rich affixation system, including a variety of prefixes, suffixes, circumfixes, and reduplications. The basic word order is SVO (Sneddon et al., 2010).

The diglossic nature of the Indonesian language exists from the very beginning of the historical

[1] http://www.delph-in.net

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 156-164, Hakodate, Japan, September 13-15, 2019

record when it is called Old Malay around the 7th century to the present day (Paauw, 2009). While much attention has been paid to the development and cultivation of the standard/formal "High" (H) variety of Indonesian, little attention has been particularly paid to describing and standardizing the informal "Low" (L) variety. Sneddon (2006) calls this variety "Colloquial Jakartan Indonesian" and states that it is the prestige variety of colloquial Indonesian in Jakarta, the capital city of Indonesia, and is becoming the standard informal style. In addition to this L variety, more than 500 regional languages spoken in Indonesia, such as Javanese, Balinese, and various local Malay languages, add to the complexity of the sociolinguistic situation in Indonesia.

The H variety is used in the context of education, religion, mass media, and government activities. The L variety is used for everyday communication. The regional vernaculars or *bahasa daerah* are used for communication at home with family and friends in the community. In this paper, the term 'informal Indonesian' or L variety refers to Colloquial Jakartan Indonesian mentioned above.

3 Traveloka Conversational Corpus

We use more than 10 millions of lines of conversation or chat data between Traveloka users and customer service agents. We have more varieties in terms of language registers in the chat data, compared with other commonly used text for corpus such as newspaper and Wikipedia articles. The customer service agents usually write in H variety, while the users or customers usually write in L variety. Many informal features which can be found in online written text such as in tweets (Le et al., 2016), also appear in the chat data. They are informal words, abbreviations, typos, discourse particles, interjections, foreign words, emojis, emoticons, and unusual word orders, as shown in Table 1.

The raw data is mainly in H and L varieties of Indonesian or Indonesian with some English words related to flights, hotels, bookings, and payments such as "booking", "check-in", "form", "payment" and sometimes they appear together with Indonesian affixes, e.g. *formnya* "the form". Very few chat lines are written entirely in foreign languages, such as English, Malay, Javanese,[2] Vietnamese, Tagalog, and German. Traveloka is expanding to countries in Southeast Asia and Australia and thus, we got chat data in various languages. In addition to the informal features, the raw data has been processed to mask sensitive information such as email addresses, phone numbers, and booking codes/numbers. The data preprocessing is described in Section 5.1. It includes text normalization, sentence segmentation (chunking the chat data into sentences), and word tokenization (chunking a sentence into words).

4 Related work

There are few open-source treebanks for Indonesian, annotated with both syntactic and semantic information. Most previous work on Indonesian treebanks focuses on the H variety and on syntactic annotation, rather than semantic annotation, e.g. the Indonesian Dependency Treebank developed by Charles University in Prague (Green et al., 2012), with manually annotated dependency structures for Indonesian; the Indonesian treebank developed by the University of Indonesia (UI) (Dinakaramani et al., 2014) which uses a part-of-speech (POS) tagged corpus as a starting point and adopts Penn Treebank bracketing guidelines; and the Indonesian treebank in the Asian Language Treebank (ALT) which was built by the Agency for the Assessment and Application of Technology (BPPT) (Riza et al., 2016), comprises about 20,000 sentences originally sampled from the English Wikinews in 2014, and uses tools such as POS tagger, syntax tree generator, shallow parser, and word alignment. The Indonesian Treebank in the ParGram Parallel Treebank (ParGramBank) (Sulger et al., 2013) is based on Lexical Functional Grammar (LFG) (Kaplan and Bresnan, 1982; Dalrymple, 2001) and publicly available via the INESS treebanking environment but contains only 79 sentences and 433 words.

Similar to the Indonesian Treebank in ParGramBank, another treebank called JATI (Moeljadi, 2017) was built based on a computational grammar for Indonesian called the Indonesian Resource Grammar (INDRA) (Moeljadi et al., 2015).[3] The raw cor-

[2]Regional languages such as Javanese are treated as foreign languages in this paper.

[3]http://moin.delph-in.net/IndraTop

Feature	Example
Informal word	*gak* (*tidak* "NEG"), *mulu* (*melulu* "only, just"), *uda* (*sudah* "PERF"), ...
Abbreviation	*sy* (*saya* "1SG"), *cm* (*cuma* "only"), *yg* (*yang* "REL"), *jg* (*juga* "too"), ...
Typo	*tikey* (*tiket* "ticket"), *abntu* (*bantu* "help"), *sata* (*saya* "1SG"), *fi* (*di* "at"), ...
Discourse particle	*koq*, *lho*, *nich*, *yach*, *sich*, *donk*, *deh*, *kek*, *mah*, *nah*, *tuh*, *yuk* ...
Interjection	*hahaha* (*haha* "ha-ha"), *wkwkwk* (*haha* "ha-ha"), *hehehe*, *hihi*, *wowww*, ...
Foreign word	within (English), *semakan* (Malay), *ngono* (Javanese), *trong* (Vietnamese), *maawain* (Tagalog), ...
Emoji/emoticon	:), :(, :-|, ^_^

Table 1: Informal features in Traveloka chat data

pus data are dictionary definition sentences related to food and beverages, extracted from the official Indonesian dictionary (KBBI) fifth edition. INDRA is open-source and it is developed within the framework of HPSG and MRS, using tools and resources developed by the DELPH-IN research consortium. The creation of Cendana is similar to the one of JATI but deals with both H and L varieties. Cendana uses INDRA to parse the data. During the treebank development, INDRA was developed with informal lexicon, morphology, and syntax rules (see Section 5.4). Similar to JATI, Cendana uses an approach called "parse and select by hand", in which lines of corpus data are parsed and the annotator selects the best parse from the full analyses derived by the grammar.

5 Treebank development

Treebanking is a part of grammar development process (Bender et al., 2011), as shown in Figure 1. The motivation is to develop a broad-coverage grammar together with the treebank, which allows the grammar developer to immediately identify problems in the grammar and the treebanker to improve the quality of the treebank (Oepen et al., 2004). The process starts from preparing the corpus data or test-suite.

Section 5.1 describes the data preprocessing part before creating the test-suite, which is mentioned in Section 5.2. Afterwards, the lexical acquisition, linguistic type classification, linguistic phenomena analysis, and implementation, are described in Section 5.3 and Section 5.4. Lastly, the annotation/treebanking part is written in Section 5.5.

5.1 Data preprocessing

Data preprocessing includes text normalization, sentence segmentation, and word tokenization, as illustrated in Figure 2.

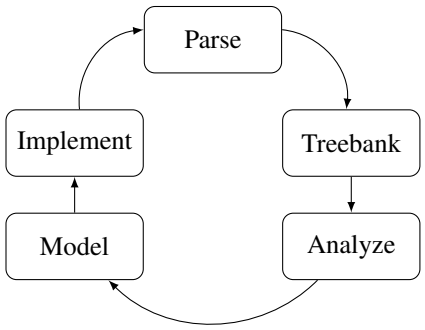


Figure 1: Grammar engineering spiral

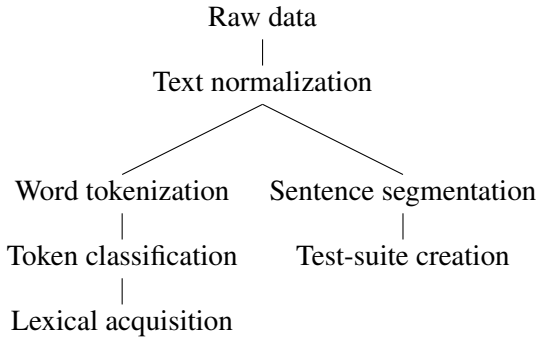


Figure 2: Data preprocessing and lexical acquisition

Text normalization: In order to ensure privacy of any user data within the linguistic corpus outlined in Section 3, we encoded email addresses into a token EMAIL, phone numbers into a token PHONE_NUMBER, website addresses into a token SITE, URIs into a token URI, image into a token image, @ sign into a token AT, and booking numbers into a token NUMBER. We normalized repetitive punctuations, removed spaces in abbreviations and within a single token, added spaces between numerals and nouns, removed excessive characters, con-

verted Unicode into ASCII characters, and removed non-printable characters like emoticons.

Sentence segmentation and word tokenization: We used Python 3 and Natural Language Toolkit (NLTK) (Bird et al., 2009) for sentence segmentation and word tokenization. After that, we counted the number of sentences and tokens. We had 13,372,929 sentences and 111,175,597 tokens. There are duplicates of sentences and tokens, thus we counted the number of unique sentences and tokens, too. There are 8,527,072 unique sentences (63.8% of total number of sentences) and 693,718 unique tokens (0.6% of total number of tokens).

5.2 Test-suite creation

For the purpose of building Cendana, only a representative subset of the chat data having the most lexically dense tokens, is extracted. We extracted a sample of data consisting of two thousand sentences having at least ten tokens in a sentence. The lexical density is measured by dividing the number of lexical word tokens (tokens written in alphabet other than stop words and foreign words) by the number of all tokens. We used NLTK for stopwords and added more stopwords from spaCy.[4] Since the available sources for stopwords are for formal Indonesian, we added more stopwords for informal Indonesian.

We made a test-suite, i.e. a sample of text, selected and formatted for treebanking. The format is explained in the DELPH-IN page.[5] Each line in the test-suite consists of an ID number, a sentence, the number of tokens in that sentence, an optional comment, and information on author and date.

5.3 Linguistic type classification and lexical acquisition

After word tokenization with NLTK, we extracted 63,294 tokens (0.09% of the total number of unique tokens) which have at least two characters and have frequency more than ten. Before lexical acquisition from the chat data, INDRA had 16,751 lexical items. Out of 63,294 unique tokens extracted, 3,059 tokens were already in INDRA's lexicon. Thus, there is a potential to add more lexical items, especially the informal ones, into INDRA. We did lexical acquisition firstly for tokens having a circumfix *pe-...-an*, *ke-...-an*, enclitic *-nya*, *-ku*, and those with reduplication (marked with a hyphen). These tokens are usually nouns. Afterwards, we added tokens having a prefix *me-*, *di-*, *nge-*, and a suffix *-kan* and *-in*. These tokens are usually verbs. This lexical acquisition process was not done at once, instead it was done throughout the treebanking project, before and during treebanking. During lexical acquisition, we grouped the tokens based on lexical types in INDRA, e.g. inanimate noun, temporal noun, intransitive verb, ditransitive verb, and transitive verb with an optional or obligatory complement.

We keep in mind that the same semantic predicate is applied to the lexical items having the same concept, regardless their varieties (H or L). For example, the negation word with non-nominal predicates is *tidak* "NEG". Sneddon (2006) notes this as a word which mostly appears in the H variety. He lists six counterparts of it in the L variety: *enggak*, *nggak*, *ngga*, *gak*, *kagak*, and *ndak*. We found 32 more variants in the data, including abbreviations and typos: *nda*, *nd*, *dk*, *nfk*, *ndk*, *tda*, *tijdvak*, *tidaj*, *tidar*, *tidk*, *tida*, *tdak*, *tdk*, *tidsk*, *ngaak*, *ngaa*, *nggaj*, *nggah*, *nggal*, *nggk*, *ngg*, *ngak*, *ngal*, *nga*, *ngk*, *ngx*, *ngakk*, *kgk*, *gag*, *ga*, *gk*, and *g*. All these 39 lexical items, although they are orthographically different, have the same concept semantically and thus, they are given the same MRS semantic predicate. After lexical acquisition, INDRA has 7,181 more lexical items, thus the total number of lexical items in INDRA became 23,932.

5.4 Linguistic phenomena analysis and implementation in INDRA

Linguistic phenomena in the test-suite are identified and analyzed based on reference grammars and other linguistic literature. The analyses are modeled in HPSG and implemented in INDRA.

Text normalization in INDRA: Beside text normalization mentioned in Section 5.1, we did more detailed text normalization using INDRA, dealing with typos, morphology, and token boundaries (see Table 3). In addition, we added more regular expression patterns to detect dates and currencies.

Active voice prefixes: Formal Indonesian has transitive verbs in active voice which take prefix

[4] https://github.com/explosion/spaCy/blob/master/spacy/lang/id/stop_words.py

[5] http://moin.delph-in.net/ItsdbReference

Action	Example	
	Before	After
Normalize repetitive punctuation	,,,+++!!!	,+!
Remove spaces	␣e␣-tiket␣␣a␣.␣n␣␣Mr␣.␣John	e-tiket a.n Mr.John
	Rp␣800000␣,-␣01␣/␣09	Rp800000,- 01/09
Add spaces	30Juni 2org 1anak 1kmr 2bed 1hr	30 Juni 2 org 1 anak 1 kmr 2 bed 1 hr
Remove excessive characters	ruangannnya hahahaha	ruangannya haha
Encode emails etc	abc@abc.com, http://abc, www.ab.co,	EMAIL, SITE, URI,
into respective tokens	+62-1234-5678, image.jpg, @	PHONE_NUMBER, IMAGE, AT

Table 2: Text normalization

Action	Example	
	Before	After
Fix words	*bantuaanya*, *danannya*, *kodebya*	*bantuannya*, *dananya*, *kodenya*
-nya as a separate token	*hotel nya*, *uang ny*, *tiket nyq*, *namax*	*hotel -nya*, *uang -nya*, *tiket -nya*, *nama -nya*
Fix token boundary	*kal omau di gantii tiketsaya 7an CGKPDG*	*kalo mau diganti tiket saya tujuan CGK PDG*

Table 3: Text normalization in INDRA

meN-, where N symbolizes a nasal which assimilates to the first sound of the verb stem. Moeljadi et al. (2015) show how this is dealt with in INDRA, in terms of morphological rule and inflectional rule. In informal Indonesian, the situation is more complex, Sneddon (2006) notes there are four possibilities:

- without any prefix
- with prefix *meN-*, as in formal Indonesian
- prefix *N-*, just drop the *me*, except for stems started with *c* and *per*
- prefix *nge-*, which occurs before all initial consonants except *p*, *t*, *s*, *c*, *k* if the stems have more than one syllable. The initial *h* is often lost. Prefix *nge-* occurs before *p*, *t*, *s*, *c*, *k* when the stems are one-syllable or the stems are borrowings, either assimilated or unassimilated borrowings.

In addition to these four possibilities, we found another one in our chat text data:

- prefix *m(N)-*

Table 4 shows these five possibilities with examples. We analyzed the patterns and implemented the rules. INDRA's lexicon lists down only the stems or the forms without prefixes.

Using the morphological and inflectional rules, INDRA can parse and generate all surface forms both with and without prefixes. All surface forms having different surface forms but derived from the same stem, have the same MRS semantic predicate. For example, *proses*, *memproses*, *mproses*, *mroses*, *ngeproses* have the same semantic predicate `_proses_v_rel`.

Because of this, given a formal sentence as input, INDRA can generate all informal sentences. For example, given an input: *Traveloka memproses pesanan saya* "Traveloka processes my booking", INDRA can generate the outputs: *Traveloka proses pesanan saya*, *Traveloka ngeproses pesanan saya*, *Traveloka mproses pesanan saya*, *Traveloka mroses pesanan saya*, ...

Compound rules for proper names Two rules for proper name (PROPN) compound were made. The first was given an underspecified semantics predicate because this type of compound can have a different meaning in different context, similar to a noun-noun compounds which are often highly ambiguous and thus, it seems necessary to have a large degree of 'world knowledge' to understand them (Ó Séaghdha, 2007).

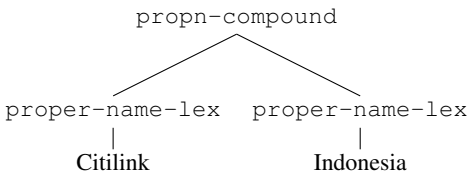
It may have a semantic relation IN, e.g. *CGK JKT* and *PLM Palembang* as in *rute CGK JKT ke PLM Palembang* "the route (from) CGK (airport) (IN) JKT (Jakarta) to PLM (airport) (IN) Palembang"; it may also have a semantic relation SPECIFICALLY, e.g. *Surabaya Juanda* as in *menuju Surabaya Juanda* "towards Surabaya SPECIFICALLY Juanda (airport)"; another possibility is a semantic relation BELONG, e.g. *Citilink Indonesia* as in *maskapai penerbangan Citilink Indonesia* "Citilink airline (which BELONGS TO) Indonesia"; and the last one is a relation which connects name parts e.g. *F Budi Warsito*. Similar to noun-noun compound analysis and implementation in INDRA (Moeljadi, 2018), the underspecified semantics is represented by `compound_p_rel`, which

stem	without prefix	*meN-*	*m(N)-*	*N-*	*nge-*
p-initial (also m-initial)	*panggil* "call"	*memanggil*	*mmanggil*	*manggil*	(none)
b-initial	*bantu* "help"	*membantu*	*mbantu*	*mbantu*	*ngebantu*
t-initial (also n-initial)	*tunggu* "wait"	*menunggu*	*mnunggu*	*nunggu*	(none)
d-initial (also j-initial)	*dapat* "get"	*mendapat*	*mdapat*	*ndapat*	*ngedapat*
c-initial	*cuci* "wash"	*mencuci*	*mcuci*	*nyuci*	(none)
s-initial (also ny-initial)	*sewa* "rent"	*menyewa*	*mnyewa*	*nyewa*	(none)
k-initial (also ng-initial)	*kirim* "send"	*mengirim*	*mngirim*	*ngirim*	(none)
g-initial	*ganti* "replace"	*mengganti*	*mganti*	*ngganti*	*ngeganti*
h-initial	*hitung* "count"	*menghitung*	*mhitung*	*ngitung*	*ngehitung*
l-initial (also r-initial)	*lempar* "throw"	*melempar*	*mlempar*	*nglempar*	*ngelempar*
vowel initial	*ambil* "take"	*mengambil*	*mngambil*	*ngambil*	(none)
borrowing	*proses* "process"	*memproses*	*mproses*	*mroses*	*ngeproses*
one syllable	*cek* "check"	*mengecek*	*mngecek*	(none)	*ngecek*

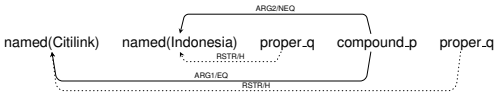
Table 4: Morphology process of active voice prefixes

takes two proper names as its arguments, as shown in (1).

(1) a.

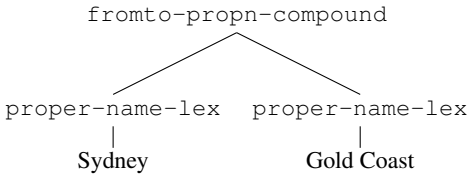


b.

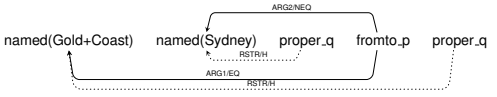


The second one has a special semantics predicate for directions (FROM one place TO another place) and appears a lot in the data, e.g. *pesawat JOG CGK* "plane FROM JOG (airport) TO CGK (airport)", also *jur sydney gold coast* "direction FROM Sydney TO Gold Coast", as illustrated in (2).

(2) a.



b.



In addition to the morphology of active voice prefixes and compound rules for proper names mentioned above, new syntactic rules, e.g. imperatives and a `head-subject` rule for informal Indonesian, as well as discourse particles, were added.

5.5 Annotation

The treebanking process was done semi-automatically using an approach called "parse and select by hand" or "discriminant-based treebanking". It is a grammar-based corpus annotation, using INDRA to parse and select or reject discriminants or possible readings until one (best) parse remains. The discriminant-based treebanking produces all syntactic and semantic parses which are grammatical and consistent, it gives feedback to INDRA, and if there's some changes or updates in the grammar, it is easy to update the treebank. However, its coverage is restricted by the computational grammar (INDRA). A treebanking tool called Full Forest TreeBanker (FFTB) (Packard, 2015) was used to select the best tree with correct syntactic and semantic parse from the 'forest' of possible trees proposed by INDRA for each sentence, and store it into a database that can be used for statistical ranking of candidate parses.

The test-suite is parsed using INDRA and then the first author as the only annotator selects the correct analysis (or rejects all analyses) using FFTB. The system selects features that distinguish between different parsers and the annotator selects or rejects the features until only one parse is left. The choices made by the annotators are saved and thus, it is possible to update the treebank when the grammar changes (Oepen et al., 2004). If a sentence is ungrammatical or if INDRA cannot parse the sentence, no discriminants will be found. However, if a sentence is grammatical and no correct tree is found, all the possible trees should be rejected and the grammar has to be modified or debugged. Sentences for which no analysis had been implemented in the grammar or which fail to parse are left unannotated.

Using FFTB, we can note some interesting findings or linguistic analyses item by item. During the treebanking process, new words, especially informal words, and new rules were added into INDRA, so that INDRA can parse informal Indonesian sentences. Some phenomena in colloquial Indonesian were analyzed (see Section 5.4).

6 Result and evaluation

Cendana can be evaluated by measuring the number of coverage, i.e. how many sentences or how many percent of total sentences INDRA can parse and how many of them are good (having correct parse trees and semantics).

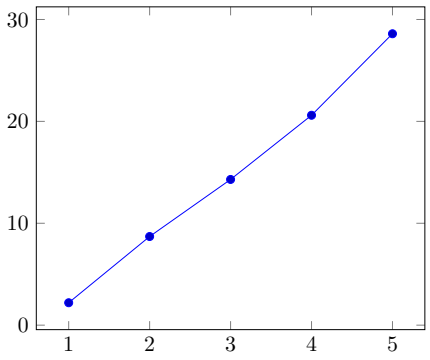


Figure 3: Evolution of coverage for the first 100,000 items (x axis = stage, y axis = coverage)

Figure 3 shows the increase in coverage of the first 100,000 items in the chat data from stage one to stage five. Out of 100,000 first items, 73.3% or 73,280 items are unique. At the **first stage**, we did coverage test before data preprocessing. The result is only 1,614 items or 2.2% could be parsed by INDRA. We got an increase to 8.7% (**second stage**) after lexical acquisition for tokens with affixes *pe-...-an*, *ke-...-an*, *-nya*, *meN-*, *di-* with frequency above 10. We got 14.3% coverage at the **third stage** after lexical acquisition for tokens with *-kan*, *N-*, *-in*, adding morphological rules for active voice prefixes and text normalization in INDRA. At the **fourth stage**, after adding compound rules for proper names and lexical acquisition for other tokens having frequency more than 1000, the coverage increased to 20.6%. At the **fifth stage**, we added more tokens which appear more than 100, including typos, as well as regular expressions for dates and time, discourse particles, and got a coverage of 28.6%. At this point, we began to make a test-suite for 2,000 representative items (see Section 5.2).

Testing INDRA on this full set of 2,000 items at the **initial stage** gave a coverage of 12.9%, as illustrated in **Figure 4**. We added rules for imperatives and added more words, and got 16.8% coverage at the **second stage**. The first big increase in coverage to 36% (**third stage**) was from lexical acquisition. At this point, we started treebanking 20 items or sentences. We kept doing lexical acquisition when treebanking and got 42% coverage with 37 items treebanked at the **fourth stage**. At the present stage (**seventh stage**), testing INDRA on the set of 2,000 items from Cendana test-suite gave a coverage of 64.1% with 715 items treebanked. INDRA has been developed too. The number of lexical items has increased

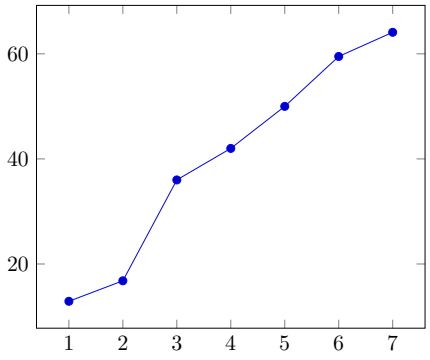


Figure 4: Evolution of coverage for 2,000 items (x axis = stage, y axis = coverage)

from 16,751 before lexical acquisition in October 2018 to 23,932 after 715 items were treebanked in March 2019. Similarly, the number of types has raised from 2,057 to 2,130; the number of lexical rules from 12 to 24; and the number of grammar rules from 63 to 85.

The treebanking result is stored in a directory consisting of several text files. The result file contains a derivation tree/phrase structure tree, node labels or POS tags from the phrase structure tree, and a MRS semantics representation for each annotated item. They can be easily edited to accommodate the changes made in INDRA. We made some of the treebank data (552 sentences/items) publicly available, licensed under the GNU General Public License, version 2 for researchers to develop Indonesian NLP.[6] We documented the treebanking process.

We run Feature Forest-based Maximum Entropy Model Trainer, using a tool developed in DELPH-IN[7] based on Miyao and Tsujii (2002). We used the model to treebank 1,000 sentences (number 9000 to 9999) automatically. The result was promising: 428 sentences could be treebanked automatically. We checked the model against the 1,000 data which contain manually annotated items. The result was 612 sentences could be treebanked automatically and the precision was around 90%.

The initial effort to leverage the treebank is by testing it for POS Tagging task. Due to the small amount of data in recent treebank, we leverage Wikipedia and manually tagged Indonesian corpus from UI (Dinakaramani et al., 2014) as our training set and use the treebank as our golden test data. The Wikipedia that we use comes from universal dependency (UD) project (Nivre et al., 2016).

[6] https://github.com/davidmoeljadi/INDRA/tree/master/tsdb/gold/Cendana

[7] http://moin.delph-in.net/FeatureForestTrainer

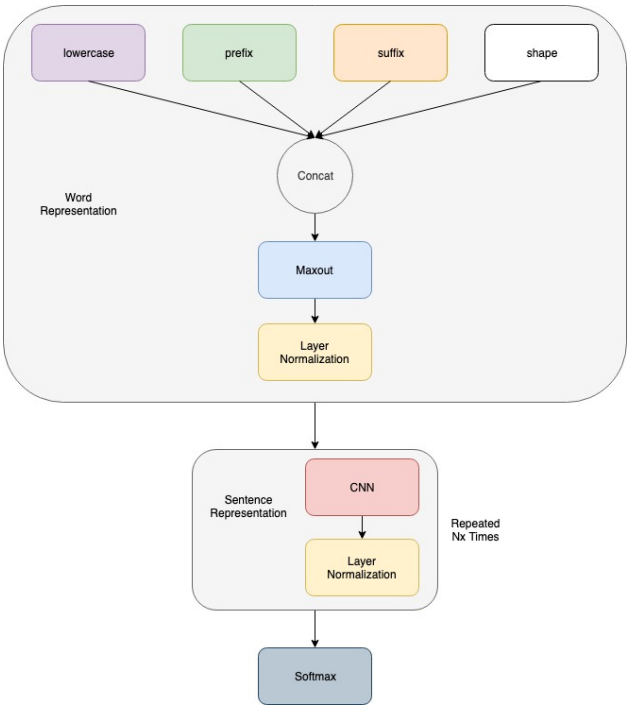


Figure 5: Machine learning model

Train data	Test data	f1-score	OOV in test
UD	UD	92.939	34.33%
UI	UI	97.630	21.68%
UD	Cendana	53.313	68.68%
UI	Cendana	52.168	71.66%

Table 5: POS Tagger experiment

We split the experiment in two folds: we set up the baseline using only UD and UI for all train, validation and test data to see the performance of the model in the same domain, and later we used the training data from UD and UI in order to make prediction on the treebank.

The machine learning models we use to do the training and inference are off-the-shelf model from spaCy. In brief, the algorithm used by spaCy is neural network based with the architecture depicted in Figure 5. The architecture consists of combining multiple features from the word such as lower case, prefix, suffix and shape embedding. Shape embedding is a transformation process by replacing numbers with token `d` and capital words with `w`. The embedding was later concatenated and used as an input to maxout layer. The result was then normalized with layer normalization. After normalization, the output was forwarded to CNN before getting the probability of the tags in softmax layer.

Type of errors	Example
Names	ulfah, mega, subagyo, hadi, heryanto, setiawan, rahayu
Typos	passanger, pkanbaru, rescedule, pembayarsn, soekarna, trransfer, tikcet
Unprocessed numerics	11-12, 20.20, 6.20, 11.10, 29-11-20, 2017, 12.25
Cases (uppercase/ lowercase)	Pemesan, Airliner, Booking, TRINUSA, CGK, DENGAN, Simpati
Abbreviations	tlpn, jog, cgk, jogja, kenapa, kmrn, sya
Tokens	EMAIL, DATE, URL, NUMBER, PHONE, SITE

Table 6: OOV Examples

The number of out-of-vocabulary (OOV) affects the performance of the model quite significantly, as shown in Table 5. We classify the OOV into six different types (see Table 6). We assume that the performance of the model could be improved by adding more data from Cendana.

7 Summary

This paper has described the construction of Cendana treebank, created from a subset of Traveloka chat data, parsed using INDRA, and annotated using FFTB. The construction of Cendana improved the development of INDRA with lexical items and rules for informal Indonesian. At the present stage, the coverage is 64.1% and 35.8% was treebanked, with correct syntactic parses and semantics (715 out of 2,000 items). The treebank was employed to build a Feature Forest-based Maximum Entropy Model Trainer and to develop a POS tagger. The results were promising. Adding more treebank data could improve the performance of the model. Cendana is available on GitHub, under the GNU General Public License.

Acknowledgments

We gratefully acknowledge the support of Traveloka (PT Trinusa Travelindo and Traveloka Services Pte. Ltd.) and the NLP, Vision and Speech department where this research was conducted, and for giving us the opportunity to work with the chat data. The first author also gratefully acknowledges the support of the European Regional Development Fund-Project "Sinophone Borderlands – Interaction at the Edges" CZ.02.1.01/0.0/0.0/16_019/0000791.

References

Emily M. Bender, Dan Flickinger, and Stephan Oepen. 2011. Grammar Engineering and Linguistic Hypothesis Testing: Computational Support for Complexity in Syntactic Analysis. In *Language from a Cognitive Perspective: Grammar, Usage and Processing*, pages 5–29. CSLI Publications, Stanford.

Steven Bird, Edward Loper, and Ewan Klein. 2009. *Natural Language Processing with Python*. O'Reilly Media Inc.

Ann Copestake, Dan Flickinger, Carl Pollard, and Ivan A. Sag. 2005. Minimal Recursion Semantics: An Introduction. *Research on Language and Computation*, 3(4):281–332.

Mary Dalrymple. 2001. Lexical-Functional Grammar. *Syntax and Semantics*, 34. Academic Press.

Arawinda Dinakaramani, Fam Rashel, Andry Luthfi, and Ruli Manurung. 2014. Designing an Indonesian part of speech Tagset and Manually Tagged Indonesian Corpus. In *2014 International Conference on Asian Language Processing (IALP)*, pages 66–69. IEEE.

Nathan Green, Septina Dian Larasati, and Zdeněk Žabokrtský. 2012. Indonesian Dependency Treebank: Annotation and Parsing. In *26th Pacific Asia Conference on Language, Information and Computation*, pages 137–145.

Ronald Kaplan and Joan Bresnan. 1982. Lexical Functional Grammar: A formal system for grammatical representation. In *The Mental Representation of Grammatical Relations*, pages 173–281. the MIT Press, Cambridge.

Tuan Anh Le, David Moeljadi, Yasuhide Miura, and Tomoko Ohkuma. 2016. Sentiment analysis for low resource languages: A study on informal Indonesian tweets. In *Proceedings of the 12th Workshop on Asian Language Resources*, pages 123–131.

M. Paul Lewis. 2009. *Ethnologue: Languages of the World*. SIL International, Dallas, Texas, 16 edition.

David Moeljadi, Francis Bond, and Sanghoun Song. 2015. Building an HPSG-based Indonesian Resource Grammar (INDRA). In *Proceedings of the Grammar Engineering Across Frameworks (GEAF) Workshop, 53rd Annual Meeting of the ACL and 7th IJCNLP*, pages 9–16.

David Moeljadi. 2017. Building JATI: A Treebank for Indonesian. In *Proceedings of The 4th Atma Jaya Conference on Corpus Studies (ConCorps 4)*, pages 1–9, Jakarta.

David Moeljadi. 2018. *An Indonesian resource grammar (INDRA): and its application to a treebank (JATI)*. Ph.D. thesis, Nanyang Technological University.

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan T McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *LREC*.

Diarmuid Ó Séaghdha. 2007. Annotating and learning compound noun semantics. In *Proceedings of the 45th Annual Meeting of the ACL: Student Research Workshop*, pages 73–78. Association for Computational Linguistics.

Stephan Oepen, Dan Flickinger, and Francis Bond. 2004. Towards holistic grammar engineering and testing–grafting treebank maintenance into the grammar revision cycle. In *Beyond Shallow Analyses–Formalisms and Statistical Modelling for Deep Analysis (Workshop at IJCNLP-2004)*, Hainan Island.

Scott H. Paauw. 2009. *The Malay contact varieties of Eastern Indonesia: A typological comparison*. PhD dissertation, State University of New York at Buffalo.

Woodley Packard. 2015. Full forest treebanking. Master's thesis, University of Washington.

Carl Pollard and Ivan A Sag. 1994. *Head-driven phrase structure grammar*. University of Chicago Press.

Hammam Riza, Michael Purwoadi, Gunarso, Teduh Uliniansyah, Aw Ai Ti, Sharifah Mahani Aljunied, Luong Chi Mai, Vu Tat Thang, Nguyen Phuong Thái, Vichet Chea, Rapid Sun, Sethserey Sam, Sopheap Seng, et al. 2016. Introduction of the Asian Language Treebank. In *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Technique (O-COCOSDA)*, pages 1–6. IEEE.

Ivan A. Sag, Thomas Wasow, and Emily M. Bender. 2003. *Syntactic Theory: A Formal Introduction*. CSLI Publications, Stanford, 2 edition.

James Neil Sneddon, Alexander Adelaar, Dwi Noverini Djenar, and Michael C. Ewing. 2010. *Indonesian Reference Grammar*. Allen & Unwin, New South Wales, 2 edition.

James Neil Sneddon. 2006. *Colloquial Jakartan Indonesian*. Pacific Linguistics, Canberra.

Sebastian Sulger, Miriam Butt, Tracy Holloway King, Paul Meurer, Tibor Laczkó, György Rákosi, Cheikh M Bamba Dione, Helge Dyvik, Victoria Rosén, Koenraad De Smedt, et al. 2013. Pargrambank: The pargram parallel treebank. In *ACL (1)*, pages 550–560.

Miyao Yusuke and Tsujii Jun'ichi. 2002. Maximum entropy estimation for feature forests. In *Proceedings of the second international conference on Human Language Technology Research*, pages 292–297. Morgan Kaufmann Publishers Inc.

Simulating Segmentation by Simultaneous Interpreters for Simultaneous Machine Translation

Akiko Nakabayashi
The University of Tokyo
akinkbys@phiz.c.u-tokyo.ac.jp

Tsuneaki Kato
The University of Tokyo
kato@boz.c.u-tokyo.ac.jp

Abstract

One of the major issues in simultaneous machine translation setting is when to start translation. Inspired by the segmentation technique of human interpreters, we aim to simulate this technique for simultaneous machine translation. Using interpreters' output, we identify segment boundaries in source texts and use them to train a predictor of segment boundaries. Our experiment reveals that translation based on our approach achieves a better RIBES score than conventional sentence-level translation.

1 Introduction

Simultaneous interpreters listen to speech in a source language, translate it into a target language, and deliver the translation simultaneously. In this process, it is important to minimize the burden on the interpreters and to keep up with the original speech, particularly between language pairs with different word orders, such as English and Japanese. One of the tactics often used by interpreters is segmentation, which is to split a sentence according to "units of meaning" into multiple segments and translate those segments in sequence. Reformulation, simplification, and omission are further applied to generate natural translation (Jones, 1998; He et al., 2016).

In simultaneous machine translation, where speeches and lectures are translated simultaneously, this segmentation technique is also effective for minimizing translation latency. If translation is generated per sentence, as in a standard machine-translation process, there is a substantial delay between the original speech and its translation. By contrast, if a sentence is segmented into excessively short pieces, it becomes difficult to produce a meaningful translation from snippets of information. It is therefore important to determine the appropriate translation segment length that defines the correct timing to start the translation.

This study aims to learn the segmentation technique—which is necessary to realize fluent simultaneous machine translation with low latency—from human interpreters by examining their output, and to analyze this technique and propose a method to simulate it. The problem, however, is that the segments identified by interpreters are not always self-evident. Segments are considered to be produced by interpreters by splitting a sentence into "units of meaning," which is a minimal unit for interpreters to process information, but the linguistic characteristics of these "units of meaning" remain unclear. After discussing the background in Section 2, we return to present this issue in Section 3, where we propose an approach to identify segment boundaries in source texts using interpreters' output. Then, we demonstrate that segments identified by the proposed approach are plausible and that this segmentation approach can produce fluent translation with low latency. In Section 4, we describe the analysis of segment boundaries in the source texts. This analysis aims at understanding what factors determine those boundaries, as source texts are the only available means of identifying them in actual simultaneous-machine-translation settings. The result reveals that

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 165-173, Hakodate, Japan, September 13-15, 2019

an intricate set of linguistic factors define segment boundaries. Based on this analysis, we propose a framework to simulate the segmentation technique of interpreters using a predictor based on a Recurrent Neural Network (RNN) and analyze the result in Section 5. Although this predictor does not reproduce interpreters' segmentation perfectly, the translation generated per predicted segment achieves better RIBES scores[1] (Isozaki et al., 2010) compared with conventional sentence-level translation.

Overall, the contributions of this study are

- We analyze the segmentation tactic of human interpreters and clarify what linguistic factors define segment boundaries.
- We propose a framework to simulate this segmentation. Our experiment reveals that this approach of learning segmentation tactic from human interpreters benefits simultaneous machine translation.

2 Background

The process of simultaneous interpreting involves segmentation and reformulation. How sentences are segmented defines the appropriate timing to start the translation, and identifying these timings is one of the major issues in the research on simultaneous machine translation. Various strategies have been explored to address this issue. Fügen et al. (2007) and Bangalore et al. (2012) tried to find clues based on linguistic and a non-linguistic features (such as commas and pauses) in the original speech, whereas segments defined assuming that the segmentation depends solely on a specific syntactical feature of a source text may not provide the best timing for the target text. Fujita et al. (2013) suggested determining the translation timings by referring to a translation phrase table. Oda et al. (2015) built a classifier to find segment boundaries that maximized the sum of translation quality indices. Recent studies focused more on end-to-end approaches by training translation timings and translation models all together to produce translations with high translation scores (Cho and Esipova, 2016; Ma et al., 2018). Among them, Cho and Esipova (2016), rather than segmenting the original speech before the translation process, chose to translate the original text incrementally and define the appropriate timing to fix the translation based on a prescribed criterion. In those approaches, sentence-level translation corpora were used to train and evaluate the models, which did not necessarily produce good results in simultaneous-machine-translation settings. However, interpreting corpora are limited in size; therefore, it is not realistic to use them in these approaches.

In this study, we propose utilizing a simultaneous-interpreting corpus to find segment boundaries on source texts and building a segmentation predictor based on them. As transcripts of simultaneous interpreters provide insight on how they split sentences into segments at appropriate timings, simulating this tactic and translating based on those segments are expected to yield translation close to actual simultaneous interpreting. Shimizu et al. (2014) also referred to interpreters' output to identify segment boundaries, but used a non-linguistic feature to find patterns of segmentation. We, by contrast, utilize linguistic features that appear in interpreters' output not only to identify segment boundaries, but also to predict them. Tohyama and Matsubara (2006) and He et al. (2016) conducted descriptive studies on the tactics of simultaneous interpreters.

After using a model to split sentences into segments, we assume that each segment is translated independently using a conventional translation model and that reformulation is applied if the segment is syntactically incomplete. This process produces the final translation output.

For analysis and experiment, we used CIAIR Simultaneous Interpreting Corpus (Toyama et al., 2004), which contains transcripts of interpreters who simultaneously interpreted monologues and conversation. Regarding monologues, there are 136 simultaneous-interpretation transcripts with 5,011 utterances for 50 English speeches with 2,849 utterances. We used 24 speeches recorded in 2000, which have the transcripts of four interpreters each.

[1] We use RIBES scores as our evaluation criteria because it factors in word order, which is critical in simultaneous interpreting.

3 Segment Boundaries in Simultaneous Interpreting

In this section, we address our approach to identifying segment boundaries, which focuses on the interpreting results. After describing our motivation, we explain our method and demonstrate its effectiveness.

3.1 What are Segment Boundaries?

Interpreters split a sentence according to "units of meaning" into segments and translate them in sequence, but what is a "unit of meaning?" Is it defined by speakers' pauses, or by punctuations, as previous works suggest? A "unit of meaning" cannot be systematically related to grammatical categories, and it changes depending on the speaker's utterance speed and languages pairs (Jones, 1998). Based on the idea that a "unit of meaning" is a cognitive representation in the listener's mind (Jones, 1998), we see that the segment boundaries identified by interpreters appear in their output. The following example shows that it is difficult to identify segment boundaries by solely examining a source text, whereas the interpretation output provides some clues about the segments recognized by the interpreters.

Source text:
If you do that, the ups and downs seem to level out and you build more. It's a natural way of making money. (SXPSX006.NX02.ETRANS)
Interpretation transcript:
その間にはいろいろな上下があると思いますがその長い期間の間にそういったものは平均が取れて最終的にはお金が儲かっていくと思います。(SXPSX006.L.IA08.JTRANS)
"During that time, there will be various ups and downs, I think, but during that long period, such things will be leveled out, and at the end, we can make money, I think."

The source text seems to suggest that it can be syntactically split between the conditional clause and the main clause in the first sentence, and between the first and the second sentence. However, when we look at the interpreter's transcripts, we can see that the interpreter split the sentence immediately after "いろいろな上下があると思いますが (there will be various ups and downs, I think)," which corresponds to the *the ups and downs* in the source text.

Jones (1998) further claimed that interpreters can start the translation "once they have enough material from the speaker to finish their own (interpreted) sentence." Given that sentences tend to become long with coordinate conjunctions connecting multiple clauses, and sentence boundaries are not clear in a spoken language, a sentence can be rephrased as a clause. In light of this idea, we believe that once interpreters identify a "unit of meaning',' they translate it and produce a clause. In other words, a clause in interpreters' output is the translation of a "unit of meaning" that they recognize. Therefore, we propose the following approach to identifying segment boundaries:

- Split interpreting results into clauses
- Identify segments on source speech texts by finding corresponding word strings

Clauses in the interpreters' output and the corresponding segments in the source speech texts were annotated manually in this study; however, we believe that these processes can be automated.

In the previous example, segment boundaries are considered to appear at the following positions in the source text.

Source text:
If you do that, the ups and downs seem / to level out / and you build more. It's a natural way of making money. /

3.2 Identified Segment Boundaries

We identified segment boundaries based on the aforementioned approach. We examined the segmentation distribution in the transcripts of four interpreters associated with the speech text file (SXPSX005.NX02.ETRANS). As shown in Table 1, the cumulative total of segment boundaries identified by four interpreters was 441 with 153 distinct places. All four interpreters agree to split segments at 64 distinct places, i.e., 256 places in total. Three out of four interpreters agree to split segments at 36 distinct places, i.e., 108 places in total. The cumulative total of segments on which three or more inter-

preters agree was 364 (82.5%) out of 441 segments. We believe that this number is sufficiently large to confirm that segment identification by the aforementioned approach is objective and usable. For further research, we extracted segment boundaries shared by three or four interpreters.

Number of Interpreters Agreed	Number of Segments (Total)
4	64 (256)
3	36 (108)
2	24 (48)
1	29 (29)
Total	153 (441)

Table 1: Distribution of segments (File SX-PSX005.NX02.ETRANS)

Table 2 shows an overview of the number of sentences and segments in 10 files. The first two lines show the total word count and total time spent for the 10 speeches. The total number of sentences in 10 files, average word count per sentence, average time spent per sentence, and those for segments follow.

Speech	Word Count	12,859
	Total Time (min:sec)	101:49
Sentence	Number of Sentences	510
	Average Word Count	25.2
	Average Time (min:sec)	0:12
Segment	Number of Segments	1,127
	Average Word Count	11.4
	Average Time (min:sec)	0:05

Table 2: Overview of sentences and segments

The average word count in a sentence is 25, while that in a segment is 11. Given that the average time spent on a segment is 5 seconds, and that spent on a sentence is 12 seconds, segment-level translation is considered to reduce translation latency by 7 seconds compared with sentence-level translation. This shows that the proposed segmentation approach contributes to reducing translation latency.

We compared the RIBES scores of the segment-level translation with those of the interpreters' transcript to prove that the translation generated by the proposed approach resembles the interpreters' output. After segment boundaries were identified, each segment was translated using Google Translate[2]. The translated segments were concatenated and used as final translation. Reformulation was not applied in this experiment. The RIBES score of this segment-level translation was 0.7755, with the transcripts of three interpreters used as the reference translation. The RIBES score of the other interpreter's transcript was 0.7754 and that of the sentence-level translation was 0.7412 using with the same reference translation.

The fact that the RIBES score of the segment-level translation with the proposed approach was close to that of an interpretation transcript suggests that the proposed approach can generate translation comparable to interpreters' output and that considering a clause in such output to be a "unit of meaning" is plausible and realistic.

4 Characteristics of Segments

While we identified segments through a relationship with interpreters' output as in the previous section, the interpreters' output is not available in actual simultaneous-machine-translation settings when predicting segment boundaries. We analyzed the segments and their characteristics to understand what factors determine the segment boundaries in the source texts.

We extracted part-of-speech (POS) bigrams before and after the segment boundaries to determine where such boundaries tend to appear. Table 3 shows the top six patterns of segment boundaries. The numbers in parentheses show the proportion of segment boundaries to the places where each pattern appears.

Feature	Number of Segment Boundaries
After "."	1,207 (98 .3%)
Before Coordinate Conjunction	927 (55.7%)
After ","	545 (39.5%)
Before Wh	114 (20.3%)
Before Adverb	240 (11.3%)
Before Preposition/ Subordinate Conjunction	377 (10.8%)

Table 3: Characteristics of segment boundaries

While periods are a strong indication for defin-

[2]https://translate.google.com [Accessed: 9 Jan, 2019]

ing segment boundaries, they also appear elsewhere, such as before coordinate conjunctions and after commas[3]. However, not all positions with these features become segment boundaries, and various linguistic factors other than POS also seem to play an important role in the decision. For example, a conjunction may coordinate noun phrases or clauses. If the conjunction coordinates clauses, it is more likely that word strings before the conjunction become a segment than when it coordinates noun phrases. However, this information cannot be captured by examining POS n-grams alone.

The last five bigram patterns are discriminative in simultaneous interpreting (Tohyama and Matsubara, 2006). We focused on relative pronouns, which include wh-determiners and subordinate conjunctions, and further analyzed what factors influence decisions on whether a relative clause with a relative pronoun becomes a segment or not. In Japanese, a relative clause comes before the antecedent, and sentence-level translation usually employs this word order. However, in simultaneous interpreting, the antecedent is often translated before the relative clause is fully uttered.

We built a logistic regression classifier to predict segment boundaries and investigated the weight of each feature to determine what factors contribute to the decision on segmentation. The features used in the classification model were: the number of words in the relative clause, the syntactic role of the antecedent in the main clause, the syntactic role of the relative pronoun in the relative clause, and the presence of comma before the relative pronoun. Concerning the syntactic role of an antecedent in the main clause and that of a relative pronoun in the relative clause, if the antecedent or the relative pronoun appeared before the verb, we assumed its syntactic roles to be "subject (SBJ)"; otherwise, it was "object (OBJ)." The values of the features were normalized before training the logistic regression. We used the NLTK[4] package to predict the POS and the scikit-learn[5] package to build the logistic regression models. A total of 45 POS, including symbols, were defined in NLTK. The accuracy of this model was 0.687, although classification was not the primary objective.

Table 4 shows the weight of each feature. A large absolute value of the weight means high contribution of the feature to the decision on segmentation, and a large positive value of the weight means that the position with the feature is likely to become a segmentation boundary.

Feature	Weight	p-value
Number of Words in Relative Clause	0.3048	$<$0.001
Role in Main Clause (SBJ)	-0.4502	$<$0.001
Role in Relative Clause (SBJ)	0.1169	0.11
Comma	0.3132	$<$0.001

Table 4: Factors influencing segment boundaries

This result shows that the number of words in the relative clause, the role of the antecedent in the main clause, and the presence of a comma are within the level of significance and are important for the definition of segment boundaries. Rather than a single factor, multiple linguistic factors contribute to the decision of simultaneous interpreters on where to split a sentence.

5 Predicting Segment Boundaries

In this section, we describe our segmentation framework for simultaneous machine translation. The data presented in the previous section show that clues on segmentation cannot be explained by a single feature. To integrate such intricate features, we built an RNN-based predictor of segment boundaries. After explaining its architecture, we show the results of experiments performed to examine whether it can capture these linguistic features and simulate interpreters' tactics.

It is worth noting that in simultaneous-machine-translation settings, words in the source text become available one by one, and the entire sentence is not available to be parsed as in the analysis presented in Section 4. Hence, all the linguistic features stated in Section 4 may not be readily available when predicting segment boundaries. We expect that those features are somehow represented in and related to the already available context. The predictor predicts

[3]Commas and periods are annotated in the transcriptions. We believe corresponding information can be captured through acoustic information in actual simultaneous-machine-translation settings.

[4]http://www.nltk.org

[5]https://scikit-learn.org/stable/

segment boundaries using segmented source texts generated by the approach proposed in Section 3 as training data. It was modeled as a binary classification with the task of predicting whether a segment boundary appears in front of an input word, when a word sequence is given as input.

5.1 Experiment Settings

The model uses the long short-term memory (LSTM) (Hochreiter and Schmidhuber, 1997) of an RNN. We use a uni-directional RNN, given that the whole sentence is not available when predicting segment boundaries and words in the source texts become available one by one in sequence in simultaneous-machine-translation settings. Figure 1 shows an overview of the model.

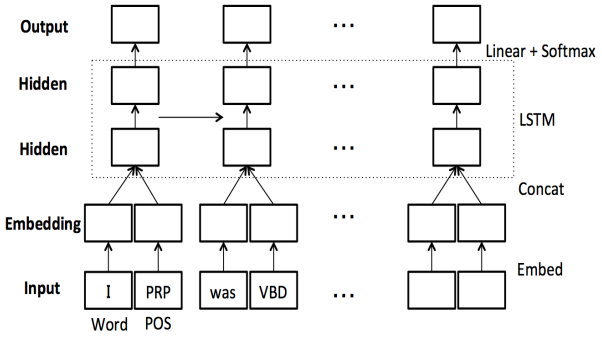


Figure 1: Model for predicting segment boundaries

Each word and its POS in the input layer are represented as one-hot vectors. The word and POS are concatenated after being mapped to the embedding layer with a lower dimension; this concatenated embedding is then used as the input for the next LSTM layer. The output of the LSTM is mapped to a two-dimensional vector and the result of applying the softmax function to the vector shows the probability of the input being assigned to each class. We used cross-entropy as a loss function.

The Chainer[6] package is used to implement the model. The dimension of word embeddings is set to 300, while that of POS embeddings is 4. The input and the output of the LSTM have 304 dimensions. The dropout rate is set to 0.1 and the class weight is set to 3 to deal with biased samples.

[6]https://chainer.org

5.2 Training Data and Test Data

Out of 24 speeches, 22 were used as a training dataset, one as a development set, and one as a test set. Table 5 provides an overview of the data used.

	Training Data	Test Data
Tokens	30,151	1,197 (OOV: 231)
Types	3,278	410
Sentences	1,097	70
Segment Boundaries	2,413	130
Non Segment Boundaries	27,738	1,067

Table 5: Size of data

The numbers of unknown words that appear in the test set but not in the training set (Out-of-vocabulary; OOV) are large.

Words on segment boundaries constitute only 8.0% of the total number of words, so the number of words in each class is biased.

Each datum is labeled as described below. Training and testing are executed per sentence. Class "1" shows that a segment boundary comes before the corresponding word.

Words: "I", "was", "traveling", "in", "Europe", "and", "when", "I", "was", "in", "Greece", ",", "I", "met", "a", "man", "from", "Holland", "."
Label: (1, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 0)

5.3 Experiment Results

Table 6 shows the precision, recall, and F-measure of the experiment. The F-measure for class "1" was 0. 80.

Class	Precision	Recall	F-measure
1	0.82	0.78	0.80
0	0.97	0.98	0.98

Table 6: Results of segment boundary prediction

We further analyzed the prediction results for the comma, coordinate conjunctions, and wh-clauses by twelve-fold cross-validation. Often, but not always, interpreters split sentences before those words in simultaneous interpreting, and various factors are considered to influence their decisions as we discussed

in the previous section. Table 7 shows the baseline, accuracy, recall, and precision for the comma, coordinate conjunctions, and wh-clauses. We used the most common prediction, which is the probability of predicting the biggest category by chance, as our baseline.

	Before Coordinate Conjunction	After ","	Before Wh-clause
Baseline	55.7	60.5	79.7
Accuracy	66.0	66.2	80.9
Precision	66.1	55.1	52.5
Recall	80.3	76.7	64.9

Table 7: Results for coordinate conjunctions, comma, and wh-clauses

The accuracy for all three cases was higher than baseline, and we can say that the model could capture more information than POS has. To further investigate the results, we picked up two relative pronouns, *which* and *that*, to see if there are any significant differences between words. Table 8 shows the number of boundaries, as well as the baseline, accuracy, recall, and precision for *which* and *that*.

	which	that
Total Number	69	116
Number of Boundaries	33	16
Baseline	52.2	86.2
Accuracy	68.1	75.0
Precision	63.4	15.8
Recall	78.8	18.8

Table 8: Results for relative pronouns, *which* and *that*

Segment boundaries often appear before *which*, while they do not before *that*. The accuracy for the relative pronoun *which* was higher than baseline. The following example shows a case where the model correctly predicted the segment boundary before the relative pronoun *which*.

Label: People are a little more casual, / they take their time / and they're really very friendly / which is something that makes me feel a lot better. / (SXUSX012)

Predicted: People are a little more casual, they take their time / and they're really very friendly / which is something that makes me feel a lot better. /

By contrast, the accuracy for *that* was lower than baseline. This may be attributed to the small positive training dataset available for this relative pronoun.

5.4 Translation Results

After segmenting sentences at the predicted segment boundaries, we translated each segment using Google Translate. Then, we concatenated the translation outputs, and analyzed it by calculating their RIBES scores. The transcripts of four interpreters were used as the reference translation. Although reformation should be applied separately before yielding the final output, this analysis was conducted on the translation texts without reformation.

The RIBES score of segment translation with predicted segment boundaries was 0.7683, which is higher than the score of sentence translation, i.e., 0.7610. The RIBES score of segment translation with correct segment boundaries was 0.7964. Table 9 shows some examples. Translation results generated by the proposed approach had a word order similar to that of the interpreters' transcripts. By applying reformation, translation outputs are expected to become more natural.

Table 10 shows an example with a low RIBES score for segment translation with the predicted segment boundaries. The predictor failed to split the sentence before the preposition, splitting it at a wrong position, which caused a reduced RIBES score. These issues can be resolved by improving the accuracy of the segmentation.

6 Conclusion and Further Study

Segmentation is one of the key issues in the area of simultaneous machine translation. To resolve it, we proposed a method that uses interpreters' output. Specifically, we assumed that a "unit of meaning" appears as a clause in interpreters' output and identified segment boundaries by marking the corresponding position in the source texts. We analyzed them in the source texts and pointed out that various linguistic factors determine those boundaries.

We used segment boundaries in the source texts as training data to build a segmentation predictor that reproduces interpreters' segmentation strategies. The F-measure of the segmentation predictor

Segment Boundaries	The bagpipe is most commonly heard at highland games / where many bands gather to play music / and perform Scottish games such as the caber toss or the hammer throw .
Translation based on Segment Boundaries	バグパイプはハイランドゲームでよく聞かれます たくさんのバンドが集まって音楽を演奏する そして、キャバートスやハンマースローなどのスコットランドの試合を行います。 “The bagpipe is most commonly heard at highland games. Many bands gather and play music. And perform Scottish games, such as the caber toss or the hammer throw.”
Predicted Segment Boundaries	The bagpipe is most commonly heard at highland games / where many bands gather to play music / and perform Scottish games such as the caber toss / or the hammer throw.
Translation based on Predicted Segment Boundaries	バグパイプはハイランド ゲームでよく聞かれます 音楽を演奏するために多くのバンドが集まる場所 キャバートスなどのスコットランドのゲームを実行する またはハンマー投げ “The bagpipe is most commonly heard at highland games. The place many bands gather to play music. Perform Scottish games, such as the caber toss. Or the hammer throw.”
Interpreting Transcript	バグパイプが通常聞かれるのはハイランドゲームでです。そこで多くのバンドが集まりましてそして音楽を演奏します。そしてスコットランドのゲーム例えばケーバートスあるいはハンマースローなどを行います。 “The bagpipe is most commonly heard at highland games. There, many bands gather and play music. And perform Scottish games, such as the caber toss or the hammer throw.”

Table 9: Translation results

Segment Boundaries	This happened again and again / until several people were, of course, killed.
Translation based on Segment Boundaries	これは何度も何度も起こった もちろん数人が殺されるまで。 “This happened again and again. Of course until several people were killed.”
Predicted Segment Boundaries	This happened again and again until several people were, / of course, killed.
Translation based on Predicted Segment Boundaries	これは何人かの人々がなるまで何度も何度も起こりました、もちろん、殺した。 “This happened again and again until several people became. Of course killed.”
Interpreting Transcript	これが何度も何度も繰り返されました 。何人かの人が撃ち殺されるまでやったわけです。 “This happened again and again. Did it until several people were killed.”

Table 10: Translation results with low RIBES scores

was 0.80. Interpreters often split sentences before relative pronouns, and in many cases the predictor could predict segment boundaries correctly at such positions. When we split sentences at the predicted positions and translated each segment using Google Translate, the output had a word order similar to that of the interpreters' transcripts and its RIBES score was higher than that of sentence-level translation. This underscores that the proposed approach benefits simultaneous machine translation. However, incorrectly predicted segment boundaries degraded the translation quality. Therefore, further improvement in the accuracy of the segmentation is required. The reformation model is another topic for further study.

References

Bangalore, S., Rangarajan Sridhar, K. V., Kolan, P., Golipour, L., and Jimenez, A. 2012. Real-time Incremental Speech-to-Speech Translation of Dialogs. *Pro-*

ceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: 437-445.

Cho, K., and Esipova, M. 2016. Can neural machine translation do simultaneous translation? *arXiv:1606.02012*.

Fügen, C., Waibel, A., and Kolss, M. 2007. Simultaneous translation of lectures and speeches. *Machine translation*, 21(4): 209-252.

Fujita, T., Neubig, G., Sakti, S., Toda, T., and Nakamura, S. 2013. Simple, lexicalized choice of translation timing for simultaneous speech translation. *Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH*: 3487-3491.

He, H., Boyd-Graber, J., and Daume III, H. 2016. Interpretese vs. Translationese: The Uniqueness of Human Strategies in Simultaneous Interpretation. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*: 971-976.

Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. *Neural computation*9(8): 1735-1780.

Isozaki, H., Hirao, T., Duh, K., Sudoh, K., and Tsukada, H. 2010. Automatic Evaluation of Translation Quality for Distant Language Pairs. *Proceedings of the Conference on Empirical Methods on Natural Language Processing (EMNLP)*: 944-952. Association for Computational Linguistics.

Jones, R. 1998. *Conference interpreting explained*. Routledge.

Shimizu, H., Neubig, G., Sakti, S., Toda, T., and Nakamura, S. 2014. Doujitsuuyaku Deta wo Riyou Shita Onsei Doujitsuuyaku no tameno Yakushutsu Taimingu Kettei Shuhou (A Method to Decide Translation Timing for Simultaneous Speech Translation Using Simultaneous Interpreting Data). *Proceedings of the Association for Natural Language Processing*: 294-297.

Ma, M., Huang, L., Xiong, H., Liu, K., Zhang, C., He, Z., Liu, H., Li, X., and Wang, H. 2018. Stacl: Simultaneous translation with integrated anticipation and controllable latency. *arXiv:1810.08398*.

Oda, Y., Neubig, G., Sakti, S., Toda, T., and Nakamura, S. 2015. Syntax-based Simultaneous Translation through Prediction of Unseen Syntactic Constituents. *Proceedings of the 53rd ACL*(1): 198-207.

Toyama, H., Matsubara, S., Ryu, K., Kawaguchi, N., and Inagaki, Y. 2004. CIAIR Simultaneous Interpretation Corpus. *Proceedings of the oriental chapter of the International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques for Speech Input/Output (Oriental COCOSDA 2004)*.

Tohyama, H., Matsubara, S. 2006. Collection of Simultaneous Interpreting Patterns by Using Bilingual Spoken Monologue Corpus. *International Conference on Language Resources and Evaluation (LREC2006)*: 2564-2569.

Attention mechanism for recommender systems

First Author
Nguyen Huy Xuan
nguyenhx@jaist.ac.jp

Second Author
Le Minh Nguyen
nguyenml@jaist.ac.jp

Abstract

Sparseness of user item rating data affects the quality of recommender systems. To solve this problem, many approaches have been proposed. They added supplemental information to increase the accuracy. We propose a recommendation model namely attention matrix factorization (AMF) that integrates attention mechanism of the both item reviews document and item genre information into probabilistic matrix factorization (PMF). Consequently, AMF attends features which are mentioned in item reviews document and further increases the rating prediction accuracy by adding item genre information. Our experiments on the Movielens and Amazon instant video datasets show that AMF outperforms the previous traditional recommendation systems. This reveals that our model can capture subtle features of item reviews although the rating data is sparse.

1 Introduction

The sparseness of item rating is still a challenge for recommender systems. Eventually, this problem affects the rating prediction accuracy of traditional collaborative filtering (CF) approaches (Adomavicius and Tuzhilin, 2005; Herlocker et al., 2004). Recently, to improve the accuracy, several methods are proposed such as Latent Dirichlet Allocation (LDA) and Stacked Denoising AutoEncoder (SDAE). These approaches added item description information such as reviews, abstracts or synopes (Ling et al., 2014; McAuley and Leskovec, 2013; Wang and Blei, 2011; Wang et al., 2015).

Wang et al. have proposed collaborative topic regression (CTR) method which unites Latent Dirichlet Allocation (LDA) and collaborative filtering (CF) in a probabilistic approach (Wang and Blei, 2011). The author also proposed collaborative deep learning (CDL) which integrates Stacked Denoising AutoEncoder (SDAE) into probabilistic matrix factorization (PMF) (Salakhutdinov and Mnih, 2008; Wang et al., 2015). Variants of CTR were integrated with topic modeling (LDA) into collaborative filtering to analyze item description with different approaches (Ling et al., 2014; McAuley and Leskovec, 2013). However, the integrated models do not fully capture document information.

In order to overcome the issue, Donghyun Kim et al. have proposed ConvMF (Kim et al., 2016) which uses item reviews document in CNN model and further enhances the rating prediction accuracy. However, it does not mention another information such as item genre information. It also does not capture attended features of item reviews document.

The most recently, the combination between deep learning methods with CF and content-based filtering methods is also proposed. Yu Liu et al. have proposed a novel deep hybrid recommender system framework based on auto-encoders (DHA-RS) by integrating user and item side information to construct a hybrid recommender system and enhance performance (Liu et al., 2018). The author has proposed two models based on the DHA-RS framework which integrates user and item side information. Libo Zhang et al. have proposed a model combining a CF algorithm with deep learning technology (Zhang et al., 2018). This approach uses a fea-

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 174-181, Hakodate, Japan, September 13-15, 2019

ture representation method based on a quadric polynomial regression model, which obtains the latent features more accurately by improving upon the traditional matrix factorization algorithm. These latent features are regarded as the input data of the deep neural network model, which is the second part of the proposed model and is used to predict the rating scores.

In this paper, we propose attention matrix factorization (AMF) model which integrates attention mechanism into probabilistic matrix factorization. Our model is different from previous approaches. We use attention mechanism which uses the both item reviews and item genre information to enhance rating prediction accuracy and attend features which are mentioned in item reviews information. For example, we have item reviews document and item genre as follows.

*Item reviews: He **license** to **kill** bond race to russia in search of the **stolen** access code.*

Item Genre[1]*: GoldenEye (1995)::**Action, Adventure, Thriller***

By adding item genre: *Action, Adventure, Thriller*, our AMF model captures attended features such as *license, kill, stolen* which are mentioned from item reviews document. Our contributions are summarized as follows.

- We propose an attention matrix factorization model which exploits ratings, item reviews documents and item genre information.
- We extensively demonstrate that AMF is a combination of PMF with attention mechanism on three datasets with more effective features representation.
- We conduct different experiments and show that AMF can facilitate the data sparsity problem in CF.

The rest of the paper is described as follows. Section 2 reviews preliminaries on the CF technique and attention neural network. Section 3 describes the AMF model and optimization method. Experimental results and evaluation AMF are presented in Section 4. Finally, we present our conclusion in Section 5.

[1]http://www.imdb.com/

2 Our baseline

In this section, we shortly describe the most common CF technique that is Matrix Factorization (MF) and attention network.

2.1 Matrix Factorization

Matrix factorization is one of the most popular methods in CF (Koren et al., 2009). Generally, MF model can learn low-rank representations (i.e., latent factors) of users and items in the user-item matrix, which are further used to predict new ratings between users and items. Assume that: N is a set of users; M is a set of items, and R is a rating matrix of users for items ($R \in \mathbb{R}^{N \times M}$). MF discovers the k-dimensional models, which is the latent models of user u_i ($u_i \in \mathbb{R}^k$) and item v_j ($v_j \in \mathbb{R}^k$). The rating r_{ij} of user i on item j can be approximated by equation: $r_{ij} \approx \hat{r}_{ij} = {u_i}^T v_j$. The loss function $\mathcal{L}$ is calculated by equation as below.

$$\mathcal{L} = \sum_{i}^{N} \sum_{j}^{M} f_{ij}(r_{ij} - {u_i}^T v_j)^2 + \lambda_u \sum_{i}^{N} \| u_i \|^2 + \lambda_v \sum_{j}^{M} \| v_j \|^2 \quad (1)$$

Where $f_{ij} = 1$ if user u_i rated v_j; otherwise, $f_{ij} = 0$

2.2 Attention neural network

Parikh et al., proposed decomposable attention model for Natural Language Inference (Parikh et al., 2016). Inputs are two phrases represented as a sequence of word embedding vectors $a = (a_1, ..., a_{la})$ and $b = (b_1, ..., b_{lb})$. The goal of attention model is to estimate a probability that two phrases are in entailment or contradiction to each other. The core model architecture is to compose of three steps: 1) attention for generating soft-aligned to the second sentence, 2) comparison for comparing soft-aligned sentence matrices, 3) aggregation for column-wise sum over the output of the comparison step so that we obtain a fixed-size representation of every sentence.

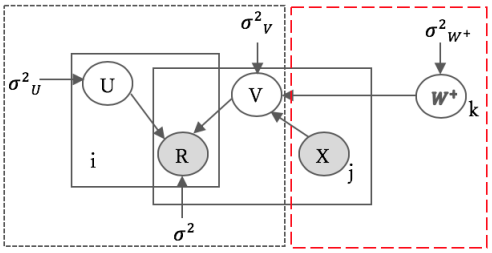


Figure 1: AMF architecture: PMF in left (dotted black); Attention neural architecture part in right (dashed red)

3 Attention mechanism for MF

In this section, we introduce our attention matrix factorization (AMF), following 3 steps: 1) We describe the probabilistic model of AMF, and introduce the main idea to combine PMF and attention mechanism in order to use ratings, item reviews documents and item genre information. 2) We describe the architecture of our attention mechanism, that generates document latent model by analyzing item reviews document and item genre. 3) We explain how to optimize our AMF.

3.1 Probabilistic Model of AMF

Our AMF is described in Figure 1, that combines an attention mechanism and PMF model. This part is cited from previous research in (Kim et al., 2016). The conditional distribution over observed ratings is given by

$$\rho(R|U,V,\sigma^2) = \prod_i^N \prod_j^M \mathcal{N}(r_{ij}|{u_i}^T v_j, \sigma^2)^{f_{ij}} \quad (2)$$

$\mathcal{N}(x|\mu,\sigma^2)$ is the Gaussian normal distribution with mean μ and variance σ^2, and f_{ij} is described in Section 2.1. The item latent model is given below.

$$v_j = att^+(W^+, X_i) + \epsilon_j \quad (3)$$

$$\epsilon_j = \mathcal{N}(0, {\sigma^2}_V f)) \quad (4)$$

Where $att^+()$ represents the output of attention architecture; X_i representing the document of item i and epsilon variable as Gaussian noise. For each weight ${w_k}^+$ in W^+, we set zero-mean spherical Gaussian prior.

$$\rho(W^+|{\sigma^2}_{W^+}) = \prod_k \mathcal{N}({w_k}^+|0, {\sigma^2}_{W^+}) \quad (5)$$

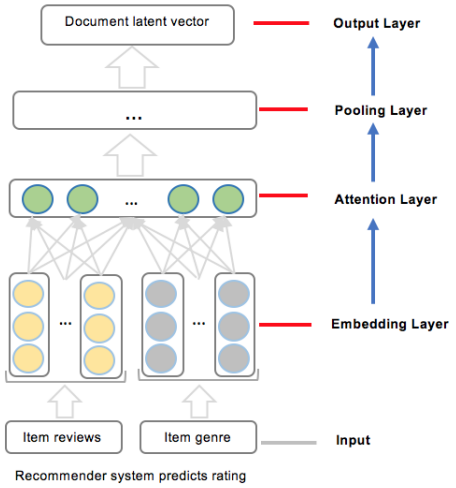


Figure 2: Our Attention neural architecture for AMF

$$\rho(V|W^+, X{\sigma^2}_V) = \prod_j^M \mathcal{N}(v_j|att^+(W^+, X_j), {\sigma^2}_V f) \quad (6)$$

where X is the set of item reviews.

3.2 Attention mechanism of AMF

In this paper, our attention mechanism uses item reviews and item genre information. Figure 2 introduces our attention architecture that consists of 4 layers described as follows.

Input of our model is both items reviews document of user for item, and item genre information.

1) Embedding Layer.

This layer is to convert a raw document into a vector. For example, we have a document with number of words is l, then we can concatenate a vector of each word into a matrix in accordance with the sequence of words. The word vectors are initialized with pre-trained word embedding model such as Glove (Pennington et al., 2014). Then, the document matrix $D \in \mathbb{R}^{q \times l}$ can be visualized as follow:

$$\begin{bmatrix} w_{11} & \cdots & w_{1i} & \cdots & w_{1l} \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ w_{q1} & \cdots & w_{qi} & \cdots & w_{ql} \end{bmatrix} \quad (7)$$

in which q is the dimension of word embedding and $w[1:q,i]$ represents raw word i in the document.

2) Attention Layer.

Our attention layer is cited from previous reseach in (Parikh et al., 2016). Let $a = (a_1, ..., a_{la})$ and

$b = (b_1, ..., b_{lb})$ are the two inputs of item review and item genre with length l_a and l_b, respectively. Each $a_i, b_j \in \mathbb{R}^d$ is a word embedding vector of dimension d. Our attention mechanism is followed by three steps below.

a) Attend.

We first obtain unnormalized attention weights e_{ij}, computed by a function F', which decomposes as:

$$e_{ij} := F'(\bar{a_i}, \bar{b_j}) := F(\bar{a_i})^T F(\bar{b_j}). \quad (8)$$

Where $\bar{a} := a$ and $\bar{b} := b$. We take F to be a feed-forward neural network with ReLU activation function (Glorot et al., 2011). These attention weights are normalized as follows:

$$\beta_i := \sum_{j=1}^{lb} \frac{exp(e_{ij})}{\sum_{k=1}^{lb} exp(e_{ik})} \bar{b_j}, \quad (9)$$

$$\alpha_j := \sum_{i=1}^{la} \frac{exp(e_{ij})}{\sum_{k=1}^{la} exp(e_{kj})} \bar{a_i}, \quad (10)$$

β_i is the subphrase in $\bar{b}$ that is (softly) aligned to $\bar{a_i}$ and vice versa for α_j.

b) Compare.

Next, we separately compare the aligned phrases $\{(\bar{a_i}), \beta_i\}_{i=1}^{la}$ and $\{(\bar{b_j}), \alpha_i\}_{j=1}^{lb}$ using a function G.

$$v_{1,i} := G([\bar{a_i}, \beta_i]); \forall i \in [1, ..., la], \quad (11)$$

$$v_{2,j} := G([\bar{b_j}, \alpha_i]); \forall j \in [1, ..., lb]. \quad (12)$$

where the brackets $[.,.]$ denote concatenation. Thus G can jointly take into account both $\bar{a_i}$, and βi.

c) Aggregate.

Finally, we now have two sets of comparison vectors $\{v_{1,i}\}_{i=1}^{la}$ and $\{v_{2,j}\}_{j=1}^{lb}$. We first aggregate over each set by summation:

$$v_1 = \sum_{i=1}^{la} v_{1,i}; v_2 = \sum_{j=1}^{lb} v_{2,j}. \quad (13)$$

and feed the result through a final classifier H, that is a feed forward network followed by a linear layer:

$$\hat{y} = H([v_1, v_2]), \quad (14)$$

where $\hat{y} \in \mathbb{R}^C$represents the predicted (unnormalized) scores for each class and consequently the predicted class is given by $\hat{y}$ = argmax$x_i \hat{y}_i$.

3) Pooling Layer.

The pooling layer extracts representative features from the attention layer, and also deals with variable lengths of documents via pooling operation that constructs a fixed-length feature vector. After the attention layer, a document is represented as n_c contextual feature vectors. However, such representation has two problems: 1) there are some contextual features might not help enhance the performance, 2) the length of contextual feature vectors varies, which makes it difficult to construct the following layers. Therefore, we utilize max-pooling, which reduces the representation of a document into a n_c fixed-length vector by extracting only the maximum contextual feature from each contextual feature vector as follows.

$$d_f = [max(c^1), max(c^2), \cdots, max(c^j), \cdots, max(c^{n_c})] \quad (15)$$

where c^j is a contextual feature vector of length $l - ws + 1$ extracted by jth shared weight W_c^j.

4) Output Layer.

From output layer, the high-level features are extracted. A document latent vector is generated by equation as below.

$$s = tanh(W_{f_2}\{tanh(W_{f_1} d_f + b_{f_1}\} + b_{f_2}) \quad (16)$$

where W_{f_1} is projection matrices ($W_{f_1} \in \mathbb{R}^{f \times f}$); b_{f_1} and b_{f_2} are a bias vector of W_{f_1}, W_{f_2} with $s \in \mathbb{R}^k$ ($b_{f_1} \in \mathbb{R}^f, b_{f_2} \in \mathbb{R}^k$). Our Attention architecture becomes a function that exports a document latent vectors s_j of item j:

$$s_j = att^+(W^+, X_j) \quad (17)$$

where W^+ denotes all the weight and bias variables; and X_j denotes a raw document of item j.

3.3 Optimization Methodology

Our optimization is based on previous research in (Kim et al., 2016). We utilize maximum a posteriori estimation to optimize the variables of attention.

The optimization function $\mathcal{L}$ is given below.

$$\mathcal{L}(U,V,W^+) = \sum_i^N \sum_j^M \frac{f_{ij}}{2}(r_{ij} - {u_i}^T v_j)_2 + \frac{\lambda_U}{2}\sum_i^N \parallel u_i \parallel_2 + \frac{\lambda_V}{2}\sum_j^M \parallel v_j - att^+(W^+, X_j) \parallel_2 + \frac{\lambda_{W^+}}{2}\sum_k^{|{w_k}^+|} \parallel {w_k}^+ \parallel_2 \tag{18}$$

where $\lambda_U = \sigma^2/{\sigma^2}_U$, $\lambda_V = \sigma^2/{\sigma^2}_V$, and $\lambda_{W^+} = \sigma^2/{\sigma^2}_{W^+}$.

The optimal solution of U (or V) is given by equations below.

$$u_i \leftarrow (VI_iV^T + \lambda_U I_K)^{-1} VR_i \tag{19}$$

$$v_i \leftarrow (UI_jU^T + \lambda_V I_K)^{-1}(UR_j + \lambda_V att^+(W, X_j)) \tag{20}$$

where I_i is a diagonal matrix with $I_{ij}, j = 1, ..., M$ and R_i is a vector with $(rij)_{j=1}^M$ for user i. For item j, I_j and R_j are similarly defined as I_i and R_i, respectively.

L is interpreted as a squared error function with L_2 regularized terms as follows.

$$\varepsilon(W^+) = \frac{\lambda_V}{2}\sum_j^M \parallel v_j - att^+(W^+, X_j) \parallel^2 + \frac{\lambda_{W^+}}{2}\sum_k^{|{w_k}^+|} \parallel {w_k}^+ \parallel^2 + const \tag{21}$$

The back propagation algorithm is used to optimize W^+. Finally, the prediction of unknown ratings of users on items is given by equation below.

$$\begin{aligned} r_{ij} &= \mathbb{E}[r_{ij}|{u_i}^T v_j, \sigma^2] = {u_i}^T v_j \\ &= {u_i}^T(att^+(W^+, X_j) + \epsilon_j) \end{aligned} \tag{22}$$

Recall that $v_j = att^+(W^+, X_j) + \epsilon_j$

4 Experiment

In this part, we evaluate our AMF and compare with four start-of-the-art algorithms.

4.1 Experimental Setting

1) Datasets.

To evaluate rating prediction of our models, we used the MovieLens datasets[2] (ML) and Amazon Instant Video[3] (AIV). Each dataset contains user's ratings on items. Each rating value is 1-5. AIV dataset has item reviews and item descriptions. For ML data, we obtained item reviews of corresponding items from imdb site[4]. For the genre information, we extract from the item files (*_movies.dat) (i.e., $itemID$ :: $itemtitle$ :: $genre1|genre2|genre3|...$).

We also pre-processed item reviews documents for all datasets similar to previous approaches (Wang and Blei, 2011; Wang et al., 2015). We removed users and items that have less than 3 ratings and do not have their description documents. Table 1 shows the statistics of each dataset. We see that even when several users are removed by preprocessing, AIV is still sparse compared with the ML dataset.

2) Baselines.

We compared our AMF model with two previous methods, which are PMF (Salakhutdinov and Mnih, 2008), CTR (Wang and Blei, 2011) as well as two deep learning methods, which are CDL (Wang et al., 2015) and ConvMF (Kim et al., 2016).

3) Evaluation Metrics.

To evaluate the performance of each model, we randomly divided each dataset into three sets: 10% for test, 10% for validation and 80% for training. The training set contains at least one ratings on each user and item so that PMF deals with all users and items. Since our purpose is to conduct rating prediction, we use root mean squared error (RMSE) as the evaluation metrics.

$$RMSE = \sqrt{\frac{\sum_{i,j}^{N,M}(r_{ij} - \hat{r}_{ij})^2}{\#\ of\ ratings}} \tag{23}$$

4) Parameter Settings.

We set the training data with different percentage (20%, 40%, 80%). For the latent dimension of U and V, we set 50 according to previous work in (Wang et al., 2015) and initialized U, V randomly

[2]https://grouplens.org/datasets/movielens/

[3]http://jmcauley.ucsd.edu/data/amazon/

[4]http://www.imdb.com/

Dataset	Item information	Genre information	# Users	# Items	# Ratings	Density
ML-1m	Item reviews	Item genre	6,040	3,544	993,482	4.641%
ML-10m	Item reviews	Item genre	69,878	10,073	9,945,875	1.413%
AIV	Item reviews	Item genre	29,757	15,149	135,188	0.030%

Table 1: Data statistic on three real-world datasets

	ML-1m		ML-10m		AIV	
Model	λ_U	λ_V	λ_U	λ_V	λ_U	λ_V
PMF	0.01	10000	10	100	0.1	0.1
CTR	100	1	10	100	10	0.1
CDL	10	100	100	10	0.1	100
ConvMF	100	10	10	100	1	100
AMF	**10**	**60**	**10**	**60**	**1**	**60**

Table 2: Parameter Setting of λ_U and λ_V

Model	ML-1m	ML-10m	AIV
PMF	0.8961	0.8312	1.412
CTR	0.8968	0.8276	1.552
CDL	0.8876	0.8176	1.3694
ConvMF	0.8578	0.7995	1.209
AMF	**0.8359**	**0.7834**	**1.106**
Improvement	2.19%	1.61%	10.3%

Table 3: RMSE

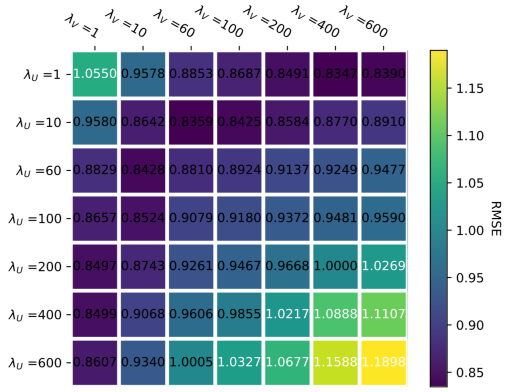


Figure 3: Parameter analysis of λ_U and λ_V on ML-1m dataset

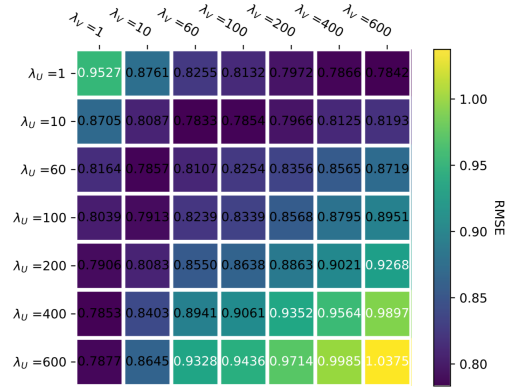


Figure 4: Parameter analysis of λ_U and λ_V on ML-10m dataset

from 0 to 1. The best performance values of parameters λ_U, λ_V of each model are described in Table 2.

4.2 Experimental Results

1) Evaluate Results.

Table 3 evaluates rating prediction error of our AMF model and four competitors. Note that "Improvement" shows the relative improvements of AMF over the best competitor. AMF achieves better performance than ConvMF, CDL, CTR, PMF. Specifically, our AMF has strong effectiveness on sparse dataset that is AIV data.

With MovieLens, the improvements of AMF over the best competitor, ConvMF, are 2.19% on ML-1m and 1.61% on ML-10m.

With AIV data, the improvement of AMF over the best competitor, ConvMF, is 10.3%.

2) Evaluate Results Over Sparseness Datasets. We set the different sparsenesses by randomly sampling with ML-1m, ML-10m and AIV datasets. Table 4 shows AMF still has robust and good performance when compared with the best competitor (ConvMF). This implies the effectiveness of incorporating item genre information in attention mechanism. Specifically, we observe that the improvements of AMF over ConvMF are 2.81% on ML-1m and 2.35% on ML-10m and 14.81% on AIV when training set is only 20%.

3) Impact of Parametes.

Figure 3, 4, 5 show the impact of λ_U and λ_V for three datasets ML-1m, ML-10m and AIV. We see

	ML-1m			ML-10m			AIV		
model	20%	40%	80%	20%	40%	80%	20%	40%	80%
ConvMF	0.9477	0.8949	0.8578	0.8896	0.8515	0.7995	1.4426	1.3584	1.2090
AMF	**0.9196**	**0.8755**	**0.8359**	**0.8661**	**0.8255**	**0.7834**	**1.2945**	**1.2171**	**1.1060**
Improvement	2.81%	1.94%	2.19%	2.35%	2.6%	1.61%	14.81%	14.13%	10.30%

Table 4: RMSE over sparseness of datasets

Model	Using Information	ML-1m	ML-10m	AIV
ConvMF	Item reviews	0.8578	0.7995	1.209
Concatenation	Item reviews + Item genre with concatenation	0.8513	0.8161	1.1891
AMF	**Item reviews + Item genre with attention**	**0.8359**	**0.7834**	**1.106**

Table 5: Comparing RMSE between AMF, Concatenation and ConvMF

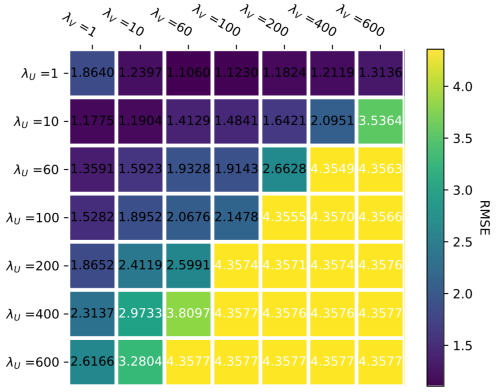


Figure 5: Parameter analysis of λ_U and λ_V on AIV dataset

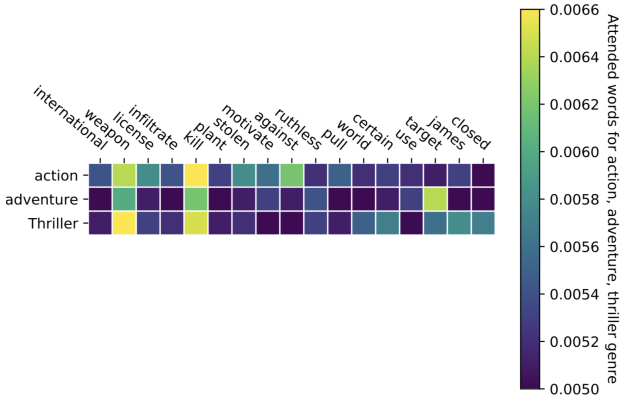


Figure 6: Attended feature of text comments

that when rating data becomes sparse, λ_U and λ_V decrease to produce the best results. In fact, the values of (λ_U, λ_V) of AMF are (10, 60), (10, 60) and (1, 60) on ML-1m, ML-10m and AIV, respectively.

4) Impact of Attention

We analyze the effectiveness of attention mechanism to document latent vector which improve rating prediction accuracy. We compare with another implementation that is *concatenation* between item reviews and item genre information.

In Table 5, our AMF model still has better performance than concatenation method. The results also show that concatenation method has better performance than ConvMF on ML-1m and AIV datasets. Specifically, we observe that the improvements of AMF over concatenation method are 1.54% on ML-1m and 3.27% on ML-10m. In the case AIV, it has strong effectiveness with 8.31% of improvement.

Figure 6 is our case study, we figure out the output of our model using attention mechanism with item genre information. The highlight points are the features attended by item genre information. These features have strong effectiveness in improving rating prediction accuracy. Moreover, they also help us to understand reviews document for items easily.

In Figure 6, we observed as follows.

- When item genre is **action**, the words **weapon**, **kill**, **against** are attended.
- When item genre is **adventure**, the words **weapon**, **kill**, **target** are attended.
- When item genre is **thriller**, the words **weapon**, **kill** are attended.

5 Conclusion

In this paper, we proposed AMF model that combines attention mechanism into PMF to enhance the rating prediction accuracy. Extensive results demonstrate that attention mechanism of AMF significantly outperforms the other competitors, which implies that AMF deals with the data sparsity problem by adding item genre information. Moreover, our model can figure out attended features for item reviews document which make us understand which information is attended from item reviews document.

References

Gediminas Adomavicius and Alexander Tuzhilin. 2005. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Transactions on Knowledge and Data Engineering*, 17(6):734–749.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In Geoffrey J. Gordon, David B. Dunson, and Miroslav Dudk, editors, *AISTATS*, volume 15 of *JMLR Proceedings*, pages 315–323. JMLR.org.

J.L. Herlocker, J.A. Konstan, L.G. Terveen, and J.T. Riedl. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems*, 22(1):5–53.

Dong Hyun Kim, Chanyoung Park, Jinoh Oh, Sungyoung Lee, and Hwanjo Yu. 2016. Convolutional matrix factorization for document context-aware recommendation. In Shilad Sen, Werner Geyer, Jill Freyne, and Pablo Castells, editors, *RecSys*, pages 233–240. ACM.

Y. Koren, R. Bell, and C. Volinsky. 2009. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37.

Guang Ling, Michael R. Lyu, and Irwin King. 2014. Ratings meet reviews, a combined approach to recommend. In Alfred Kobsa, Michelle X. Zhou, Martin Ester, and Yehuda Koren, editors, *RecSys*, pages 105–112. ACM.

Y. Liu, S. Wang, M. S. Khan, and J. He. 2018. A novel deep hybrid recommender system based on auto-encoder with neural collaborative filtering. *Big Data Mining and Analytics*, 1(3):211–221, Sep.

Julian J. McAuley and Jure Leskovec. 2013. Hidden factors and hidden topics: understanding rating dimensions with review text. In Qiang Yang, Irwin King, Qing Li, Pearl Pu, and George Karypis, editors, *RecSys*, pages 165–172. ACM.

Ankur P. Parikh, Oscar Tckstrm, Dipanjan Das, and Jakob Uszkoreit. 2016. A decomposable attention model for natural language inference. *CoRR*, abs/1606.01933.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *EMNLP*, volume 14, pages 1532–1543.

Ruslan Salakhutdinov and Andriy Mnih. 2008. Probabilistic matrix factorization. In *Advances in Neural Information Processing Systems*, volume 20.

Chong Wang and David M. Blei. 2011. Collaborative topic modeling for recommending scientific articles. In Chid Apt, Joydeep Ghosh, and Padhraic Smyth, editors, *KDD*, pages 448–456. ACM.

Hao Wang, Naiyan Wang, and Dit-Yan Yeung. 2015. Collaborative deep learning for recommender systems. In Longbing Cao, Chengqi Zhang, Thorsten Joachims, Geoffrey I. Webb, Dragos D. Margineantu, and Graham Williams, editors, *KDD*, pages 1235–1244. ACM.

L. Zhang, T. Luo, F. Zhang, and Y. Wu. 2018. A recommendation model based on deep neural network. *IEEE Access*, 6:9454–9463.

Identifying Adversarial Sentences by Analyzing Text Complexity

Hoang-Quoc Nguyen-Son[1], Tran Phuong Thao[2], Seira Hidano[1], and Shinsaku Kiyomoto[1]
[1]KDDI Research, Inc.
2-1-15 Ohara, Fujimino, Saitama, 356-8502 Japan
{ho-nguyen, se-hidano, kiyomoto}@kddi-research.jp
[2]The University of Tokyo
7-3-1 Hongo, Bunkyo, Tokyo, 113-8656, Japan
tpthao@yamagula.ic.i.u-tokyo.ac.jp

Abstract

Attackers create adversarial text to deceive both human perception and the current AI systems to perform malicious purposes such as spam product reviews and fake political posts. We investigate the difference between the adversarial and the original text to prevent the risk. We prove that the text written by a human is more coherent and fluent. Moreover, the human can express the idea through the flexible text with modern words while a machine focuses on optimizing the generated text by the simple and common words. We also suggest a method to identify the adversarial text by extracting the features related to our findings. The proposed method achieves high performance with 82.0% of accuracy and 18.4% of equal error rate, which is better than the existing methods whose the best accuracy is 77.0% corresponding to the error rate 22.8%.

1 Introduction

The computer-generated text has achieved remarkable success in replacing human roles in interactive systems such as question answering and machine translation. However, aside the positive impacts, an adversary takes advantage of the text to fool the judgment systems which are even unrecognized by human-beings themselves. The fake political attitudes and product previews, for instances, have significantly affected the awareness of real audiences. It raises the urgent task which has to efficiently identify the adversarial text before it is spread through the public media.

Previous methods have focused on proposing classifications to detect computer-generated text that is used in various unscrupulous purposes. More particularly, Labbé and Labbé (2013) created a detector[1] to identify a dummy paper by using text similarity. Other methods recognized untrusted information from machine translation text based on N-gram model (Aharoni et al., 2014) and word matching (Nguyen-Son et al., 2018; Nguyen-Son et al., 2019). However, according to the dramatic development of enhanced technologies, especially in deep learning era, adversarial text generated from the deep networks (Iyyer et al., 2018; Liang et al., 2018) can bypass in both existing work and human recognition.

We aim to investigate the adversarial texts (Iyyer et al., 2018) which have different syntax structure with the original text due to the difficulty in tracing their origins. Such texts are more dangerous than other easily tractable texts which are simply created by word or phrase modification (Liang et al., 2018; Ebrahimi et al., 2018). Moreover, the adversarial texts considered in this paper really fool AI systems. One of the actual adversarial texts is picked from the development set as represented in Figure 1. The samples are movie reviews whose interest rates are represented with the number of stars determined by a common sentiment analysis system (Socher et al., 2013). In these reviews, the level of the adversarial text is labeled with four-star, the original is more interested with five.

The original sentence is often more complex than the adversarial text in both word usage and text

[1]http://scigendetection.imag.fr/main.php

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 182-190, Hakodate, Japan, September 13-15, 2019

Human text ★★★★	*For the most part, **it's** a work of incendiary genius, steering clear of knee-jerk reactions and quick solutions.*
Adversarial text ★★★★★	*For the most part **is** the work of incendiary genius, steering clear of various responses and quick solutions.*

Figure 1: Human-created original vs machine-generated adversarial text.

structure. Human writers express their intentions via fashionable and modern words, such as "*knee-jerk*" and "*reactions*". In contrast, the adversarial text is optimized its readability by simple common words, i.e. "*various*" and "*responses*." Moreover, human tends to use flexible structural utterances. The flexibility is illustrated with the use of the complex expression "*most part, it's a work*" instead of "*most part is the work.*"

Contribution In this paper, we analyze the most threatening adversarial text which not only fools the recent AI system but also is difficultly tracked because of changing the original structure. We thereafter propose the following process to distinguish the adversarial with the original text:

- We estimate the text coherence by matching words and measuring the word similarities. Only the high similarities which mainly construct the coherence are distributed into certain groups depending on the part of speech (POS) of the matched words. In each group, the similarities are normalized by the means and the variances that represent for the coherence features.

- We suggest using frequencies to exploit the difference of word usage in original and adversarial text. In particular, the frequencies are allocated to appropriate POS groups and are used as frequency features.

- We design other features to address the optimization problems of generating adversarial text. More especially, we notice that the adversarial text is short and may contain successive duplicate phrases. We thus integrate sentence length with the number of duplicate phrases in order to extract the optimization features.

- We combine our features with the features extracted from the N-gram language model for determining whether the input sentence is an original text written by a human or an adversarial text generated by a machine.

To evaluate our method, we use 11K original sentences from a common movie review corpus[2]. We then generate the corresponding adversarial text using the syntax-based system by Iyyer et al. (2018). Afterward, we select approximate 1500 pairs which are classified in different sentiment labels by a Stanford system (Socher et al., 2013). The result shows that our method achieves high performance with 82.0% of the accuracy and 18.4% of the equal error rate. It outperforms the existing state-of-the-art method whose accuracy is 77.0% and equal error rate is 22.8%.

Roadmap The rest of this paper is organized as follows. The related work is presented in Section 2. Our proposed method is described in Section 3. The evaluation is given in Section 4. Finally, Section 5 summarizes some main key points and mentions future work.

2 Related Work

In this section, we present previous work in two aspects: methods for generating adversarial texts and methods for identifying adversarial texts.

2.1 Adversarial Text Generation

There are two approaches for generating adversarial text: non-syntax-based and syntax-based. In the first approach, an adversary modifies some parts of the original text but still preserve its structure. On the other hand, in the second approach, the adversary changes the text's syntax to deceive the AI systems.

Non-syntax-based Approach Liang et al. (2018) changed the salient text components via white-box attack using cost gradients or black-box attack using occluded samples. The modification can apply for

[2] http://nlp.stanford.edu/~socherr/stanfordSentimentTreebank.zip

both *character* and *word* levels in order to generate robust adversarial text which efficiently fools a multiclass classifier related to news posts. Also targeting on this classifier, Ebrahimi et al. (2018) generated the adversarial text by using a set of operations on *characters* such as flip, insertion, and deletion based on the one-hot vectors extracted from the input text. Besides the multiclass classification, the adversarial text is also able to attack the question answering (QA) system. For instance, Jia and Liang (2017) padded a distract *sentence* into text for changing the answers of the QA system. Furthermore, Ribeiro et al. (2018) can generate the adversarial text which deceives both AI systems including QA and sentiment analysis. More especially, the authors used a set of rules on *phrase* to generate adversarial questions which look alike to the origins but change the results of these systems. Alzantot et al. (2018) also proved that the adversarial text affords to fool various AI systems, namely sentiment analysis and textual entailment, using *word* replacements.

Syntax-based Approach Most previous work from the non-syntax-based approach that we mentioned above adapts the operations such as modifications, insertions, or deletions on various text levels: characters, words, phrases, and sentences. Due to the unchanged structures, such texts easily trace back to their origins and these generated texts are easily filtered. In the opposite way, the syntax-based approach addresses more serious adversarial texts when the structures are changed, so such texts are easy to mix with their origins without being detected. We, therefore, focus on this approach instead of the other. In the syntax-based approach, Iyyer et al. (2018) generated a paraphrase with a desired syntax by using attention networks to transfer the text structure. Such adversarial texts can target to two current popular risks: (i) fake reviews by fooling sentiment analysis system, and (ii) political posts by deceiving the textual entailment.

2.2 Adversarial Text Detection

Previous work is categorized into four approaches: parse tree, word distribution, N-gram model, and word similarity.

Parse Tree Li et al. (2015) prove that the syntactic structure of a human-written sentence is more complex than that of computer-generated one because the simple artificial text is often created to prevent the mistakes in both grammar and semantic. The structure of the simple text is well-balanced, so the authors extracted some related features, i.e. the ratio of right/left-branching notes in various scopes: main constituents and whole sentence. Some surface and statistical features were also used to including parse tree depth, sentence length, and out-of-vocabulary words. The main drawback of parse tree approach is that it only investigates on text syntax but ignore the semantics itself.

Word Distribution The word distribution in the large text is used to classify computer- and human-generated text. For example, Labbé and Labbé (2013) indicate the high similarity of the distribution in artificial documents. They suggested a metric, namely inter-textural distance, to measure the similarity between two texts. It can be used to identify fake academic papers with impressive accuracy. More general, Nguyen-Son et al. (2017) used Zipfian distribution to identify other texts. Additional features extracted from humanity phrases (e.g., idiom, cliché, ancient, dialect, and phrasal verb) and co-reference resolution were also applied to improve their result. The main drawback of word distribution approach is that a large number of words are required. This limitation is also confirmed by the authors of both the inter-textual distance and the Zipfian distribution.

N-gram Model The common method to estimate the fluency of continuous words is to use the N-gram model. Many researchers have measured this property on discontinuous words and combine with the N-gram model. For instance, Arase and Zhou (2013) used sequential pattern mining to extract fluent human patterns such as "*not only * but also*" and "*more * than*." They contrasted with the weird patterns (e.g., "*after * after the*" and "*and also * and*") in machine-generated texts from low-resource languages. Nguyen-Son and Echizen (2017) extracted features from two types of noise words: (i) the humanity words from a user message, such as misspelled (e.g., comin, hapy) and short-form/slang words (e.g., tmr, 2day), (ii) the untranslated words from a machine message. This approach, however, is only suitable for social network texts that abun-

dantly contained substantial noise words. On the other hand, Aharoni et al. (2014) targeted functional words that are often chosen by a machine for improving the readability of the generated text.

Word Similarity Nguyen-Son et al. (2018) proposed the classification based on the idea that: the coherence between words in a computer-generated text is less than that in a human-generated text. They matched similar words in every pair of sentences in a paragraph using Hungarian maximum matching algorithm. More particularly, each word was matched with the most similar word in another sentence. The drawback of this work is that the relationships between words inside a sentence are not considered so it cannot be applied for individual sentences as targeted in this paper. Nguyen-Son et al. (2019) overcome the limitation by matching similar words in the whole text. The maximum similarity for each word was used to estimate the coherence while the other similarities are dismissed. For coherence features mentioned in the Section 3.1, we indicate that the other high similarities are also useful to measure the text coherence and efficient to identify the adversarial text.

3 Proposed Method

The overview of the proposed method is formalized as four parallelizable steps:

- **Step 1 (*Matching words*)**: Each word is matched with the other words, and their similarities are estimated by Euclidean distances in word embedding. These similarities represent the connections of words and are thus used to extract coherence features.

- **Step 2 (*Estimating frequencies*)**: The frequencies of individual words are inferred from corresponding items of Web 1T 5-gram corpus. The frequency indicates the popularity of the word in various context usages.

- **Step 3 (*Finding optimization issues*)**: Optimization issue features which result from the optimization process of adversarial text generation are extracted from sentence length and successive duplicated phrases.

- **Step 4 (*Extracting word N-gram*)**: The text fluency is measured by using the word N-gram model. N continuous words with N between 1 to 3 are used for this model.

The details of each step using the examples mentioned in Figure 1 are described in the following subsections.

3.1 Matching Words (Step 1)

Words in the input text are separated and tagged with the part-of-speeches (POSs) using a Stanford tagger (Manning et al., 2014). Each word is matched with the other words, then their similarities are estimated. For inferring the similarity between two words, we measure the Euclidean distance of their vectors in word embeddings. The higher similarity of two words results in the lower distance. The GloVe corpus (Pennington et al., 2014) is used to map the words, collected from Wikipedia and Gigaword, with 300-dimensional vectors. The similarities of some words in the human-generated text are illustrated in Figure 2 with the POS taggers denoted by subscripts.

A machine tends to create a simple text so that the text's meaning and readability are preserved. The generated text is thus generally shorter than the human-generated one, as also claimed by Volansky et al. (2013). The long expression of the original text compared to the short of the adversarial text is shown in Figure 2 and Figure 3, respectively. The additional words and their connections with other words are marked in bold to emphasize the difference. The small values of the distances demonstrate the tight connections that are created by these padding words. These connections do not influence the overall meaning but slightly improve the text coherence.

The high similarities with small distances have higher impacts on the text coherence than the low similarities. Therefore, we only choose the highs and eliminate the others. We suggest a threshold of α to determine the ratio between them. The α is set to 0.7 after being optimized from the development set as mentioned in Section 4. It means that only 70% of the high similarities are selected while the remaining is removed as presented by double strikethough numbers linked to dashed-line connections.

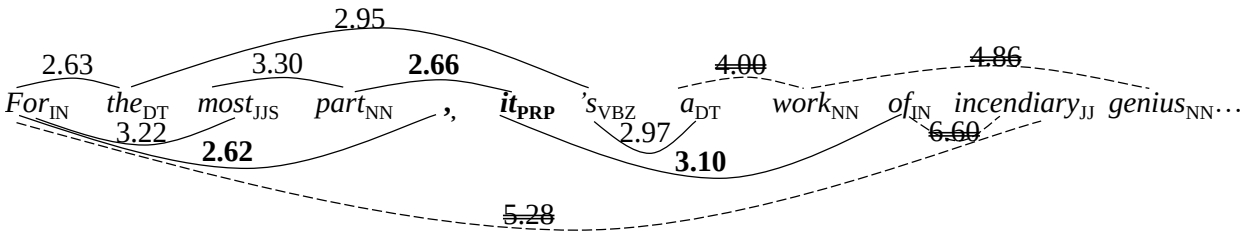


Figure 2: Matching words in the human original text.

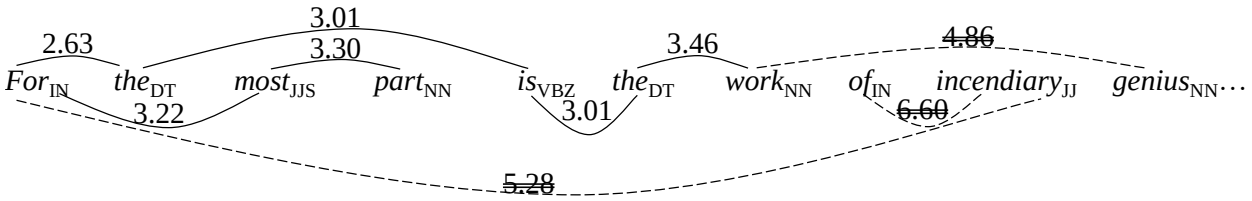


Figure 3: Matching words in the machine adversarial text.

According to the POS tags, words play different roles in a certain sentence. Major words like nouns (NN) and verbs (VB) have higher influential than minor ones such as determiners (DT) and prepositions (IN). We thus distribute the high similarities to appropriate POS groups. In Figure 4, two pairs "*the*$_{DT}$–*'s*$_{VBZ}$" (2.95) and "*'s*$_{VBZ}$–*a*$_{DT}$"(2.97) are allocated to the same group, i.e. DT–VBZ. We use all of 45 POS tags containing in the training set and produce 1035 possible combinations in total.

The individual POS groups often contain different numbers of similarities. The numbers in each group are normalized by using the means and the variances. Total 2070 statistical values are calculated for all POS groups. These values are used as the features representing the coherence of the text.

3.2 Estimating Frequencies (Step 2)

After splitting and tagging POSs, we estimate the popularity of the words by using their frequencies in Web 1T 5-gram corpus[3]. This corpus counts the number occurrences of around 14 million common words in approximately 95 billion sentences extracted from available web pages. The frequencies of several words in the human and the adversarial text are shown in Figure 5. The words occurring in the only human text are underlined while the other differences are marked in bold. All non-displayed words are identical in the two texts.

The machine-generated text is often optimized with "safe words" which are commonly used in other contexts. It explains that the frequencies of the adversarial words are slightly higher than those of human words. More especially, among the synonyms, the adversarial text tends to select high-frequency words, for instance, "*responses*" (19E+6) instead of "*reactions*" (7E+6). On the other hand, in case of the same word meaning in the context, the standard words such as "*is*" and "*the*" have higher selection's priorities than the words "*'s*" and "*a*," respectively. The writing styles of native speakers are very flexible, they can creatively choose "fashionable words" fitting to the context. For example, since "*knee-jerk*" is rarely used, it is out of the highest frequency words even in the large 1 terabyte tokens of the Web 1T 5-gram corpus.

Like the process of Step 1, we distribute the word frequencies into specific groups based on the POS tags. For instance, the two nouns "*part*" and "*work*" are delivered to the same group of nouns (NN) as illustrated in Figure 6. We also normalize the frequencies within the individual groups by the means and the variances for extracting the final frequency features.

3.3 Finding Optimization Issues (Step 3)

The optimization process of adversarial text generation may cause the appearances of successive dupli-

[3] https://catalog.ldc.upenn.edu/LDC2006T13

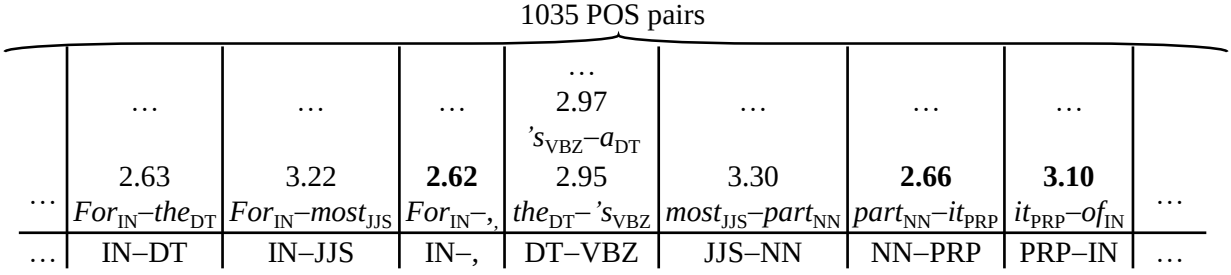


Figure 4: Distributing high similarities of the human text.

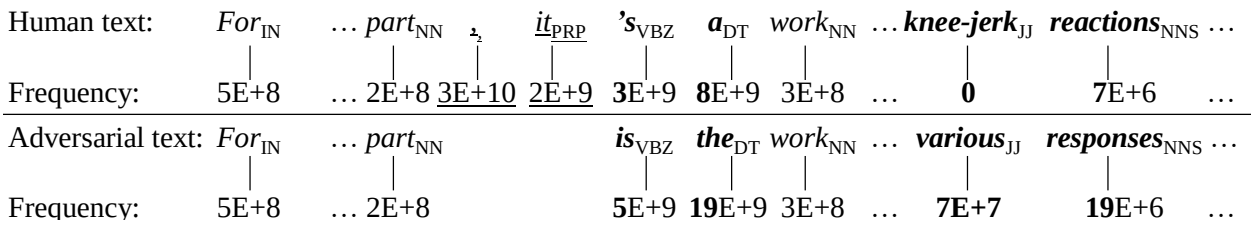


Figure 5: Calculating word frequencies in the human and the adversarial text.

cated phrases. We, therefore, extracted the phrase-related features by counting the numbers of such phrases with the length varying from 1 to 5. In Figure 7, since the adversarial sentence have two successive duplicate phrases "***own***," the 1-phrase-length feature is equal to 2.

Another issue of the optimization process is that a machine often generates a simple short text. In other words, the machine practically selects candidates with a minimal number of words to express a certain intention. Consequently, such generated texts are shorter than analogous texts written by a human. To deal with this problem, we simply count the number of words and denote them as the length feature. This length feature is integrated with the phrase-related features above and they are served as the optimization features.

3.4 Extracting Word N-gram (Step 4)

To evaluate the frequency of the text, we inherit the N-gram model to extract continuous POS phrases with the length up to 3. Some extracted phrases from the human text are listed in Figure 8. We use the POS tags for the model instead of the word because the similar phrases having the same structure can be recognized. For example, the first pattern "IN DT" represents not only for the phrase "*For the*" but also for other identical structural ones such as "*For a*" and "*In the*."

4 Evaluation

4.1 Dataset

We created the experimental data by using 11,855 sentences from a movie review corpus[4]. These sentences were inputted to the syntactically controlled paraphrase networks (SCPN)[5] to generate adversarial text. We only proceeded the input sentences which can produce the adversarial texts actually fooled the well-known sentiment analysis system (Socher et al., 2013). At the result, the 1489 inputs were considered as human-written text while the corresponding sentences were denoted as machine-generated text.

The 2978 satisfied sentences were split into three sets: for training, for development, and for test phases with the ratio as 60%, 20%, and 20%, respectively. To balance the human and the machine sentences in each set, we put a pair of the original and the adversarial sentences into the same set. The development test was used to determine the threshold α described in Section 3.1. Two pairs in development set are shown in Figure 1 and Figure 7.

[4] http://nlp.stanford.edu/~socherr/stanfordSentimentTreebank.zip

[5] https://github.com/miyyer/scpn

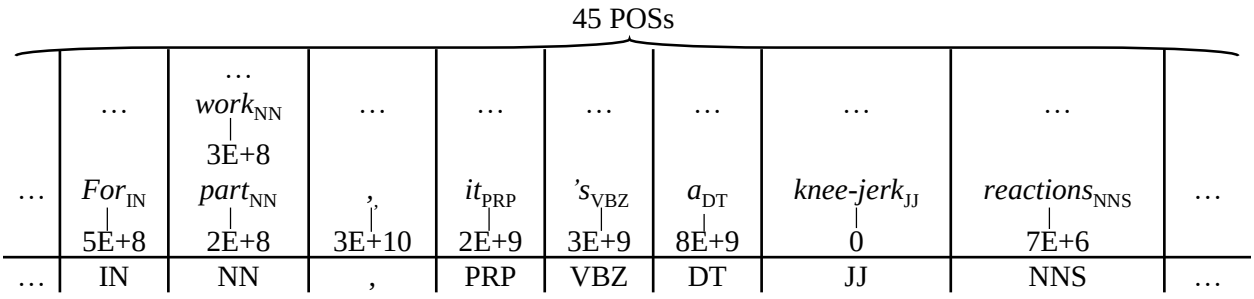


Figure 6: Distributing word frequencies of the human text.

Human text ★★	*the whole mildly pleasant outing -- the r rating is for brief nudity and a grisly corpse -- remains aloft not on its own self-referential hot air, but on the inspired performance of tim allen.*
Adversarial text ★★★★	*the whole mildly pleasant thing - the r rating is for short nudity and a grisly corpse - is still not on its **own** **own** hot air , but on the inspired the above the possible the right.*

Figure 7: Successive duplicated phrases in the adversarial text.

Human text: For_{IN} the_{DT} $most_{JJS}$ $part_{NN}$ …
POS *N*-gram={"IN," "DT," "JJS," "NN,"… …"IN DT," "DT JJS," "JJS NN,"… …"IN DT JJS," "DT JJS NN,"…}

Figure 8: Extracting POS N-gram from the human text.

4.2 Individual Features and Combinations

We conducted experiments on our individual features and their combinations. The experiments were run with three common machine learning algorithms including logistic regression (LOGISTIC), support vector machine (SVM) optimized by stochastic gradient descent (SGD(SVM)), and SVM optimized by sequential minimal optimization (SMO(SVM)). The results are summarized in Table 1 with individual features in the top rows and their combinations in the bottoms. For assessing the performances on the test set, we used two standard metrics: the accuracy and the equal error rate (EER).

In four groups of individual features, the experiment on optimization gives low results. It indicates that the surface information extracted from the internal input sentence is insufficient to identify adversarial text. The use of external knowledge such as the frequency can improve the performances. However, the frequency is limited to separate words and ignore the mutual connections of them. On the other hand, the coherence features based on these connections improve both accuracy and EER metrics. The POS N-gram features achieve better performances and point out the low fluency of the adversarial text.

In combinations, while the frequency features target on individual words, the coherence features examine the mutual connections among them. They support each other to raise the overall performances. The combination with the features based on optimization problems of adversarial generators even achieves better results. Finally, each individual exploits the different aspects of adversarial problems, so all features put together can establish the new milestone with the best accuracy up to 82.0%.

4.3 Comparison

We compare our method with previous work on identifying machine-generated text. The results of the comparison are provided in Table 2 with the highest performances marked in bold. The first method (Nguyen-Son et al., 2017) verified the word distribution with Zipf's law. In the second method, Li et al. (2015) extracted features from the parsing tree and used them for classifiers. The most similar method to our coherence features (Nguyen-Son et al., 2019) matched similar words within the text and manipulates on maximum similarity. Finally, the last method (Aharoni et al., 2014) combined POS

No	Features	LOGISTIC		SGD(SVM)		SMO(SVM)	
		Accuracy	EER	Accuracy	EER	Accuracy	EER
1	Coherence features	60.5%	39.7%	68.7%	28.0%	**73.8%**	**25.6%**
2	Frequency features	**71.5%**	**28.3%**	69.0%	28.4%	69.2%	30.0%
3	Optimization features	**70.2%**	**29.8%**	69.2%	32.6%	66.7%	36.6%
4	POS N-gram features	57.8%	41.7%	65.4%	35.1%	**75.2%**	**25.1%**
5	1 + Frequency features	60.8%	38.3%	72.5%	**22.7%**	**74.0%**	26.6%
6	5 + Optimization features	61.0%	39.0%	76.2%	26.3%	**77.7%**	**23.2%**
7	All features	67.3%	32.3%	81.2%	**14.7%**	**82.0%**	18.4%

Table 1: Individual features and combinations.

Method	LOGISTIC		SGD(SVM)		SMO(SVM)	
	Accuracy	EER	Accuracy	EER	Accuracy	EER
Nguyen-Son et al. (2017)	66.5%	33.0%	64.5%	32.9%	**67.3%**	**25.9%**
Li et al. (2015)	67.5%	32.3%	66.3%	34.1%	**68.7%**	**31.1%**
Nguyen-Son et al. (2019)	60.2%	40.0%	64.0%	35.9%	**73.3%**	**21.1%**
Aharoni et al. (2014)	59.5%	40.3%	66.0%	34.2%	**77.0%**	**22.8%**
Our (All features)	67.3%	32.3%	81.2%	**14.7%**	**82.0%**	18.4%

Table 2: Comparison with other methods.

N-gram model with functional words to identify the machine text.

The word-distribution-based method (Nguyen-Son et al., 2017) is suitable for large text, e.g., document and web page, because of needing large words to adapt to the Zipf's law; but the performance is positively affected on the sentence level. On the other hand, the syntax-based method (Li et al., 2015) seems more appropriate with this task but this work merely focuses only on text structure and dismisses the intrinsic meaning. Besides that, the previous coherence-based method (Nguyen-Son et al., 2019) only used a maximum similarity of each word rather than near-optimal similarities. Therefore, this work is more appropriate to a paragraph than a sentence, which has a limit number of words. In another approach, the adding of function words into POS N-gram model (Aharoni et al., 2014) can improve the SMO(SVM) classifier. Among these classifiers, our method is the most stable especially in both SVM classifiers with the highest accuracy of 82.0%.

5 Conclusion

We have investigated the issues from one of the most harmful adversarial texts which are generated by changing the structures of the original texts. Although the adversarial generators can produce understandable texts, which preserve the meaning of the origins; the coherence and fluency of the generated texts still have limits. Moreover, a person has a higher probability to create a professional text by using flexible words. In another aspect, the optimization process leads to the adversarial texts incurring some artificial phenomenal such as a shortage in length or duplication in phrases. Based on the findings, we propose a method to identify the adversarial texts by suggesting distinguishable features with the original texts. The results of the evaluation on the adversarial texts generated from the movie review corpus show that our proposed method achieves high performance: 82.0% of the accuracy and 18.4% of the EER which are greater than related methods with the best accuracy 77.0% and EER 22.8%.

In future work, we improve the proposed features by using deep learning networks and identify other harmful adversarial texts such as product reviews and political comments. We also improve the quality of useful machine-generated texts based on our analysis in this paper.

References

Roee Aharoni, Moshe Koppel, and Yoav Goldberg. 2014. Automatic detection of machine translated text and translation quality estimation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 289–295.

Moustafa Alzantot, Yash Sharma Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang. 2018. Generating natural language adversarial examples. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2890–2896.

Yuki Arase and Ming Zhou. 2013. Machine translation detection from monolingual web-text. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1597–1607.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou. 2018. Hotflip: White-box adversarial examples for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 31–36.

Mohit Iyyer, John Wieting, Kevin Gimpel, and Luke Zettlemoyer. 2018. Adversarial example generation with syntactically controlled paraphrase networks. In *Proceedings of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 1875–1885.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2021–2031.

Cyril Labbé and Dominique Labbé. 2013. Duplicate and fake publications in the scientific literature: how many scigen papers in computer science? *Scientometrics*, 94(1):379–396.

Yitong Li, Rui Wang, and Hai Zhao. 2015. A machine learning method to distinguish machine translation from human translation. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation (PACLIC)*, pages 354–360.

Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. 2018. Deep text classification can be fooled. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4208–4215.

Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. 2014. The Stanford CoreNLP natural language processing toolkit. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL) - System Demonstrations*, pages 55–60.

Hoang-Quoc Nguyen-Son and Isao Echizen. 2017. Detecting computer-generated text using fluency and noise features. In *Proceedings of the International Conference of the Pacific Association for Computational Linguistics (PACLING)*, pages 288–300.

Hoang-Quoc Nguyen-Son, Ngoc-Dung T Tieu, Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. 2017. Identifying computer-generated text using statistical analysis. In *2017 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC)*, pages 1504–1511.

Hoang-Quoc Nguyen-Son, Ngoc-Dung T Tieu, Huy H Nguyen, Junichi Yamagishi, and Isao Echizen. 2018. Identifying computer-translated paragraphs using coherence features. *ArXiv Preprint arXiv:1812.10896*.

Hoang-Quoc Nguyen-Son, Tran Phuong Thao, Seira Hidano, and Shinsaku Kiyomoto. 2019. Detecting machine-translated paragraphs by matching similar words. In *ArXiv Preprint arXiv:1904.10641*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2018. Semantically equivalent adversarial rules for debugging nlp models. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 856–865.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642.

Vered Volansky, Noam Ordan, and Shuly Wintner. 2013. On the features of translationese. *Digital Scholarship in the Humanities*, 30(1):98–118.

Phi-Agreement by C in Japanese:

Evidence from Person Restriction on the Subject

Miki Obata
Hosei University
3-7-2 Kajino-cho
Koganei, Tokyo
obata@hosei.ac.jp

Mina Sugimura
Kyoto Notre Dame University
1 Minaminonogami-cho
Shimogamo Sakyo-ku, Kyoto
sugimura@notredame.ac.jp

Abstract

The goal of this work is to present additional empirical evidence for C-agreement in Japanese. We specifically focus on the interaction between discourse modals and the verb *give* in Japanese. By extending Miyagawa's (2017) analysis of politeness marking in Japanese, we demonstrate that C in Japanese triggers phi-agreement with the subject without transferring phi-features to T.

1 Phi-Agreement by C in Japanese

Since Chomsky (2000), phi-agreement has played an important role, especially in constructing syntactic dependencies between linguistic elements. Under Chomsky's (2007) feature inheritance, unvalued phi-features are introduced to the derivation with phase heads (C/v) and inherited by non-phase heads (e.g. T/V). In English, for example, phi-features transferred from C to T agree with the subject, and nominative Case is assigned as the reflex of agreement. By extending this system, Miyagawa (2017) suggests that languages can be categorized into the following four types:

(1) a. Category I: $C_{ø}$, T_{δ} - Japanese
b. Category II: C_{δ}, $T_{ø}$ - English
c. Category III: C, $T_{\delta/ø}$ - Spanish
d. Category IV: $C_{\delta/ø}$, T - Dinka
(Miyagawa 2017: 18)

In Category I and IV, phi-features are not inherited by T but stay at C, in contrast to Category II and III. (δ stands for topic/focus features, which we put aside for ease of discussion in this paper.) (See also Ouali 2006 for relevant discussion.) As supporting evidence for Category I, Miyagawa demonstrates that phi-agreement by C with a speech act head takes place for politeness marking in Japanese:

(2) a. Watasi-wa pizza-o tabe-mas-u.(formal)
I-Top pizza-Acc eat-MAS-Pres
"I will eat pizza."
b. Watasi-wa pizza-o tabe-ru. (colloquial)
I-Top pizza-Acc eat-Pres
"I will eat pizza."
(Miyagawa 2017: 18)

The morpheme *-mas-* in (2a) is the politeness marker in Japanese. An appropriate form needs to be chosen depending on whom the speaker is talking to.

Miyagawa finds allocutive agreement in Souletin (an eastern dialect of Basque) similar to Japanese politeness marking:

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 191-195, Hakodate, Japan, September 13-15, 2019

(3) a. *To a male friend*
Pettek lane gin
Peter.Erg work.Abs do-Pres
di**k**.
Aux-3Sg.Abs-2Sg.Colloq.M-3Sg.Erg
allocutive agr. subj. agr.
'Peter worked.'

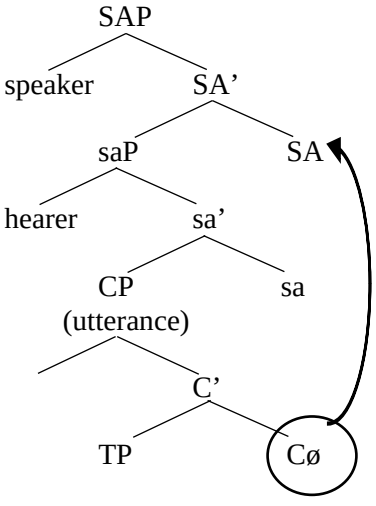
b. *To someone higher in status (formal)*
Pettek lane gin
Peter.Erg work.Abs do-Pres
diz**ü**.
Aux-3Sg.Abs-2Sg.Formal-Sg.Erg
'Peter worked.'

(Miyagawa 2017: 22)

If the speaker is talking to someone higher in status, the auxiliary is marked with the formal form as in (3b) in contrast to (3a), in which the speaker is talking to a (male) friend. According to Oyharçabal's (1993) analysis, allocutive agreement is triggered by C, since allocutive agreement morphemes are in competition with other materials at C, including question morphemes.

By extending Oyharçabal's analysis to Japanese politeness marking, Miyagawa (2017) proposes that C undergoes head movement to the head of the Speech Act Phrase (SAP), which is proposed in Speas and Tenny (2003) as well as in Haegeman and Hill (2011), and that phi-features on C are valued 2nd person formal, in which politeness marking is allowed:

(4)



(modified from Miyagawa 2017: 26)

In the following sections, we consider another type of phi-agreement by C in Japanese by examining cases of person restriction on the subject imposed by discourse modals and by the Japanese verb *give*, which supports Miyagawa's (2017) view.

2 Two types of Person Restrictions in Japanese: Discourse Modals and the Verb *kure(ru)* 'give'

2.1 Person Restriction on the Subject in Japanese Discourse Modals

Although phi-morphemes are rarely observed in Japanese, discourse modals (Inoue 2006), which express the speaker's attitude toward the utterance or the hearer, impose a specific person on the subject (i.e. person restriction):

(5) *Prohibition*
{*boku/✓kimi/*Taro}-wa sonnakoto
I/you/Taro-Top such a thing
kinisuru-**na**.
care-never
'You, not others, don't worry about such a thing.'

(6) *Intention*
{✓boku/*kimi/*kare}-ga sugu ik(u)-**oo**
I/you/he-Nom right now go-Intention
'I, not others, go there right now.'

(Ueda 2008: 134)

The prohibition marker *-na* in (5) requires a 2nd-person subject, while the intention marker *-oo* in (6) requires a 1st-person subject. It has been widely assumed that the discourse modals occur in the CP-domain (part of C, Rizzi 1997).

Although Ueda (2008) demonstrates that U(tterance)-modals including the morphemes *-na* in (5) and *-oo* in (6) impose a specific person on the subject, she does not discuss how exactly the agreement relation is constructed between the subject and C/Modal. In order to elucidate the dependency between them, we overview another type of person restriction observed in the verb *give* in Japanese and combine both types of person restriction in the next sections, which have important implications for how features on C and T are distributed in Japanese under Chomsky's (2007) feature inheritance.

2.2 Person Restriction on the Subject in the Japanese Verb *kure(ru)* 'give'

Person restriction on the subject is also triggered by the verb *give* (and only *give*) in Japanese (cf. Kuno and Kaburaki 1977). Since this type of person restriction is unique to this specific verb, it is arguably pre-specified in the lexicon:[1]

(7) *English 'give'*
{✓I/✓You/✓Hanako} gave Taro a book.

(8) *Japanese 'kure(ru)'*
{*Watasi/✓Anata/✓Hanako}-ga Taro-ni
I/you/Hanako-Nom Taro-Dat
hon-o kure-ta
book-Acc give-Past
'I/You/Taro gave Taro a book.'

Unlike in the English example (7), the verb *kure(ru)* 'give' disallows the 1st person subject as in (8), which exhibits person restriction.

How does person restriction imposed by the verb apply to the subject in the syntactic derivation? In English, on one hand, unvalued phi-features on T are valued by phi-features on the subject and no person restriction arises, as in (7). In (8), on the other hand, the specific verb *kureru* requires [2nd person] or [3rd person] for the subject. One might think that English also shows person restriction in the case of the 3rd person singular subject. The 3rd person singular subject requires overt inflection on T in English: once phi-features on T are valued by the 3rd person singular subject, the morpheme *-s* is inserted into T at the morpho-phonological level. This looks like person restriction, but note that neither T nor V limits the person of the subject in this case. Rather, T's realization varies depending on phi-features of the subject, which means this is not person restriction.

How can the agreement in (8) be implemented? Although agreement takes place between T and the subject, person restriction is a property of the verb *kureru*, not T. Following Obata and Sugimura's (2014) head movement analysis, a V-v-T amalgam is formed by head movement and the amalgam including T (not solely T) agrees with the subject and only [2nd] or [3rd] subject is ruled in:

(9) a. [T [Subj. v V …] *by head movement*

b. [V-v-T [Subj. <v> <V> …]
The V-v-T amalgam is formed and agrees with the subject for phi-features.

Before phi-agreement takes place, the V-v-T amalgam bears unvalued features: person is not specified as to [2nd] or [3rd], and number has no value yet, as in (10). Through phi-agreement with the subject, person is specified as either [2nd] or [3rd], and number gains a value.

(10) V-v-T
Per: [2nd/3rd]
Num: [---]

(11) a. *Subj.
Per: [1st]
Num: [Sg]

b. ✓Subj.
Per: [2nd]
Num: [Sg]

If the amalgam in (10) agrees with the subject in (11a), the underspecified person feature is never specified, which causes the derivation to crash. If the amalgam agrees with the subject in (11b), person is specified as [2nd] and number gains a value, satisfying the person restriction. Note that [3rd] on the amalgam becomes inactive after specification through phi-agreement with the subject and is no longer available. This is how person restriction is applied to the subject through phi-agreement by the V-v-T amalgam in the case of the verb *kureru*.[2]

3 Phi-Agreement by C vs. Phi-Agreement by T: Evidence from the Verb *kure(ru)* 'give'

In Section 2, we overviewed two types of person restrictions in Japanese. In the first case, discourse modals (i.e. C/Modal) impose a person restriction on the subject. In the second case, the verb *kure(ru)*, the V-v-T amalgam imposes a person

[1] Another verb, *ageru*, which also means 'to give', imposes person restriction on the dative object: the first person dative object is not allowed to occur. In Obata and Sugimura (2014), we suggest that person restriction of *ageru* is also pre-specified in the lexicon, just like in the case of *kureru*.

[2] In this paper, we assume that person restriction is applied through phi-agreement. Although Chomsky (2000) suggests that Case is assigned as a consequence of phi-agreement, we limit our discussion only to phi-agreement and do not go into controversial issues of Case-assignment in Japanese in this paper.

restriction on the subject. What happens if a person restriction is imposed on the subject both by discourse modals and by the verb *kure(ru)* in a single sentence?

(12) *Prohibition*
a. (kimi-wa) musuko-ni sonna hon-o
you-Top son-Dat such a book-Acc
kureru-na
give-never
'Don't give such a book to my son.'
b. {*watasi/*Taro}-wa musuko-ni
I/Taro-Top son-Dat
sonna hon-o kureru-na
such a book-Acc give-never

(13) *Intention*
{*boku/*kimi/*kare}-ga Taro-ni hon-o
I/you/he-Nom Taro-Dat book-Acc
kure-yoo
give-Intention
'I/you/he gives Taro a book.'

The verb *kure(ru)* co-occurs with the prohibition marker in (12) and the intention marker in (13). With respect to the person restriction, the verb requires either [2nd] or [3rd] and the modal requires [2nd] in (12), so that the sentence becomes grammatical only when the subject is [2nd] as in (12a). In (13), the verb imposes either [2nd] or [3rd] on the subject while the modal imposes [1st]. The sentence is ungrammatical with any subject.

What do (12) and (13) imply concerning phi-agreement by C and/or T? (14) and (15) show the logical possibilities for phi-feature distributions on C and T in (12) and (13), respectively.

(14) a. *NO inheritance from C to T in (12):*

C/Modal	V-v-T
[2nd]	[2nd] or [3rd]

b. *AFTER inheritance from C to T in (12):*

C/Modal		V-v-T
[---]	→	[2nd], [2nd] or [3rd]

(15) a. *NO inheritance from C to T in (13):*

C/Modal	V-v-T
[1st]	[2nd] or [3rd]

b. *AFTER inheritance from C to T in (13):*

C/Modal		V-v-T
[---]	→	[1st], [2nd] or [3rd]

No feature inheritance from C to T takes place in (14a)/(15a), while phi-features are inherited by T in (14b)/(15b). If phi-features on C/Modal are inherited by T, as in (14b)/(15b), all the phi-features are combined to form a single probe, which applies person restriction to the subject. If feature inheritance does not occur, as in (14a)/(15a), T (amalgam) and C serve as probes independently.

What does this difference predict? In (14b), if phi-features on C are inherited by T, either the [2nd] or [3rd] subject in (14b) (i.e. *kimi* 'you', *Taro*, respectively) should be allowed, despite the fact that the [3rd] subject in (12) is ungrammatical. Also in (15b), if feature-inheritance takes place from C to T, any subject (i.e. *boku* 'I', *kimi* 'you', *kare* 'he') should be allowed, contrary to fact. (Note that inherited [1st] in (15b) needs to be listed with [2nd] and [3rd] by disjunction (or), not by conjunction (and), since persons are mutually exclusive in nature and combining two different persons is logically impossible, independent of our discussion.)

If, however, C and T agree with the subject separately for phi-features, as in (14a)/(15a), the overgeneration mentioned above never happens. In both (14a) and (15a), T (amalgam) first imposes [2nd] or [3rd] on the subject. Then, C bearing [2nd] agrees with the subject in (14a). Thus, only when the subject is [2nd] does the derivation converge, which explains the (un)grammaticality of (12b). Also in (15a), after T-agreement, C bearing [1st] agrees with the subject. However, since only the [2nd] or [3rd] subject can survive after agreement with T, C's agreement for [1st] always fails. As a result, any subject in (13) is ungrammatical.

In this section, we examined if phi-features stay at C or are inherited from C to T by focusing on discourse modals and the verb *kure(ru)* in Japanese. These cases both impose specific person restrictions on the subject. In the case of discourse modals, C/Modal requires a specific person for the subject. In the case of the verb *kure(ru)*, on the other hand, person restriction is one of the properties the verb *kure(ru)* specifically bears, so that V undergoes head movement to T and the resulting V-v-T amalgam imposes specific persons on the subject. We combined these two elements in a single sentence and demonstrated that phi-features on C/Modal are never inherited by T in these cases.

4 Consequences and Conclusion

The proposed system has several theoretical consequences. First, if the proposed system is on the right track, it lends further empirical support to Miyagawa's (2017) view that phi-features stay at C for agreement in Japanese without being inherited by T, in contrast to languages like English. Furthermore, under our analysis, the verb *kure(ru)* undergoes head movement to T forming an amalgam, which enables T to bear V's properties (i.e. person restriction). Also, the subject can never be included in the search domain of V for agreement if no head movement takes place. These points imply that head movement is a syntactic movement, in contrast to Boeckx and Stjepanović (2001), where head movement is phonological movement. Finally, it was demonstrated that the amalgamated heads (and inherited features from T to C if inheritance happens) serve as a single probe by keeping the person restriction each of the heads originally displays (not by prioritizing one of them). This is why phi-features on C need to be separated from those on T, as in (14a)/(15a). Hiraiwa (2001) also shows that the amalgam C-T-V serves as a single probe/Case-assigner, maintaining the heads' original properties. C and T originally assign genitive Case and nominative Case, respectively. As a result of amalgamation, C-T-V can assign both Cases by keeping their original Case-assignment abilities. Therefore, the proposed analysis is compatible with Hiraiwa's (2001) view on how amalgams work for agreement in the syntactic derivation.

In this work, we presented additional evidence for phi-agreement by C in Japanese, focusing on the person restriction observed in discourse modals and the verb *kure(ru)*, although it is still unexplained why only the verb *kure(ru)*, and not other verbs, imposes person restriction on the subject. Also, we clarified several theoretical consequences obtained from the proposed analysis.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP17K02823. We are very grateful to Samuel Epstein for his insightful comments and suggestions.

References

Boeckx, C. and S. Stjepanović (2001) Head-ing toward PF. *Linguistic Inquiry* 32:2, 345-355.

Chomsky, N. (2000) Minimalist Inquiries: The Framework. In *Step by Step: Essays on Minimalist Syntax in Honor of Howard Lasnik*, ed. by R. Martin, D. Michaels, and J. Uriagereka, 89-155. Cambridge, MA.: MIT Press.

Chomsky, N. (2007) Approaching UG from below. In *Interfaces+Recursion=Language?*, ed. by U. Sauerland and H. Gärtner, 1-29. Mouton de Gruyter.

Haegeman, L. and V. Hill (2011) The Syntacticization of Discourse, Ms., Ghent University and University of New Brunswick-SJ.

Hiraiwa, K. (2001) On Nominative-Genitive Conversion. In *MIT Working Papers 39: A Few From Building E-39*, ed. by E. Guerzoni, O. Matushansky and W. O'Neil, 66-124. Cambridge, MA: MITWPL.

Inoue, K. (2006) Ninongo no Zyookensetu to Syubun no Modaritii. *Scientific Approaches to Language* 5: 9-28. Center for Language Sciences, Kanda University of International Studies.

Kuno, S. and E. Kaburaki (1977) Empathy and Syntax. *Linguistic Inquiry* 8: 627－72.

Miyagawa, S. (2017) *Agreement beyond Phi*, MIT Press.

Obata, M. and M. Sugimura (2014) Phi-Agreement in Japanese: On the Person Restriction of Case-valuation. *Proceedings of the 40th Western Conference on Linguistics, Vol. 22*, 111-119.

Ouali, H. (2006) Unifying Agreement Relations: A Minimalist Analysis of Berber, Ph.D. Thesis, University of Michigan, Ann Arbor.

Oyharçabal, B. (1993) Verb Agreement with Non Arguments: On Allocutive Agreement. In *Generative Studies in Basque Linguistics*, ed. by J. I. Hualde and J. Ortiz de Urbina, 89-114. Amsterdam: John Benjamins.

Rizzi, L. (1997) The Fine Structure of the Left Periphery. In *Elements of Grammar: A Handbook of Generative Syntax*, ed. by L. Haegeman, 281–337. Dordrecht: Kluwer.

Speas, M. and C. Tenny (2003) Configurational Properties of Point of View Roles. In *Asymmetry in Grammar*, ed. by M. Di Sciullo, 315-444. Amsteram: John Benjamins.

Ueda, Y. (2008) Person Restriction and Syntactic Structure of Japanese Modals. *Scientific Approaches to Language* 7: 123-150. Center for Language Sciences, Kanda University of International Studies.

Towards the Non-predicate Modification Analysis of the Expressive Small Clause in Japanese

Kenji Oda
Syracuse University / Syracuse, NY, USA
Institute for the Study of Language and Information, Waseda University / Tokyo, Japan
koda100@syr.edu

Abstract

This paper concerns an understudied aspect of the grammar of Japanese termed "expressive small clause." In particular, it shows that the analysis of the expressive small clause in Japanese using the composition rule of Predication Modification is not empirically tenable, and the construction should instead be analyzed as an instance of applying a nominal argument to an expressive predicate.

1 Introduction

This paper concerns the structure of what is termed as *expressive small clause* in Japanese (Izumi and Hayashi, 2018) illustrated in (1).

(1) Kenji-no kusottare!
Kenji-GEN shit.dripper
'Kenji, asshole!'

In particular, this paper shows that the analysis based on the semantic operation of Predicate Modification that Izumi and Hayashi (2018) (I&H hereafter) propose fails to capture the descriptive generalizations of the construction, and it instead argues for an analysis based on more "traditional" predication.

The organization of this paper is as follows. Section 2 provides some introductory descriptive accounts of the expressive small clause. Section 3 reviews what I call the Predicate Modification analysis that I&H propose, which I argue is not empirically tenable in section 4. Section 5 provides an alternative account that is quite similar to what Potts and Roeper (2006) argue for English. Section 6 concludes.

2 Expressive Small Clause in Japanese

Let us first discuss some fundamental properties of the expressive small clause construction in Japanese. Potts and Roeper (2006) use the term *expressive small clause* (ESC) for expressions like (2).[1]

(2) You idiot! (Potts and Roeper, 2006)

Example (2) is *expressive* in that it is used to express the speaker's attitude/emotion towards the addressee or towards a situation. It is a *small clause* in the sense that it is verbless. I&H identify that expressions like (1) are the Japanese equivalent of (2). Thus, (1) is *expressive* in that it is used to express the speaker's emotion or attitude towards the referent (*Kenji*) or his act, rather than to describe or assert the referent's idiocy, and it lacks a verb or a copula. In Japanese, unlike in English, the ESC is mainly used to express the speaker's *negative* attitude towards the addressee or his act. Thus, positive descriptions such as *tensai* 'genius' and *hansamu* 'handsome' are far less acceptable, unless, perhaps, they are used clearly as a sarcasm.[2]

(3) # Kenji-no tensai/hansamu!
Kenji-GEN genius/handsome!
'Kenji, genius!'

Japanese differs from English in two other regards. First, the particle *-no* links the two constituents, and it is obligatory. I call this particle

[1] Gutzmann (2019) uses the term *expressive vocative* to refer to expressions like (2).

[2] Some Japanese speakers may find (3) rather acceptable, as a reviewer pointed out, while the native speakers that I consulted share the judgment given in this paper.

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 196-202, Hakodate, Japan, September 13-15, 2019

genitive marker for the sake of convenience, merely reflecting the fact that it typically marks the genitive case. A typical ESC in Japanese has the form of NP-*no* XP. The NP appears to be the argumental referential expression, whereas the XP is seemingly the predicate which typically carries some derogatory sense.[3] Second, the Japanese ESC allows non-second person argument, and in fact, use of a second person pronominal form is disallowed:

(4) ?? omae-no kusottare!
you-GEN shit.dripper
'You asshole!'

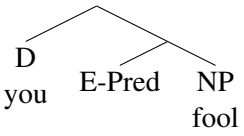
If a speaker wishes to use an ESC to a hearer, s/he can use the hearer's name. Thus, example (1) can be used to express the speaker attitude to the hearer whose name is *Kenji*. Finally, the Japanese expressive small clause can be used as an argument as in (5) below.

(5) Kenji-no kusottare-ga kita.
Kenji-GEN shit.dripper-NOM came
'The asshole Kenji came.'

3 The Predicate Modification Analysis of Japanese Expressive Small Clause

Let us now consider how the Japanese ESC has been analyzed. Taking Potts and Roeper's (2006) analysis as a starting point, I&H propose that the English ESC is mediated by E-Pred as shown in (6) and (7).

(6)



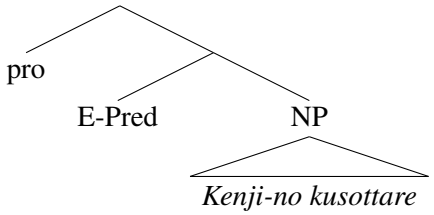
(7) $[\![\text{E-Pred}]\!]^c = \lambda P_{\langle e,t\rangle}.\lambda x.\{c' : c'_S = c_S$ and $c'_A = c_A$ and x is c'_A and c'_S considers c'_A to be unfavourably describable as being $P\}$
(where c_S is the speaker of context c and c_A is the addressee of c)
(Izumi and Hayashi, 2018: (20a), (21))

[3] It has been argued recently that Japanese is an NP language that lacks the D layer in its syntax. I use the term NP for expository purposes only, and I remain agnostic regarding this issue. See Izumi (2016) and references there in for further details. Also, I use the term XP for the "predicate" component of the Japanese ESC to abstract away from the exact nature of this entity, although nouns often appear in this position.

In short, E-Pred is an abstract predicate which turns an ordinary predicate into an expressive one. The composition of the structure in (6) results with an expressive meaning of type E.

I&H claim that Japanese differs from English in that E-Pred appear above the two constituents in the expressive expression:

(8)



(Izumi and Hayashi, 2018: (20b))

This follows from the idea that proper names are underlyingly predicates of type $\langle e,t\rangle$ (Izumi, 2016; Oda, 2018, and references therein). Given that both the proper name *Kenji* and the predicate *kusottare* 'shit dripper' are of type $\langle e,t\rangle$, they are processed by Predicate Modification:

(9) Predicate Modification
If α is a branching node, $\{\beta,\gamma\}$ is the set of α's daughters, and $[\![\beta]\!]$ and $[\![\gamma]\!]$ are both in $D_{\langle e,t\rangle}$, then
$[\![\alpha]\!] = \lambda x \in D_e.[\![\beta]\!](x) = [\![\gamma]\!](x) = 1$
(Heim and Kratzer, 1998: 65)

Predicate Modification yields a single type $\langle e,t\rangle$ predicate out of two. Thus, assuming that the name *Kenji* roughly means a set of entities that are called "Kenji", and leaving aside the exact nature of the particle *-no*, the NP *Kenji-no kusottare* in (8) roughly means the following:

(10) $\lambda x.x$ is called *Kenji* and x is an asshole.
(cf. Izumi and Hayashi, 2018: (24b))

After applying (10) and the abstract null pronoun to E-Pred, we obtain the following meaning.

(11) $\{c'$: the individual referred to by pro is $c'_A(= c_A)$ and $c'_S(= c_S)$ considers c'_A to be unfavourably describable as being called Kenji and as being an asshole$\}$
(cf. Izumi and Hayashi, 2018: (24c))

I&H claim that this analysis captures some of the observations that we made in the previous section. First, since the ESC always target the addressee in

the context, so that the use of second person pronoun in the ESC would be redundant, which accounts for the ill-formedness of example (4). In addition, I&H account for the argumental use of the ESC illustrated in (5) by claiming that it is in fact not an ESC, but it is the NP component of the structure in (8) composed by Predicate Modification, which receives a referential meaning by a type-shifting rule.

4 Against the Predicate Modification Analysis

While there are some clear advantages to I&H's Predicate Modification-based analysis laid out above, I argue in this section that it is not tenable as it fails to account for many empirical facts.

First, let us consider the structure of the Japanese ESC and the meaning of E-Pred in (7) and (8). Since E-Pred turns an ordinary predicate to an expressive one, and it takes scope over both the argumental NP and the predicative XP in the structure in (8), E-Pred crucially takes scope over the NP and it necessarily anticipates that the speaker has an unfavourable attitude towards being called *Kenji* by uttering (1). This prediction is not borne out since the name *Kenji* in (1) carries no derogatory sense. I&H acknowledge this observation and add that the NP can have a negative referential expression such as *teisupe pasokon* 'low-spec PC' in (12).

(12) teisupe pasokon-no baka!
low.spec PC-GEN idiot
'Low spec PC, idiot!'
(cf. Izumi and Hayashi, 2018: (9d))

It should be pointed out, however, that the way in which the E-Pred is construed necessarily makes the entire ESC expressive, and therefore the predicative content of the argumental NP component would also necessarily be interpreted derogatory.

Now let us consider some of the consequences of deriving the structure of the Japanese ESC using Predicate Modification. The composition rule of Predicate Modification is typically used to account for restrictive modification, and thus the elements that enter into the semantic composition via Predicate Modification, such as *red* in *red shirts* are taken to be adjuncts. This means that at least type-theoretically there is no reason for the Japanese ESC to have both constituents of the frame [NP-*no* XP] to be present, if the Predicate Modification-based analysis is on the right track. Example (13a), which lacks the argumental NP, is felicitous only when the speaker is facing the addressee and expressing the addressee's idiocy. This contrasts with (1), which is possible even when the individual whose name is *Kenji* is not present in the scene where it is uttered. Similarly, (13b) may be used as a vocative expression, but it does not by itself express any negative impression towards the addressee or the name.

(13) a. Baka! 'Idiot!'
b. Kenji!

In addition, the two constituents are both of type $\langle e, t \rangle$. This means that they are semantically very similar and they should be switched around when everything else is equal. Thus we expect that the following example to be able to function as an ESC, just like (1) does.

(14) kusottare-no Kenji
shit.dripper-GEN Kenji
'Kenji, who is an asshole'

While it is possible to interpret this example as a phrase where the expressive expression *kusottare* 'shit dripper' modifies the proper name *Kenji*, it does not have the expressive function that (1) invokes.[4]

Finally, let us consider the fact that the ESC can be used as a nominal argument, illustrated in (5) which is repeated as (15) below.

(15) Kenji-no kusottare-ga kita.
Kenji-GEN shit.dripper-NOM came
'The asshole Kenji came.'

[4]It should be noted that Japanese allows modification of proper names.

i. yakyuusenshu-no Ichiro
baseball.player-GEN Ichiro
'Ichiro who is a baseball player'

It should be reminded that the form NP-*no* XP in Japanese is highly ambiguous, as an reviewer points out that there are non-expressive cases like (ii) below where the second entity appears to be predicated of the first entity.

ii. bara-no hana
rose-GEN flower
'rose/rose flower'

See footnote 5 for another construction that has the form NP-*no* XP.

I&H argue that the semantics of the nominal argument *Kenji-no kusottare* 'the asshole Kenji' should be derived by Predicate Modification just like an ordinary nominal expression of type $\langle e, t\rangle$ with a modifier. This then would suggest that the derogatory expression *kusottare* 'shit dripper' can be substituted with any other type $\langle e, t\rangle$ (nominal) expression. However, it is not the case. Consider (16) below.

(16) * Kenji-no {kyooju / sensei / tensai}-ga kita.
Kenji-GEN professor / teacher / genius-NOM came
'Kenji, who is a professor/teacher/genius, came.'

In (16), all of the nominal predicates following the proper name does not necessarily convey any negative interpretation that *kusottare* 'shit dripper' in (15) does, and the sentence is ungrammatical with any of those words with the intended meaning.[5] In fact, it is possible to interpret (16) with the intended appositive-like meaning, once it is made clear that the predicate is sarcastically used.

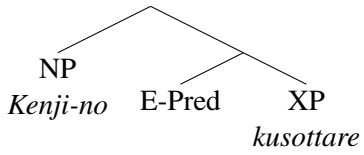
(17) Kenji-no sensei-**sama**-ga mata nanika itteru.
Kenji-GEN teacher-HON-NOM again something say.PROG
'Kenji, the teacher-preacher, is saying something again.'

The honorific suffix *-sama* in (17) makes it clear that the phrase is used sarcastically and the sentence is perfectly salient with the intended meaning in a situation where the speaker is annoyed by Kenji's mansplaining behaviour. These examples show clearly that the predicate XP entity in the string [NP-*no* XP] must have a clear negative expressive meaning, and thus the use of the string as an argument cannot be derived simply by applying the rule of Predicate Modification, and thus we need to treat this as one of the characteristics of the Japanese ESC.

5 Towards the "Small Clause" analysis of the Japanese Expressive Small Clause

Having seen that the Predicate Modification analysis of the Japanese ESC, which makes use of Predicate Modification and places E-Pred above the argument NP in the frame of [NP-*no* XP], encounter numerous issues, it is fair to conclude that the analysis is not tenable, at least in its current shape. In this section, I would like to argue that the Japanese ESC should be very similar to its English counterpart like *You idiot!*. More specifically, I argue that the Japanese ESC should have a structure very similar to (6) shown in (18) below, abstracting away from the exact nature of the genitive marker *-no*.

(18)



What is most crucial in (18) is that what I have been calling the "argumental" NP is, in fact, referential, and thus it is an argument, while the predicate XP seems to carry the predicative characteristics. This simply means that the semantic composition of the phrase should proceed as a rather trivial instance of Functional Application, which applies the argument NP to the predicate XP.

5.1 Coordination in Japanese

In order to see that the NP in the Japanese ESC is, in fact, referential while the XP component is a predicate, we first need to review how coordination works in Japanese.

Japanese has numerous coordinators, and most notably they are sensitive to the syntactic/semantic categories of the conjuncts. Our main concern in this section is the coordination particles *-to* and *-de*. The particle *-to* typically appears when two referential nominal expressions are coordinated, while *-de* is used when two predicative nominal expressions

[5] Note that (16) is grammatical with the possessive interpretation. It should also be pointed out Oda (2018) reports that sentences analogous to (16) is acceptable with the intended meaning when the predicative nominal entity is one of the kinship terms.

i. Tanaka-no ojisan-ga kita.
Tanaka-GEN uncle-NOM came
'Tanaka, who is a middle aged man, came.'

As Oda (2018) argues, what is observed in the example above is underlyingly different from the ESC that we are concerned with in this paper.

are coordinated (Nishiyama, 2003: 126–128).[6] The contrast is illustrated by the examples in (19) and (20). Let us consider (19) first.

(19) a. [seijika-**to** pianisuto-wa]
politician-CONJ pianist-TOP
kanemochi-da.
rich-cop
'Politicians and pianists are rich.'

b. # ano hito-wa [Fukuda
that person-TOP Fukuda
Takehiko-**to** Kada Reitarou
Takehiko-CONJ Kada Reitarou
]-da
-COP
'That person is Takehiko Fukuda and Reitarou Kada.'

c. # [furansu bungakusha-**to**
France literature.scholar-CONJ
pianisto]-no seijika
pianist-GEN politician
'a politician who is a scholar of French literature and (also) a pianist.'
(Nishiyama, 2003: 126–128)

The subject phrase with the particle *-to* in (19a) is well-formed since it is salient to interpret the conjuncts *seijika* 'politician(s)' and *pianisuto* 'pianist(s)' are distinct (sets of) referential entities. On the other hand, (19b) is infelicitous since the coordinator *-to* forces the interpretation where there are two separate individuals, one called *Takehiko Fukuda* and the other *Reitarou Kada* which is inconsistent with the singular subject *ano hito* 'that person'. Similarly (19c) is infelicitous since the conjuncts are the predicates of the relative clause modifying the noun *seijika* 'politician', which cannot be coordinated by the particle *-to*. Once the particle *-to* in (19) is changed to *-de*, we see the completely opposite results:

(20) a. ? [seijika-**de** pianisuto-wa]
politician-CONJ pianist-TOP
kanemochi-da.
rich-COP
'Politicians and pianists are rich.'

b. ano hito-wa [Fukuda
that person-TOP Fukuda
Takehiko-**de** Kada Reitarou
Takehiko-CONJ Kada Reitarou
]-da
-COP
'That person is Takehiko Fukuda and Reitarou Kada.'

c. [furansu bungakusha-**de**
France literature.scholar-CONJ
pianisto]-no seijika
pianist -GEN politician
'a politician who is (also) a scholar of French literature and a pianist.'

Example (20a) which contrasts with (19a) is not quite acceptable, and even if a speaker finds it acceptable, it rather means that someone who is both a politician and a pianist is rich. In contrast, examples (20b, c) are quite well-formed. It should be noted that the coordinated structure in (20b) is interpreted as having two distinct names, and thus they are acting predicatively just like the coordinated phrase in (20c).

With this information in mind, let us now return to the ESC.

5.2 Coordination and Expressive Small Clause

The structure in (18) that I argue for maintains the referent–predicate asymmetry of the Japanese ESC. Thus, we should expect that the particle *-to* that coordinates referential items may appear in the argumental NP position whereas the particle *-de* that coordinates predicates may be found in the predicate XP position. This prediction is borne out. Let us first consider the argumental NP position.

(21) a. [Kenji-**to** Toshie]-no baka!
Kenji-CONJ Toshie -GEN idiot!
'Kenji and Toshie, idiot!'

b. # [Kenji-**de** Toshie]-no baka!
Kenji-CONJ Toshie -GEN idiot!
'Kenji and Toshie, idiot!'

Sentence (21a), with the referent-coordinating *-to*, is highly acceptable with the reading where there are two individuals, Kenji and Toshie, and the speaker is frustrated with the idiocy of them or their behaviour. In contrast, (21b) that has the predicate-coordinating

[6] A similar observation has also been made by Tsujimura (1996: 126–127).

-de is not acceptable at all. It certainly fails to yield the interpretation that we find with (21a). It may at best mean that the speaker is frustrated with one individual whose name is *Kenji*, who happens to be also called *Toshie*. I suggest for now that (21b) sounds strongly infelicitous because the expressive content is marred by providing rather ancillary information that s/he is also called *Kenji*. The availability of the particle *-to* within the NP position confirms that the NP constituent is referential and thus it functions as an argument. In conclusion, the data support the claim that the ESC should have the structure in (18). On the other hand, the Predicate Modification analysis fails to provide a clear explanation we observe the pattern in (21).

Let us now turn to the predicate XP position. The elements in this position are typically coordinated with *-de*.

(22) a. # Kenji-no [baka-**to**
Kenji-GEN idiot-CONJ
wakarazuya]!
bigot
'Kenji, idiot and bigot!'

b. Kenji-no [baka-**de**
Kenji-GEN idiot-CONJ
wakarazuya]!
bigot
'Kenji, idiot and bigot!'

While (22b) is completely salient for expressing the speaker's wish to inveigh against Kenji by calling out two negative properties, (22a) is not acceptable at all. I take this as a clear confirmation that the XP component of the Japanese ESC is in fact a predicate.

Putting together, it is safe to conclude that the Japanese ESC has the "small clause" structure in 18 that is analogous to its English counterpart.

5.3 Apparent Counterexamples

Some of the empirical findings that we discussed earlier in this paper are not apparently consistent with the structure in (18). First, the "small clause" analysis fails to account for the unavailability of the second person argument exemplified in (4) (repeated as (23) below).

(23) ?? omae-no kusottare!
you-GEN shit.dripper
'You asshole!'

I agree with I&H's intuition that this particular sentence is not acceptable, but it is dubious whether it has to do with having the second person argument. Japanese is known for having an array of pronominal expressions with slightly different expressive meanings. For example, *anata*, *anata-sama*, *anta*, *omae*, *kisama*, *temee*, *omae-san*, *kiden*, etc. all mean 'you', but each of these expresses a slightly different attitude towards the hearer. Crucially, it seems that some of these pronominal expressions fare better with the ESC than others. For example, the example below appears to be well-formed, although it does certainly sound rather archaic.

(24) **omaesan**-no {baka/ kusottare/ hentai/
you-GEN idiot shit.dripper pervert
ecchi}!
dirty
'You idiot/asshole/pervert/dirty!'

While the exact nature of this restriction is very unclear and it is beyond the scope of this paper, this data suggests clearly that the restriction is not simply about being the second person.

Another apparent counterexample to the small clause analysis of the Japanese ESC is that it can function as an argument of the larger structure as in (25) repeated from (15).

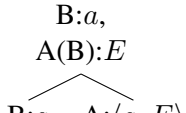
(25) Kenji-no kusottare-ga kita.
Kenji-GEN shit.dripper-NOM came
'The asshole Kenji came.'

This contrasts to the English ESC which cannot behave as a nominal argument.

(26) * You fool should read more carefully.
(Potts and Roeper, 2006: (44b))

Based on this observation, Potts and Roeper (2006) claim that the English ESC is composed via Functional Application, and rejects the two-dimensional application of Functional Application proposed by Potts (2005) shown in (27).

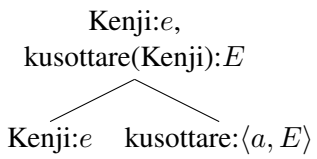
(27) Expressive Composition



(Potts and Roeper, 2006: (42))

Given the fact that the ESC in Japanese can be used nominally, we may simply propose that availability of this expressive composition rule is rather language specific, and thus the ESC in (25) should look like the following:

(28)



Thus, in one dimension, the expression expresses the speaker's attitude towards Kenji, while it functions as a part of larger structure in another dimension.[7]

6 Conclusion

This paper investigated the expressive small clause construction, a rather understudied domain in the syntax/semantics of Japanese. While it has been previously argued by Izumi and Hayashi (2018) that the construction is derived by the composition rule of Predicate Modification, this paper has shown that such an analysis is not empirically tenable and that it should be analyzed as an instance of simple application of an argument to an expressive predicate, just like its English counterpart.

Acknowledgments

I would like to thank the audience at the Syntax Project at the University of Toronto and the three anonymous reviewers for their insightful comments. All errors are mine.

References

Daniel Gutzmann. 2019. *The Grammar of Expressivity*. Oxford University Press, Oxford, UK.

Irene Heim and Angelika Kratzer. 1998. *Semantics in Generative Grammar*. Blackwell, Malden, MA.

Yu Izumi. 2016. *Namae-to Taishoo: Koyuumei-to Rameishi-no Imiron*. [Name and Referent: Semantics of Proper Names and Bare Nouns]. Keiso Shobo, Tokyo.

Yu Izumi and Shintaro Hayashi. 2018. Expressive Small Clauses in Japanese. In: Arai S., Kojima K., Mineshima K., Bekki D., Satoh K., Ohta Y. (eds) *New Frontiers in Artificial Intelligence.* JSAI-isAI 2017. Lecture Notes in Computer Science, vol 10838. 188–199. Springer, Cham.

Yuji Nishiyama. 2003. *Nihongo Meishiku-no Imiron-to Goyooron*. [Semantics and Pragmatics of Japanese Noun Phrase]. Hituji Shobo, Tokyo.

Kenji Oda. 2018. Syntax of proper names in Japanese. In: Ermolaeva, M., Hakgüder, E., Lai, J., Montemurro, K., Rhodes, B., Sankhagowit, A., Tabatowski, M. (eds) *Proceedings of the 53rd Annual Meeting of the Chicago Linguistic Society*. 261–274. The Chicago Linguistic Society, Chicago.

Christopher Potts. 2005. *The Logic of Conventional Implicatures*. Oxford University Press, Oxford, UK.

Christopher Potts and Tom Roeper. 2006. The Narrowing Acquisition Path: From Expressive Small Clauses to Declaratives. In: Progovac, L., Paesani, K., Casielles-Suárez, E., Barton, E. (eds) *The Syntax of Nonsententials: Multi-Disciplinary Perspectives*. 183–201. John Benjamins, Amsterdam.

Natsuko Tsujimura. 1996. *An Introduction to Japanese Linguistics*. Blackwell, Malden.

[7] A reviewer points out that (28) is similar to English *damn* and its Japanese counterpart *kuso* and provides the following examples.

i. The damn dog ate my homework!

ii. kuso inu-ga shukudai-o tabeta!
shit dog-NOM homework-ACC ate

On the other hand, the reviewer adds that the following examples are not acceptable.

iii. ?? inu-no baka!
dog-GEN idiot

iv. ?? inu-no baka-ga kyoo-mo sosoo-o
dog-GEN idiot-GEN today-too toilet.accident-ACC
sita
did
'(The) idiot dog had a toilet accident again today.

I suggest that this interesting contrast is observed due to the fact that the noun *inu* 'dog' is either treated indefinite or non-referential in (ii)–(iv). Thus, adding a demonstrative before the noun in (iii) and (iv) improves the acceptability:

v. ano inu-no baka-ga kyoo-mo
that dog-GEN idiot-GEN today-too
sosoo-o sita.
toilet.accident-ACC did
'(The) idiot dog had a toilet accident again today.

Syntax and Semantics of Numeral Classifiers in Japanese

Atsushi Oho
International Christian University
3-10-2 Osawa, Mitaka-shi, Tokyo 181-8585, Japan
atsushioho@gmail.com

Abstract

This paper explores the syntax and semantics of Japanese numeral classifiers in the prenominal and postnominal positions. argue that there are two transformationally unrelated structures for numeral classifiers in Japanese: one in which a numeral and a classifier form a constituent and modify a noun as an adjunct and one in which numerals and classifiers are distinct functional heads of the extended nominal projection. Evidence comes from novel data about the optionality of classifiers. It is shown that classifiers can be omitted for certain numerals in the prenominal construction but not in the postnominal one. Some theoretical consequence and implication of the analysis are also discussed.

1 Introduction

This paper discusses the syntax and semantics of numeral classifiers in Japanese by examining the word order variation. Japanese allows numeral classifiers to appear prenominally and postnominally as shown in (1a) and (1b) respectively.[1]

(1) a. *Prenominal*
John-wa **san-ko**-no ringo-o tabeta.
John-TOP 3-CL-GEN apple-ACC ate
'John ate three apples.'

b. *Postnominal*
John-wa ringo **san-ko**-o tabeta.
John-TOP apple 3-CL-ACC ate

There has been a debate in the literature whether the two constructions are transformationally related. For example, Watanabe (2006) argues that the two constructions are derived from one underlying structure, namely, they are transformationally related. By contrast, Huang and Ochi (2014) claim that they are not related by any transformational rules.

This paper aims to contribute to this debate by providing novel data about the optionality of classifiers. It shows that there is an asymmetry of the optionality of classifiers between the two constructions. It is argued that the asymmetry is due to the difference in the syntax and semantics between the two constructions. Specifically, I propose that for the prenominal construction, a numeral and a classifier form a constituent and occupy an NP adjunct position, whereas for the postnominal construction, a numeral and a classifier are functional heads of the extended nominal projection. The current paper also contributes to the debate as to why classifiers are required in this language. In Krifka (1995), it is because of numerals, whereas in Chierchia (1998a), it is because of nouns. I will discuss some implications from the analysis for this debate.

The paper is organized as follows: Section 2 presents core data showing the asymmetry of the optionality of classifiers. Section 3 makes an analysis based on the semantics of Rothstein (2013) and Sudo (2016). Section 4 discusses some implications of the analysis. Section 5 concludes the paper.

[1] Quantifier float is also possible in Japanese as in (i).

(i) John-wa ringo-o **san-ko** tabeta.
John-TOP apple-ACC 3-CL ate
'John ate three apples.'

This paper will not discuss the floating quantifier construction, though a brief comment is made in footnote 4 and 15.

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 203-211, Hakodate, Japan, September 13-15, 2019

2 Asymmetry of optionality of classifiers

Japanese is an obligatory classifier languages and classifiers are needed when numerals modify nouns.

(2) a. san*-(satsu)-no hon
3-(CL)-GEN book
'three books'

b. hon san-*(satsu)
book 3-(CL)

However, under some circumstance, classifiers can be optional. Sudo (to appear) observes that classifiers tend to be optional with numerals expressing large numbers:

(3) Daitooryoo-wa shichoosha-kara yoserareta
president-TOP viewer-from were.sent
hyaku-(ko)-no shitsumon-ni kaitooshita.
100-(CL)-GEN question-to answered
'The president answered 100 questions viewers asked.' (Sudo, to appear: 4)

Nomoto (2013) observes that relatively small numbers allow classifiers to be omitted: 9 is marginally acceptable and 15 is well-formed without classifiers.

(4) John-wa { kyuu-?(ko) / juu-go-(ko) }-no
John-TOP 9-(CL) / 10-5-(CL) -GEN
gengo-o shirabeta.
language-ACC investigated
'John investigated { nine / fifteen } languages.'
(Based on Nomoto, 2013: 16)

In addition to the case of large numbers, classifiers can be optional for non-specific numbers.[2]

(5) a. John-wa juu-suu-(ko)-no shima-o
John-TOP 10-some-(CL)-GEN island-ACC
otozureta.
visited
'John visited a dozen islands.'
(Based on Nomoto, 2013: 16)

b. John-wa suu-juu-(ko)-no shima-o
John-TOP some-10-(CL)-GEN island-ACC
otozureta.
visited
'John visited dozens of islands.'

[2] In (5b) and (7b), when the classifier appears, the final sound of *juu* 'ten' assimilates to the first consonant of the classifier, resulting in *suu-juk-ko*.

Admittedly, it is not totally clear exactly when classifiers are optional. However, the observations suggest that the optionality depends on numerals.

So far, we have seen the examples in the prenominal construction. A novel observation, however, shows that the optionality does not hold in the postnominal construction. Consider the following examples, all of which are the same as (3–5) except the position of the numeral classifiers relative to the head nouns.

(6) *Large numbers*

a. Daitooryoo-wa shichoosha-kara
president-TOP viewer-from
yoserareta shitsumon hyaku-*(ko)-ni
were.sent question 100-(CL)-to
kaitooshita.
answered
'The president answered 100 questions viewers asked.'

b. John-wa gengo juu-go-*(ko)-o
John-TOP language 10-5-(CL)-ACC
shirabeta.
investigated
'John investigated fifteen languages.'

(7) *Non-specific numbers*

a. John-wa shima juu-suu-*(ko)-o
John-TOP island 10-some-(CL)-ACC
otozureta.
visited
'John visited a dozen islands.'

b. John-wa shima suu-juu-*(ko)-o
John-TOP island some-10-(CL)-ACC
otozureta.
visited
'John visited dozens of islands.'

In the examples in (6), which contain the large numbers, the classifiers cannot be omitted.[3] Similarly, the examples in (7) containing the non-specific numbers are considerably degraded without the classifiers.

As we have seen, on the one hand, the prenominal construction shows the optionality of classifiers for

[3] Yasutada Sudo (p.c.) pointed out to me that classifiers can be omitted in the postnominal construction as in (i).

the particular types of numerals. The postnominal construction, on the other hand, does not admit the optionality and classifiers are always required. The contrast between the two constructions shows that the optionality also depends on the construction.[4]

3 Analysis

I propose that the asymmetry of the optionality and obligatoriness of classifiers in Japanese is due to the syntactic and semantic difference between the prenominal and postnominal constructions. Specifically, I suggest that in the prenominal construction, a numeral and a classifier form a constituent but in the postnominal construction, they are distinct heads of the extended nominal projection. In the following, I first analyze the prenominal construction based on Rothstein (2013) and Sudo (2016) and then move on to the postnominal construction.

3.1 Prenominal constructions

Rothstein (2013) proposes that numerals are analyzed as properties. In property theory as in Chierchia (1985), properties have multiple functions which are related via type-shifting operations. In the case of numerals, they are predicated of arguments and in this case, numerals are of type ⟨e,t⟩ just like adjectives as in (8a) with the cardinality function defined in (8b), where x ranges over plural individuals.

(8) a. $[\![\text{three}_{\langle e,t\rangle}]\!] = \lambda x.|x| = 3$

b. $|x| = n \leftrightarrow |\{y : y \sqsubseteq_{\text{ATOMIC}} x\}| = n$

Predicates have the corresponding individual property correlate of the set in (8a). Thus, numerals are also of type n, a type of numbers. This is derived by the $^{\cap}$ operation.

(9) $[\![\text{three}_{n}]\!] = 3 = {}^{\cap}(\lambda x.|x| = 3)$

On the other hand, the $^{\cup}$ operator can apply to type-n objects, deriving the corresponding predicates of type ⟨e,t⟩.

(10) ${}^{\cup}3 = {}^{\cup\cap}(\lambda x.|x| = 3) = \lambda x.|x| = 3$

Having said that, let us turn to the Japanese data. In Sudo (2016), denotations of nominals in Japanese are equivalent to English count nouns, except the number specification. They contain both singular and plural individuals. Plural individuals are sums of singular individuals (Link, 1983; Sauerland, 2005). Thus, the noun *gakusei* 'student' is true of both singular and plural entities consisting of students as indicated by the *-operator.[5]

(11) $[\![\text{gakusei}]\!] = [\![\text{students}]\!] = \lambda x.{}^{*}\text{STUDENT}(x)$

Sudo assumes that the default type of numerals is of type n.

(12) $[\![\text{san}_{n}]\!] = 3$

Numerals cannot directly modify nouns since they are type-n objects. Sudo proposes that the role of classifiers is to turn the type-n object into a modifier

(i) John-wa hohei sen-(nin)-o hiki-tsureta.
John-TOP foot.soldier 1000-(CL)-ACC took
'John took 1000 foot soldiers.'

The judgments are delicate and seem to vary among speakers, indicating that several factors seem to be involved to make classifiers optional. I should leave for future research whether there are rules governing the optionality or examples such as (i) are exceptions.

It should be noted, however, that when we make postnominal numerals vague-quantity by addition some element, classifiers would tend to be optional. Tomoyuki Yoshida (p.c.) notes that when *ijoo* 'greater than or equal to' is attached, a classifier may be omitted.

(ii) Daitooryoo-wa shichoosha-kara yoserareta shitsumon
president-TOP viewer-from were.sent question
100-(ko)-ijoo-ni kaitooshita.
hundred-(CL)-greater.than.or.equal.to-to answered
'The president answered greater than or equal to 100 questions viewers asked.'

An anonymous review provides the following example, in which an approximate expression *oyoso* 'about' is used.

(iii) Kanshuu oyoso ichi-man-(nin)-ga tsumekaketa.
spectator about 1-10000-CL-NOM crowded
'About 10000 spectators crowded.'

Though it is not straightforward to capture what factors are responsible for the optionality in the postnominal construction, what is clear at this point is that there is a contrast between the prenominal and postnominal construction with regard to the acceptability of numerals without classifiers as shown in (3–7).

[4] In the floating construction, when classifiers are omitted, the acceptability varies across speakers.

(i) a. John-wa gengo-o juu-go-?/??(ko) shirabeta.
John-TOP language-ACC 10-5-(CL) investigated
'John investigated fifteen languages'

An anonymous reviewer observes that when *juu-go* 'fifteen' is replaced with *juu* 'ten', the acceptability improves. The reviewer suggests that this may have to do with some sort of phonological weight. I would like to thank the reviewer for drawing my attention to the phonological factor.

[5] *P(x) is the closure of P(x) under i-sum formation $^{\sqcup}$.

of type ⟨e,t⟩. In addition, each classifier has a sortal restriction. For example, *-nin* is used for counting humans and humans only. This sortal restriction is assumed to be a presupposition.

(13) $[\![\text{-nin}]\!] = \lambda n.\lambda x : *\text{HUMAN}(x).|x| = n$

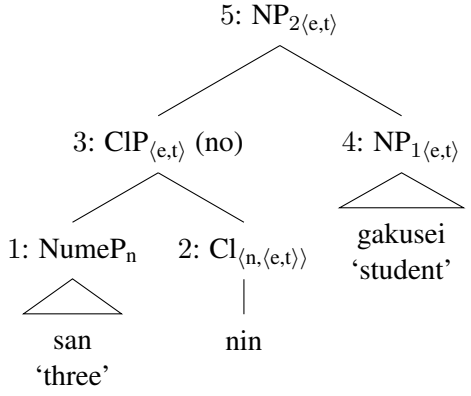
Due to the sortal presupposition, the classifier *-nin* ensures that x is a single human or an i-sum consisting of humans and counts the number of singular humans in x. A classifier and a numeral are combined via Functional Application, resulting in a function of type ⟨e,t⟩.

(14) $[\![\text{san-nin}]\!] = \lambda x : *\text{HUMAN}(x).|x| = 3$

The numeral classifier, then, combines with a noun via Predicate Modification or a type-shifting operation.

Sudo assumes that a numeral and a classifier form a constituent to the exclusion of the noun phrase. I adopt this analysis and propose the structure in (15) for the prenominal construction.[6]

(15)



In this structure, a numeral and a classifier combine first and form a constituent, ClP, which then modifies NP (the genetive marker *-no* is considered as having no semantic effects). The derivation of (15) is given in (16).

(16) a. $1 : [\![\text{NumeP}]\!] = 3$

b. $2 : [\![\text{Cl}]\!]$
$= \lambda n.\lambda x : *\text{HUMAN}(x).|x| = n$

c. $3 : [\![\text{ClP}]\!] = \lambda x : *\text{HUMAN}(x).|x| = 3$

d. $4 : [\![\text{NP}_1]\!] = \lambda x.*\text{STUDENT}(x)$

e. $5 : [\![\text{NP}_2]\!]$
$= \lambda x.|x| = 3 \;\&\; *\text{STUDENT}(x)$

[6] I assume that DP is located above the highest NP.

In non-classifier languages such as English, the $^{\cup}$ operator is applicable to type-n numerals as in (17a). The $^{\cup}$ shifted numerals can modify a noun directly just like adjectives as in (17b).

(17) a. $^{\cup}[\![\text{three}_n]\!] = [\![\text{three}_{\langle e,t\rangle}]\!] = \lambda x.|x| = 3$

b. $[\![\text{three students}]\!]$
$= \lambda x.|x| = 3 \;\&\; *\text{STUDENT}(x)$

Following Chierchia (1998a; 1998b), Sudo claims that the $^{\cup}$ operation is considered as a last resort option. When a language has overt lexical items whose function is equivalent to the $^{\cup}$ operator, the use of such lexical items is mandatory and consequently the application of the $^{\cup}$ operation is blocked. As we have seen, classifiers do the job of the $^{\cup}$ operator. Thus, in classifier languages, due to the existence of classifiers, the $^{\cup}$ operation is not applicable.

Regarding optionality of classifiers, Sudo acknowledges that his analysis cannot straightforwardly account for languages in which classifiers are optional. He notes that in optional classifier languages, the application of the $^{\cup}$ operator is not blocked, though it remains unanswered how this works.

To capture the optionality in Japanese, I suggest that the $^{\cup}$ operation is applicable in Japanese, contra Sudo (2016).[7] The application of $^{\cup}$ is, however, restricted to a subset of numerals. As seen, classifiers become optional for large numbers and non-specific numbers. As mentioned, it is not clear exact when classifiers are optional, but it is safe to say that the $^{\cup}$ operation is applicable to those numerals that express large numbers and non-specific numbers. To distinguish from the ordinary $^{\cup}$, I introduce $^{\Cup}$, a partial function version of $^{\cup}$, defined in (18).[8]

(18) Let n be a number in the domain of type n.
$^{\Cup}n$ is defined only if n expresses a "large" number or a "non-specific" number.
If defined, $^{\Cup}n = \lambda x.|x| = n$

Let us see a concrete example. The numeral *hyaku* 'hundred' can combine directly with a noun without a classifier (as in (3)). Thus, when it combines with a

[7] Recently, however, Yasutada Sudo (p.c.) has noted that the type-shifting with $^{\cup}$ should be available in Japanese.

[8] As pointed out by a reviewer, the treatment of the $^{\Cup}$ operation contains several issues. Particularly, what counts as "large" or "non-specific" numbers is vague. I have to leave this issue for future research.

noun *hon* 'book', two forms are possible: without a classifier (*hyaku-no hon*) and with a classifier *-satsu* (*hyaku-satsu-no hon*). Compare the derivations of the two forms. First, the numeral of type n forms a constituent with the classifier *-satsu* and modifies the noun *book* as shown in (19).

(19) a. $[\![\text{hyaku}_{\text{n}}]\!] = 100$
b. $[\![\text{-satsu}]\!] = \lambda n.\lambda x : *\text{BOOK}(x).|x| = n$
c. $[\![\text{hon}]\!] = \lambda x.*\text{BOOK}(x)$
d. $[\![\text{hyaku}_{\text{n}}\text{-satsu-no hon}]\!]$
$= \lambda x.|x| = 100 \ \& \ *\text{BOOK}(x)$

When the numeral modifies the noun without the classifier, the $^{\cup}$ operator applies to the numeral of type n and the corresponding predicate of type ⟨e,t⟩ is derived as in (20a), with the assumption that the $^{\cup}$ operation is defined for *hyaku*. The numeral can combine with the noun without the classifier as in (20b).

(20) a. $^{\cup}[\![\text{hyaku}_{\text{n}}]\!] = [\![\text{hyaku}_{\langle \text{e,t}\rangle}]\!]$
$= \lambda x.|x| = 100$
b. $[\![\text{hyaku}_{\langle \text{e,t}\rangle}\text{-no hon}]\!]$
$= \lambda x.|x| = 100 \ \& \ *\text{BOOK}(x)$

When $^{\cup}$ is applied, ClP is not projected, since Cl is not needed. That is, NumeP directly combines with NP just in non-classifier languages.

For non-large and non-non-specific numerals, the $^{\cup}$ operation is undefined and hence the corresponding predicts are not derived. Thus, those numerals always require classifiers to modify nouns. Further, type-shifted numerals cannot combine with classifiers as illustrated in (21).

(21) $[\![\text{hyaku}_{\langle \text{e,t}\rangle}\text{-satsu}_{\langle \text{n},\langle \text{e,t}\rangle\rangle}]\!] \rightarrow$ type mismatch

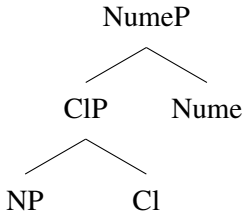
The combination results in a type mismatch. As a result, when a classifier appears, the only possible way to modify a noun is to use a numeral of the default type (type n) which a classifier turns into predicates.

3.2 Postnominal constructions

Now let us turn to the postnominal construction. As we have seen in the previous section, in the postnominal construction, classifiers are obligatory. I propose that the obligatoriness is due to the syntactic structure of the postnominal construction which is different from the one of the prenominal construction. Specifically, the obligatoriness of classifiers is due to the selectional requirement of numerals.

I follow Cheng and Sybesma (1999), Jenks (2017), Tang (1990) and among others, assuming that nominal phrases contain functional projections above NP. The extended nominal projection in Japanese has the following structure.[9]

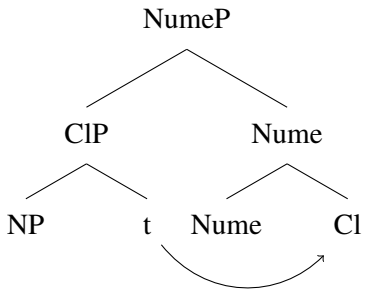
(22)



In this structure, classifiers and numerals are heads of their own projections, ClP and NumeP, respectively. I further postulate that Nume^0 selects for ClP. This ensures that whenever numerals are present, classifiers are present.

One may notice that the liner order derived by the structure in (22) is not a correct surface order. (22) produces an NP-Cl-Nume sequence but it should be NP-Nume-Cl. I suggest that Cl^0 is moved to Nume^0 obligatory.

(23)



I propose that the motivation of Cl^0-to-Nume^0 movement is the affixal status of numerals in Japanese. Nume^0 may contain a strong feature, which attracts Cl^0. A piece of evidence for the affixal nature is found in some numerals. In Japanese, there are two types of numerals: native and Sino-Japanese numerals. Native Japanese numerals, which are limited to number 1–10, cannot stand independently, except 4, 7 and 10 as shown in Table 1. Although Sino-Japanese numerals can be used independently, native Japanese numerals would suggest

[9] Similar to the prenominal construction, I assume that DP is located above NumeP. In both the construction, Case is assigned to DP. I thank an anonymous review for drawing my attention to the case assignment.

	1	2	3	4	5	6	7	8	9	10
Native	hito-	futa-	mi-	yo(n)	itsu-	mu-	nana	ya-	kokono-	too
Sino-Japanese	ichi	ni	san	shi	go	roku	shichi	hachi	kyuu	juu

Table 1: Numerals in Native and Sino-Japanese

that Japanese numerals are affixal in nature.

The result of the obligatory head movement is a complex head which behaves as a single word. In fact, the combination of numerals and classifiers shows some morpho-phonological effects. For example, the form of some classifiers alters depending on the preceding numerals. Consider the following examples, in which *-hon*, a classifier for counting cylindrical objects such as pens or fingers, shows the alternations *-pon* and *-bon*.

(24) a. ichi + hon → ip-**p**on
1 CL

b. ni + hon → ni-hon
2 CL

c. san + hon → san-**b**on
3 CL

In addition, the forms of numerals also change depending on the following classifiers. The following examples are the combination of numerals 1, 6, and 8 and a classifier *-ko*, which is used to count inanimate objects.

(25) a. ichi + ko → ik-ko
1 CL

b. roku + ko → rok-ko
6 CL

c. hachi + ko → hak-ko
8 CL

In this case, the forms of the numerals assimilate the first consonant of the classifier, yielding geminates. The morpho-phonological effects found in the combination of numerals and classifiers indicate that the tight connection between the two heads exists.[10] Kobuchi-Philip (2007) claims that the morpho-phonological effect such as (24) is accounted for by the head movement analysis.[11]

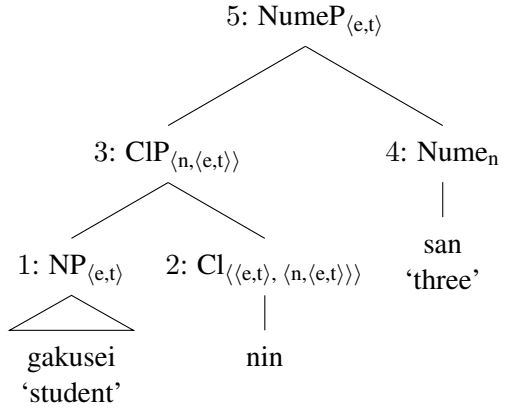
Given the syntactic structure, one may wonder whether the semantic analysis of Sudo (2016) is extendable to the postnominal construction. The biggest issue, however, is that combining Cl^0 with NP leads to a type mismatch since, in Sudo, classifiers are of type $\langle n,\langle e,t\rangle\rangle$ and nouns are of type $\langle e,t\rangle$. To solve this issue, I suggest that classifiers of type $\langle n,\langle e,t\rangle\rangle$ as in (26a) are to be type-shifted to type $\langle\langle e,t\rangle, \langle n,\langle e,t\rangle\rangle\rangle$ as in (26b).[12]

(26) a. $[\![\text{nin}_{\langle n,\langle e,t\rangle\rangle}]\!]$
$= \lambda n.\lambda x : *\text{HUMAN}(x).|x| = n$

b. $[\![\text{nin}_{\langle\langle e,t\rangle,\ \langle n,\langle e,t\rangle\rangle\rangle}]\!]$
$= \lambda P.\lambda n.\lambda x : *\text{HUMAN}(x).|x| = n$
$\&\ P(x)$

The derivation with the shifted classifier in (26b) is shown in (27) and (28).

(27)



(28) a. $1 : [\![\text{NP}]\!] = \lambda x.*\text{STUDENT}(x)$

b. $2 : [\![\text{Cl}]\!]$
$= \lambda P.\lambda n.\lambda x : *\text{HUMAN}(x).|x| = n$
$\&\ P(x)$

c. $3 : [\![\text{ClP}]\!]$
$= \lambda n.\lambda x : *\text{HUMAN}(x).|x| = n$

[10] As noted by a reviewer, the head movement is expected to occur in the prenominal construction as well. Given the affixal nature of numerals, Cl^0 moves to $Nume^0$ by lowering (Arregi and Pietraszko, 2018).

[11] Kobuchi-Philip (2007) proposes a similar head movement analysis but different semantics for numerals and classifiers.

[12] It is possible to assume that the lower type is derived from the higher type. In addition,

$\&$ *STUDENT(x)

d. $4 : [\![\text{Nume}]\!] = [\![\text{san}]\!] = 3$

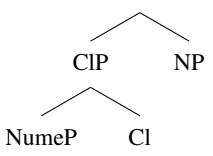
e. $5 : [\![\text{NumeP}]\!]$
$= \lambda x.|x| = 3 \ \& \ *\text{STUDENT}(x)$

Unlike the prenominal construction, Cl first combines with NP to form ClP.[13] NumeP has the denotation by combining ClP and Nume. The resultant denotation is identical to the one of the prenominal construction. In this analysis, the Cl^0-to-Nume^0 movement happens at PF. Thus, the head movement does not affect the interpretation.

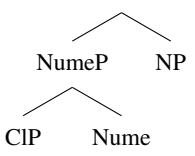
What is crucial in the postnominal construction is the assumption that Nume^0 selects for ClP. This forces the presence of both numerals and classifiers. Thus, the presence of classifiers is obligatory in the postnominal construction. The current proposal suggests that there are two syntactic types of numerals: one selects for ClP (as in the postnominal) and the other does not (as in the prenominal). Note that while syntactically numerals are different between the prenominal and postnominal constructions, semantically, when $^{\cup}$ is not applied, they are identical, namely, they are type-n objects (cf. Bale and Coon, 2014). The difference in the semantic type of NumeP between the two constructions (type n in the prenominal construction and type $\langle$e,t$\rangle$ in the postnominal one) is the result of the derivation.[14]

[13] A review raises a question why in the prenominal construction, Cl^0 selects for NumeP as in (ia) , not the other order way round as in the postnominal construction, namely, Nume^0 selects for ClP as in (ib) with the Cl^0-to-Nume^0 movement.

(i) a.



b.



It seems that the alternative structure (ib) works fine just like the structure proposed in the paper (ia). One potential challenge for the alternative structure, however, is how the optionality of classifiers is captured. As in the postnominal construction, it is assumed that the structure requires that classifiers always appear, because of the syntactic selection. We thus need to stipulate a syntactically different type of numerals which does not take ClP as its complement. In addition, the alternative structure will contain another NP in the complement of ClP. It is not obvious how we treat this NP. I leave it open to future work which structure is empirically and theoretically plausible.

[14] I would like to thank an anonymous reviewer for making the point explicit.

4 Some implications

4.1 Two structures for numeral classifiers

In the current analysis, the prenominal and postnominal constructions have different syntactic structures. This difference accounts for the asymmetry of the optionality. In contrast, in the previous literature, it has been argued that the two constructions are transformationally related (Watanabe, 2006). The transformational analysis, however, faces the difficulty of accounting for the asymmetry: since the two constructions are transformationally related, when a classifier is optional/obligatory in one construction, the same should hold in the other construction. As we have seen, classifiers are optional for the particular types of numerals in the prenominal construction, whereas obligatory in the postnominal construction. Thus, we need to hypothesize that a classifier can be dropped only in the prenominal construction. It seems not straightforward to defend this hypothesis in principled manner.[15]

The current analysis is also compatible with cross-linguistic observations on the numeral-noun constructions (Danon, 2012; Ionin and Matushansky, 2018). Danon (2012) examines a wide variety of languages, arguing that numerals are located in at least two syntactic positions: a head position and a specifier position. Though my analysis for Japanese numeral classifiers differs from Danon in that the prenominal numeral classifiers are adjuncts, what is crucial is that both the proposals assume that UG makes it possible for numerals and numeral classifiers to occupy a head and non-head positions.

4.2 The role of classifiers

There is a debate as to why classifiers are required in classifier languages. Krifka (1995) and Chierchia (1998a) provide different accounts. In Krifka (1995), classifiers are for numerals: they are needed

[15] In Watanabe (2006), floating numerals are also transformationally related to the two constructions discussed in this paper. As mentioned in footnote 4, it is not obvious whether classifiers can be optional in the floating quantifier construction. If classifiers are optional in the floating quantifier construction, it is not inconsistent with the idea that the prenominal and floating quantifier constructions are transformationally related. If, on the other hand, classifiers are obligatory in the floating construction, the postnominal and the floating constructions might be related as in Huang and Ochi (2014).

to make numerals to be able to specify which types of things that the numerals count (e.g., long and round things, flat things, inanimate things, humans etc.). Sudo's (2016) analysis is along the line of Krifka's, though in Sudo, classifiers supply numerals with a way to modify nouns by changing the type of numerals. By contrast, in Chierchia (1998a), classifiers are for nouns: they are required to enable nouns to be countable and modifiable by numerals.

Bale and Coon (2014) distinguish the two accounts by examining data from Mi'gmaq (Algonquian) and Chol (Mayan). They show that the presence/absence of classifiers in these languages depends on numerals. Bale and Coon argue that the pattern of the optionality is compatible with Krifka's classifiers-are-for-numeral analysis but not with Chierchia's classifiers-are-for-nouns one. Jenks (2017), on the other hand, observes that in Dafing (Mande: Burkina Faso), certain nouns do not appear with classifiers, concluding that in Dafing, classifiers are for nouns, not for numerals, that is, the pattern is consistent with Chierchia (1998a). The analyses of Bale and Coon, and Jenks suggest that two types of classifier languages exit.

The optionality of classifiers in the prenominal construction in Japanese indicates that the language is categorized as a classifiers-are-for-numeral language, since the presence/absence of classifiers depends on numerals. In addition, the syntactic structure for the prenominal construction reflects the assumption that classifiers are for numerals, since classifiers combine with numerals. However, the proposed structure for the postnominal construction looks as if it contradicts Krifka's theory, since Cl^0 selects for NP and is interpreted at the base position. Thus, the proposed structure for the postnominal construction appears to fit Chierchia's theory that classifiers are for nouns.

Nonetheless, I argue that the proposed analysis for the postnominal construction does not deviate from Krifka's theory. In the current analysis, as in Sudo (2016), the role of classifiers is to turn type-n objects into predicates and consequently numerals can modify nouns. This role remains intact after classifiers of type $\langle n,\langle e,t\rangle\rangle$ is shifted to type $\langle\langle e,t\rangle,\langle n,\langle e,t\rangle\rangle\rangle$. Even though a classifier combines first with a noun, the classifier turns a numeral of type n into a predicate. Thus, the proposed analysis for the postnominal is compatible with Krifka's theory.

An implication of this discussion is that whether classifiers are for numerals or for nouns seems independent of syntactic structures. In other words, syntactic aspects would not be a deciding factor for the types of languages in terms of the role of classifiers. Given this discussion, it is not surprising that there are classifiers-are-for-noun languages in which syntactically a classifier combines first with a numeral before the constituent of the numeral and classifier combines with a noun. It is expected that classifiers can make nouns countable, even though classifiers and nouns are not directly combined. Further typological investigations are called for to check the prediction.

5 Concluding remarks

This paper has analyzed the syntax and semantics of numeral classifiers in Japanese. We have seen that the asymmetry of the optionality of classifiers in Japanese between the two constructions is captured by the differences in the syntax structure. Specifically, in the prenominal construction, numerals and classifiers form a constituent, modifying a noun. In this construction, the type-shifting operation with the $^{\cup}$ operator is applicable if it is defined, resulting in the optionality of classifiers for large and non-specific numerals. In contrast, in the postnominal construction, numerals and classifiers are heads of the extended nominal projection and $Nume^0$ selects for ClP, which forces the presence of classifiers whenever numerals appear. The consequence of the analysis is that the two constructions are not transformationally related. I have also discussed some implications of the current analysis regarding the role of classifier. It is suggested that syntactic facts would not always reflect whether classifiers are for numerals or for nouns.

Acknowledgments

I would like to thank Mana Asano, Junko Hibiya, Seunghun Lee, Motoko Obata, Satoshi Tomioka and Tomoyuki Yoshida for comments and suggestions on earlier versions of this paper. I would also like to thank two anonymous reviewers for their helpful comments. All errors are mine.

References

Karlos Arregi and Asia Pietraszko. 2018. Generalized head movement. *Proceedings of the Linguistic Society of America*, 3(1):5–15.

Alan Bale and Jessica Coon. 2014. Classifiers are for numerals, not for nouns: Consequences for the mass/count distinction. *Linguistic Inquiry*, 45(4):695–707.

Lisa Lai-Shen Cheng and Rint Sybesma. 1999. Bare and not-so-bare nouns and the structure of NP. *Linguistic Inquiry*, 30(4):509–542.

Gennaro Chierchia. 1985. Formal semantics and the grammar of predication. *Linguistic Inquiry*, 16(3):417–443.

Gennaro Chierchia. 1998a. Plurality of mass nouns and the notion of "semantic parameter". In *Events and Grammar*, pages 53–103. Springer.

Gennaro Chierchia. 1998b. Reference to kinds across language. *Natural Language Semantics*, 6(4):339–405.

Gabi Danon. 2012. Two structures for numeral-noun constructions. *Lingua*, 122(12):1282–1307.

C-T James Huang and Masao Ochi. 2014. Remarks on classifiers and nominal structure in east asian. *Language and Linguistics Monograph Series*, 54:53–74.

Tania Ionin and Ora Matushansky. 2018. *Cardinals: The syntax and semantics of cardinal-containing expressions*. MIT Press.

Peter Jenks. 2017. Numeral classifiers compete with number marking: evidence from Dafing. Paper presented at Annual Meeting of the Linguistic Society of America, Austin, TX.

Mana Kobuchi-Philip. 2007. Floating numerals and floating quantifiers. *Lingua*, 117(5):814–831.

Manfred Krifka. 1995. Common nouns: A contrastive analysis of Chinese and English. In Gregory N Carlson and Francis Jeffry Pelletier, editors, *The Generic Book*, pages 398–411. University of Chicago Press.

Godehard Link. 1983. The logical analysis of plurals and mass terms: A lattice-theoretic approach. In Rainer Bäuerle, Christoph Schwarze, and Arnim von Stechow, editors, *Meaning, Use and the Interpretation of Language*, pages 303–323. de Gruyter.

Hiroki Nomoto. 2013. *Number in Classifier Languages*. Ph.D. thesis, University of Minnesota.

Susan Rothstein. 2013. A Fregean semantics for number words. In *Proceedings of the 19th Amsterdam Colloquium*, pages 179–186. Universiteit van Amsterdam Amsterdam.

Uli Sauerland. 2005. DP is not a scope island. *Linguistic Inquiry*, 36(2):303–314.

Yasutada Sudo. 2016. The semantic role of classifiers in Japanese. *Baltic International Yearbook of Cognition, Logic and Communication*, 11(1):10.

Yasutada Sudo. to appear. Countable nouns in Japanese. In Proceedings of 11th Workshop on Altaic Formal Linguistics (WAFL 11).

Chih-Chen Jane Tang. 1990. *Chinese Phrase Structure and the Extended X'-theory*. Ph.D. thesis, Cornell University.

Akira Watanabe. 2006. Functional projections of nominals in Japanese: Syntax of classifiers. *Natural Language & Linguistic Theory*, 24(1):241–306.

An emoticon is well worth a few empathetic words

Juan Pablo Rodriguez Gomez	Tomoko Iizuka	Edson T. Miyamoto	Changyun Moon	Kaoruko Ouchi
Graduate School of Humanities and Social Sciences, University of Tsukuba		Center for Meta-Learning, Future University Hakodate	Faculty of Humanities and Social Sciences, University of Tsukuba	Akamonkai Japanese Language school
pab.rdgz@gmail.com	tomoko.iizk@gmail.com	miyamoto@alum.mit.edu	moon.changyun.gf@u.tsukuba.ac.jp	o_kvein@yahoo.co.jp

Abstract

We report three studies providing evidence that Japanese college students judge emoticons to express emotions as much as empathetic verbal expressions. The effect is observed in the judgements to the message where the emoticon was included, but also in judgements to the replies to that initial message. The results hold for emails as well as for more recent messaging apps.

1 Introduction

Emoticons (short for *emotion icon*) are simple representations of faces using letters and symbols. They have been in use at least since the 1980s and started as a simple way of disambiguating rapidly typed messages (Fahlman, 2002; also McCulloch, 2019, chapter 5, for a detailed account). We report three questionnaires providing evidence that *kaomoji* (the Japanese equivalent of emoticons) are an integral part of machine-mediated communication for college students in Japan.

That emoticons can add affective content to text is not particularly striking. Even text layout or stationery choice may augment verbal information (e.g., cute, colorful stationery may imply a happy state of mind). The following describes three alternative uses for emoticons.

1. *Decorative use*: emoticons and other embellishments (e.g., stars, geometrical shapes) increase visual appeal and are similar to pretty stationery. Their role in expressing an emotion, although discernible, is limited.
2. *Utilitarian use:* limited time or physical constraints (e.g., the tiny keys of a mobile phone) lead to truncated, incomplete messages. Emoticons help express what is not explicit in words and are only effective when verbal content is incomplete. Emoticons are a quick and easy way to disambiguate the intended meaning and express what would take much longer to express explicitly in words. But emoticons are makeshift solutions and only effective when verbal content is incomplete or ambiguous.
3. *Emphatic use:* emoticons emphasize emotional content even when the words in the message explicitly express the emotion intended.

The categories above may overlap but they help us determine how integrated emoticons are in communication (see Derks, Fischer and Bos, 2008, for a review of related results and various possible uses of emoticons). Given previous results, emoticons are unlikely to be just decorative (Arakawa, et al., 2006; Derks, Bos and von Grumbkow, 2008; Thompsen and Foulger, 1996;

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 212-218, Hakodate, Japan, September 13-15, 2019

and references therein). But a utilitarian use would suggest that emoticons are poor substitutes that are ignored in normal circumstances when verbal information is explicit (but see To, 2008, who found that emoticons lead to more accurate interpretations regardless of whether the accompanying text was ambiguous or not).

A possible argument against utilitarian uses is that emoticons are not necessarily easier to type than linguistic expressions. By the time the studies reported were conducted, many mobile phones already contained canned verbal expressions, as easily retrievable as emoticons. Picking an appropriate linguistic expression may seem more complex given the nuances of language, but choosing an emoticon can be almost as daunting, given the extensive range of alternatives at users' disposal (see Kato et al., 2007, Table 2, for 163 facial emoticons). Moreover, they are not always restricted to the face alone and can include culture-specific images such as **m(_ _)m** (a bowing head; face level with the two *m*s representing the hands; eyes closed in contrition or gratitude), as well as those with the whole body such as **orz** (a person banging the head on the ground in frustration or desperation; *o* for the head, *r* for the arms, and *z* for the torso and legs).

We report data indicating that emoticons are used emphatically adding affect to explicit verbal content and imposing expectations on ensuing replies.

2 Study 1

We conducted a questionnaire to provide basic evidence for the effectiveness of emoticons uses.

2.1 Method

Participants: Sixteen native-Japanese students (9 female) at a national university in the Kanto area of Japan participated in the study for financial compensation based on on-campus rates for part-time work.

Stimuli: Twenty-four sets of messages were created. Participants were asked to rate how much each message expressed an emotion. Half of the messages described happy events (positive contexts), and the other half described upsetting events (negative contexts). The following is an example of a positive context with a smiley face at the end.

(1) この間面接に行った新しいバイト、無事に採用されたよ(^0^)
"The interview for the new part-time job, I got it without a problem [happy face]."

Each set contained four versions in a 2×2 within-participants design. The first factor was whether an emoticon or a full stop ended the message (see Kawakami, 2008, for judgements on different types of emoticons).

The second factor manipulated was the role that the participant was instructed to assume: as the sender or as the receiver of the message. We avoided using words and morphological endings that are stereotypically associated with one gender, so that both male and female participants could identify as the sender of any message.

Procedure: Each message was printed on a separate page within a frame depicting a mobile phone display. On top of the page a line of instruction indicated whether the participant was to assume the role of sender or receiver of the message. At the bottom of the page, participants rated how much the message expressed an emotion (e.g. for positive contexts: *yorokobi* "joy", for negative contexts: *ikari* "anger"; 1 not at all; 7 very much).

The 24 sets of messages (each set containing the four versions of each message) were distributed into four lists according to a Latin Square design, so that each list contained exactly one version from each set, and equal numbers of positive events (e.g., as in (1)) and negative events, with and without emoticon. Each list was stapled in a block in pseudo-random order so that items in the same condition did not follow in succession. Each participant saw one list in a within-participants design.

Analysis: All analyses were conducted on R (R Core Team, 2016). Rating was treated as an ordered factor and analyses were conducted with random-effects ordered logit models (function *clmm*, package *ordinal*; Christensen, 2015; similar trends were obtained with analysis of variance). Random structure of the models was determined through backward selection. Pairwise comparisons were conducted using least-square means with *Tukey* adjustments (function *lsmeans*, package *lsmeans*; Lenth, 2016).

Empathetic phrase	Emoticon received	Message received
-	-	Oh. Was it at a cram school?
with	-	Oh. Was it at a cram school? **Great that you got it.**
-	**with**	Oh. Was it at a cram school? **[happy face]**
with	**with**	Oh. Was it at a cram school? **Great that you got it [happy face]**

Table 1. Example of the four types of message received in Study 2.

2.2 Results and discussion

The factors included in the analysis were emoticon (with/without), role (sender/receiver), context (positive/negative) and all their interactions.

Overall, messages with emoticon elicited higher scores (mean 5.6) than messages without emoticon (4.4; β=2.24, P<.001), suggesting that emoticons help express emotions. This enhancing effect was larger in the positive contexts than in the negative contexts (β=1.42, P=.001; on emoticons being used more frequently in positive than negative contexts, see Derks, Bos and von Grumbkow, 2008; Park et al., 2013), but it was reliable in both types of contexts (Ps<.001).

All other effects were not reliable (Ps>.1). Previous reports indicate that participants tend to be egocentric and overestimate the effectiveness of their messages to express their intent such as sarcasm (Kruger et al., 2005). We failed to see such an effect in this study, perhaps because alternating between the role of sender and receiver made participants more sensitive to the effectiveness of the messages, or perhaps because we did not require the participants to type the messages they were supposed to send.

The results provide basic evidence that emoticons help express emotions. The following two studies build on this result to investigate emoticons in more detail.

3 Study 2

In this study, participants rated pairs of messages (a message sent and its reply) to determine how their reactions to the reply varied depending on the nature of the message sent. Moreover, we also manipulated the amount of verbal content to determine whether explicitly expressing empathy with words would cancel the effectiveness of emoticons.

3.1 Method

Participants: A new group of 28 native-Japanese students (11 female) from the same population as Study 1 were paid to participate in the study.

Stimuli: The 24 messages from Study 1 were used as *messages sent*, which participants were asked to assume they had sent to a friend. An item consisted of a message sent paired with the friend's reply (the *message received*). Each item had eight versions according to the following three factors in a 2×2×2 within-participants design. (See Table 1 for an example of the four types of message received in response to example (1).)

(2) Factors in Study 2
a. *emoticon sent:* whether the message sent contained an emoticon;
b. *empathetic phrase:* whether the message received contained an empathetic phrase;
c. *emoticon received:* whether the message received included an emoticon.

The message received always contained a neutral expression that did not give away the friend's feelings (e.g., (3) as a response to (1)).

(3) Neutral text in a message received without emoticon
おー。塾講だっけ。
"Oh. Was it at a cram school?"

An emoticon after (3) should have a clear effect following such a neutral expression, as it complements its meaning. But if emoticons only have decorative or utilitarian uses, in other words if they only have an effect when the words are ambiguous or insufficient to express an emotion, their effect should be neutralized by an overt expression of empathy and should have no effect

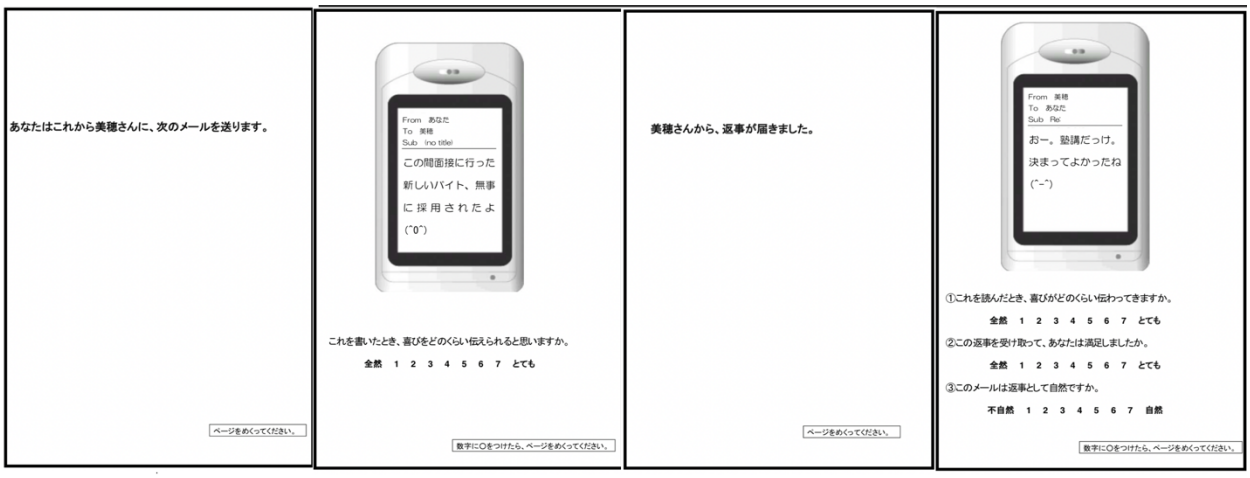


Figure 1. The four pages of an item in Study 2.

when following an explicitly empathetic phrase as in (4).

(4) Empathetic phrase with emoticon
決まってよかったね(^-^)
"Great that you got it [happy face]"

But if emoticons can be used emphatically, they may add to the emotional content already expressed by the verbal message.

Procedure and analysis: Each item was printed on four successive pages containing a message sent and a message received (see Figure 1 for an example item with the message in (1) on page 2 and the messages in (3, 4) on page 4). On the first page, the participants were told that they were about to send a message to a friend. The second page had the message sent printed within the frame of a mobile phone and Question 1 at the bottom. On page 3, the participant was told that a response from the friend was being received. Page 4 had the message received and questions 2 to 4.

Participants were instructed to assume that the person they were interacting with was a friend. The name of the person appeared as the recipient on page 2 and as the sender on page 4. Common female names were used.

Participants answered four 7-point scale rating questions ('1' not at all, and '7' very much). Question 1 was shown immediately after the message sent and asked how much this message conveyed a feeling (*yorokobi* "happiness" or *ikari* "anger"). The last three questions were shown immediately after the message received. Question 2 asked how much the message received expressed a feeling (same as in Question 1). Question 3 asked whether the message received was a satisfactory reply. Question 4 asked whether the message received was a natural reply.

The 24 sets of items (each set containing eight versions) were distributed into eight lists according to a Latin Square design, so that each list contained one version from each set and the same number of each version. Each list was stapled in a block in pseudo-random order so that items of the same type did not follow in succession. Each participant saw one list with 24 items.

Data analysis was conducted as in Study 1.

3.2 Results and discussion

Results were as follows.

Question 1 (about the message sent): replicated the results of Study 1. There was a main effect of emoticon as messages sent with emoticon (mean 5.93) were rated higher than those without emoticon (4.18; β=2.83, P<.001). There was also an interaction between context and emoticon as the emoticon effect was larger for positive than for negative contexts (β=1.37, P=.017).

Question 2 (about the message received): Results were as follows.

(5)

a. *Emoticon sent.* Messages received were rated higher if they were replies to a message sent *without* an emoticon (mean 4.45) than if they were responses to a message sent *with* an emoticon (4.23; P<.001). That is, sending a

message with an emoticon heightens the expectation for an empathetic response, leading judgements about the message received to be stricter.

Overall patterns indicated that emoticon sent did not interact with other factors. Moreover, type of context (positive or negative) only affected the effect sizes, but not their directions. Therefore, emoticon sent and context were not included in the remaining analyses reported.

b. *Empathetic phrase.* Messages received with an empathetic phrase were rated higher (mean 4.93) than those without them (3.75; β=1.75, P<.001). This guarantees that the phrases used (e.g., (4) without the emoticon) were effective in expressing an empathetic response.
c. *Emoticon received.* Messages received with an emoticon were rated higher (mean 5.03) than those without an emoticon (3.66; β=2.05, P<.001).
d. *Emoticon received vs empathetic phrase.* Messages received with emoticon and without empathetic phrase (mean 4.62) were as effective as those without emoticon and with empathetic phrase (4.43), suggesting that emoticons were as effective as the empathetic phrases (P=.75).
e. *Emoticon received plus empathetic phrase.* There was an interaction between emoticon received and empathetic phrase (β=−1.17, P<.001) as the effect of the emoticon was smaller when there was an empathetic phrase (1.0) than when there was no such a phrase (1.74). This is unsurprising. What is more crucial is that although smaller the effect of the emoticon is reliable even when there is an empathetic phrase (P<.001). In other words, the emoticon increases the empathy conveyed by the empathetic phrase.

The results to questions 3 and 4 revealed trends similar to those in question 2, therefore they are not reported.

The results suggest that already in 2010, when the ratings were collected, college-age native Japanese speakers were using emoticons to express emotional content and accepted them as much as short empathetic verbal phrases.

4 Study 3

The data for Study 2 was collected in 2010, therefore a new study was conducted in 2017 to replicate it by simulating exchanges in a messaging app commonly used in Japan these days.

Moreover, a concern in Study 2 is that the answer to Question 1 (about the message sent) may have affected the response to Question 2 (about the message received). Therefore, in this replication we asked one single question about each item: whether the reply expressed a given emotion (corresponding to Question 2 of Study 2).

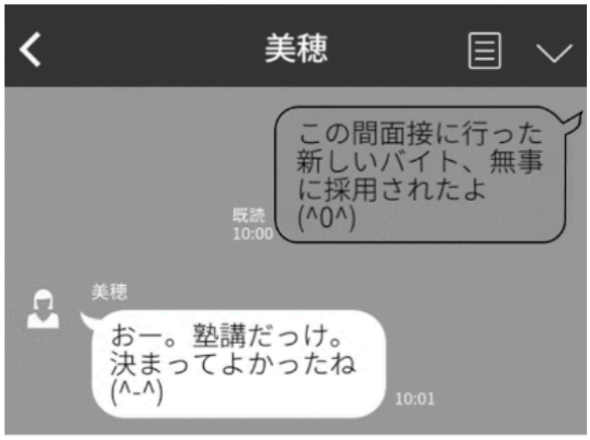


Figure 2. Example item of Study 3. The speech bubble on the top right is the message sent. The bottom left bubble is the friend's reply.

To prevent participants from going back to previous items, items were presented one a time on a computer screen using a modified version of Doug Rohde's Linger program (available from http://tedlab.mit.edu/~dr/Linger/).

4.1 Method

Participants: A new group of 40 participants (30 female) from the same population were paid to participate.

Stimuli: The messages and manipulations were the same as in Study 2, with some information updated (e.g., by removing the name of a rock band that had fallen out of favor). Figure 2 illustrates how the messages in examples (1, 3, 4) were presented as pictures (created using a freely available service at http://www.mojimaru.com/talk and simulated the appearance of a popular messaging app in Japan).

Procedure and analysis: Items were presented one a time on a computer screen in a different random order for each participant. Analyses were conducted in the same way as in Study 2.

Results: Trends replicated the results of Study 2 as summarized next.

(6)

a. *Emoticon sent.* Messages received were rated higher if they were responses to a message sent *without* an emoticon (mean 4.22) than if they were responses to a message sent *with* an emoticon (4.13; β=−.75, P=.011).
b. *Empathetic phrase.* Messages received with an empathetic phrase were rated higher (4.72) than those without them (3.62; β=1.65, P<.001). This confirms that the phrases used (e.g., (3) without the emoticon) were effective in expressing an empathetic response in this study as well.
c. *Emoticon received.* Messages received with an emoticon were rated higher (mean 4.68) than those without an emoticon (3.67; β=1.55, P<.001).
d. *Emoticon received vs empathetic phrase.* Messages received with emoticon and without empathetic phrase (mean 4.32) were as effective as those without emoticon but with empathetic phrase (4.40), suggesting that emoticons were as effective as the empathetic phrases (P=.98).
e. *Emoticon received plus empathetic phrase.* There was an interaction between emoticon received and empathetic phrase (β=−1.06, P<.001) as the effect of the emoticon was smaller when there was an empathetic phrase (1.41) than when there was no such a phrase (1.50). Like in Study 2, the effect of the emoticon is reliable even when there is an empathetic phrase *(*P<.001). In other words, the emoticon increases the empathy conveyed by the empathetic phrase in this study as well.

The results suggest that the ratings remained consistent despite the passage of time and the different types of media involved (email in Study 2, messaging app in Study 3).

5 General Discussion

Our findings can be summarized as follows.

A. Sending an emoticon creates the expectation for an empathetic response (see (5a) and (6a)). But the response need not contain an emoticon. As long as it conveys empathy (through words or an emoticon), the response is rated as an acceptable reply to the initial message sent with an emoticon.
B. Emoticons are judged to express an emotion (see (5c) and (6c)) and can be as expressive as a few empathetic words (see (5d) and (6d)).
C. Even when the verbal message is unambiguous, emoticons can emphasize their emotional content (as in (5e) and (6e); see To, 2008, for similar trends).
D. The role of emoticons has been stable between 2010 (when Study 2 was conducted with emoticons embedded in email exchanges) and 2017 (when Study 3 was conducted with the same stimuli simulating a messaging app).
E. Emoticons are likely to be more acceptable in happy, positive events than in negative ones in line with previous reports (Derks, Bos and von Grumbkow, 2008; Park et al., 2013; *inter alia*).

In sum, emoticons are like emotion-expressing punctuation and have become a form of paralinguistic information akin to prosody (Asteroff, 1987, for an early discussion; also McCulloch, 2019, chapter 5, for a discussion on emoticons as gestures).

However, some caveats are in order. First, a possible concern in all three studies reported here is that the condition without emoticon always ended with a *maru* (the Japanese equivalent of a sentence-ending full stop). Recent reports suggest that full stops tend to be judged negatively in typed messages in English (Gunraj et al., 2016). Informal judgments suggest similar trends in Japanese college students. Therefore, in our studies, it is possible that at least part of the effect was caused by the negative effect of the full stop, rather than the expressiveness of the emoticons. This possibility requires further study, but some trends in the data suggest that the negative effect of full stops may not be enough to explain the results. For example, empathetic phrases without emoticon always ended with a full stop; nevertheless, they were rated favorably (see (5b,d) and (6b,d)). Moreover, the negative effect of full stops may be restricted to short messages (McCulloch, 2019,

chapter 4; preliminary results using the items in Study 3 tend to support this possibility).

Another concern is that previous work suggests that women use more emoticons than men (Tossell et al., 2012; and references therein). Preliminary analyses did not find gender differences, but this is also an area that merits more detailed analysis in the future. A factor that is likely to be relevant is that in our studies, the people who participants were asked to interact with always had female names, because we assumed it to be easier for participants to interact using emoticons with a female friend (see Fullwood et al., 2013, for a summary of results suggesting that males are more likely to use emoticons in mixed-sex environments).

6 References

Arakawa, A., Takehara, T., & Suzuki, N. (2006). The effects of various kinds of receiver's emotions on sender's use of emoticons [in Japanese]. *Japanese Journal of Research on Emotions*, *13*(2), 49-55.

Asteroff, J. F. (1987). *Paralanguage in electronic mail: A case study*. Doctoral dissertation. Columbia University.

Christensen, R. H. B. (2015). Ordinal-Regression Models for Ordinal Data. R package version 2015.6-28. *R Foundation for Statistical Computing: Vienna*. http://www.cran.r-project.org/package=ordinal/.

Derks, D., Bos, A. E. R., & von Grumbkow, J. (2008). Emoticons in CMC: Social motives and social context. *Cyberpsychology and Behavior*, *11*, 99-101.

Derks, D., Fischer, A. H., & Bos, A. E. R. (2008). The role of emotion in computer-mediated communication: A review. *Computers in Human Behavior*, *24*, 766-785.

Fahlman, S. E. (2002). *Smiley Lore* :-) . https://www.cs.cmu.edu/~sef/sefSmiley.htm (Retrieved 30 May, 2019.)

Fullwood, C., Orchard, L., & Floyd, S. (2013). Emoticon convergence in Internet chat rooms. *Social Semiotics*, *23*, 648-662.

Gunraj, D. N., Drumm-Hewitt, A. M., Dashow, E. M., Upadhyay, S. S. N., & Klin, C. M. (2016). Texting insincerely: The role of the period in text messaging. *Computers in Human Behavior*, *55*, 1067-1075.

Kato, S., Kato, Y., Kobayashi, M., & Yanagisawa, M. (2007). Analysis of the kinds of emotions interpreted from the emoticons used in e-mail [in Japanese]. *Japan Society of Educational Information*, *22*, 31-39.

Kawakami, M. (2008). The database of 31 Japanese emoticon with their emotions and emphases [in Japanese]. *The Human Science Research Bulletin*, *7*, 67-82.

Kruger, J., Epley, N., Parker, J. & Ng, Z.-W. (2005). Egocentrism over e-mail: Can we communicate as well as we think? *Journal of Personality and Social Psychology*, *89*, 925-936.

Lenth, R. V. (2016). Least-squares means: the R package lsmeans. *Journal of Statistical Software*, *69*(1), 1-33.

McCulloch, G. (2019). *Because Internet: Understanding the New Rules of Language*. New York: Riverhead Books.

Park, J., Barash, V., Fink, C., & Cha, M. (2013). Emoticon style: Interpreting differences in emoticons across cultures. In *Seventh International AAAI Conference on Weblogs and Social Media*.

R Core Team (2016). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Thompsen, P. A., & Foulger, D. A. (1996). Effects of pictographs and quoting on flaming in electronic mail. *Computers in Human Behavior*, *12*, 225-243.

To, N. M-L. (2008). *Influence of emoticons on message interpretation in computer--mediated communication*. Master's Thesis, Trinity Western University, Langley, Canada.

Tossell, C. C., Kortum, P., Shepard, C., Barg-Walkow, L. H., & Zhong. L. (2012). A longitudinal study of emoticon use in text messaging from smartphones. *Computers in Human Behavior*, *28*, 659-663.

Utilization of histories by country in question-answering system to solve world history essay type questions

Kotaro Sakamoto[1] **Yuta Fukuhara**[1] **Madoka Ishioroshi**[2]

Kousuke Ohya[1] **Keigo Iwasaki**[1]

Hideyuki Shibuki[2] **Tatsunori Mori**[1]

[1]Yokohama National University, Japan
[2]National Institute of Informatics, Japan
{sakamoto, yuta_f, kosuke-o, i-keigo, mori}@forest.eis.ynu.ac.jp
{ishioroshi, shib}@nii.ac.jp

Abstract

We propose a query expansion method which uses the book of histories by country as a knowledge source for the question-answering system about world history complex essay type questions from the University of Tokyo's entrance examination. As the result of the experiment, the recall increased distinctly but the precision decreased instead. Therefore, there was an improvement from the viewpoint of summarization's preprocessing.

1 Introduction

Research on real-world complex question-answering (QA) has flourished in recent years. Notable examples are the QA Lab tasks at the NTCIR workshop[1] (Shibuki et al., 2017; Shibuki et al., 2016; Shibuki et al., 2014) and Todai Robot Project[2] (Fujita et al., 2014). The QA Lab has a task for essay type questions of Japanese university entrance examinations about world history. There are two kinds of essay type questions: complex essay type questions with limits of about 500 characters and simple essay type questions with limits of less than or equal to 200 characters. We are developing a complex essay type question-answering system (Sakamoto et al., 2017). The task of satisfying the information need is addressed by obtaining information from textbooks and a glossary as knowledge sources. Also, since essays have maximum numbers of characters, the system's approach is an extractive multi-document summarization aimed at satisfying the information need. The system pipeline consists of a first part which retrieves needed information from knowledge sources, and a second part which summarizes retrieved information adequately. This system achieved a certain result, but there is still room for improvement. For example it can not retrieve relevant descriptions which need to be included in the answer. In the paper (Fukuhara et al., 2017), we investigated the correspondence between gold standards[3] and knowledge sources. Also, we noticed the 85.1% of descriptions in gold standards corresponded to knowledge sources, where the rate was calculated in characters. Therefore, in regard to the problem of the former retrieval part of the pipeline, low ability to retrieve relevant descriptions from knowledge sources is worse than shortage of knowledge sources. We think that there are two reasons why the system cannot retrieve relevant descriptions. The first reason is that expressions in relevant descriptions do not match expressions in questions. The second reason is that expressions in relevant descriptions are not included in the questions. When we focus on the latter, the system needs a framework to expand phrases in a question in order to associate them to phrases in relevant descriptions. For example a person who learnt world history can associate "Napoleon" with "French campaign in Egypt" and "battle of Waterloo". We developed a module based on the above. According to the book (Sato, 2016), there are four types of complex

[1]http://research.nii.ac.jp/ntcir/index-en.html
[2]https://21robot.org/

[3]As gold standards, we used answers in Akahon, which is a book series of Japanese universities' past examinations. https://akahon.net/

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 219-227, Hakodate, Japan, September 13-15, 2019

Figure 1: A translation of an example of complex essay type questions

> Egypt, which created a brilliant ancient civilization, went on to maintain an unbroken line of history for 5,000 years. This history was based in the nation's rich land, but any discussion of it must include the political powers which came from near and far and left their profound marks, and their relationship with the active response by Egypt.
>
> Paying close attention to this background, provide an overview of the development of Egypt since the birth of its civilization, taking into consideration both 1. the interests of those arriving in Egypt and the reasons for their advances into Egypt, and 2. the policies and actions taken by Egypt in response to these advances. Limit your answer to 18 lines (540 Japanese characters) or less. Use each of the eight terms below once, and underline each term when it is used.
>
> Battle of Actium, Islam, Ottoman Empire, Saladin, Nile River, Nasser, Napoleon, Muhammed Ali

essay type questions about world history in the University of Tokyo's entrance examination. One of them asks to describe the course of history of a country or an area.[4] The Figure 1 shows an example of this type of complex essay type questions and the Figure 2 shows an example of gold standards of the question. In this question, if knowledge sources written about historical events about Egypt are available, it might be possible to conjecture phrases which are not written in the question. As one of such knowledge sources, there is a book of histories by country (Imaizumi et al., 2007) as shown in Figure 3. The book plainly describes a specific country's important events on multiple themes such as war or politics. Based on the

[4]The other three types are *"to describe historical courses and relations among several areas"*, *"to describe relations between some specific elements and historical courses in each area"* and *"to compare two or more historical events"*.

Figure 2: A translation of a gold standard of the question

> Independent dynasties flourished for many years in ancient Egypt, centered on the Nile River, but were conquered by Alexander the Great, et al. Cleopatra, in Ptolemaic dynasty Egypt, allied with Antony during the civil war in Rome, hegemon of the Mediterranean. Together, they fought Octavian, but were defeated at the Battle of Actium, and Egypt became a Roman province. The Islamic forces which unified the Arabian Peninsula during the 7th century BCE took advantage of the conflict between the Eastern Roman Empire and Sasanian Empire Persia to expand their territory, conquering Egypt. In the Fatimid Caliphate, which inherited part of the territory of the Muslim Empire, Saladin became vizier in 1196, and fought the Crusaders, who sought to recapture the holy land of Jerusalem from the Islamic forces. Similarly, the Mamluk Sultanate, established in Egypt during the fight against the Crusaders, was invaded by the Ottoman Empire and fell. In the modern era, French leader Napoleon invaded Egypt to break the communication line between Britain and India, proclaiming himself a liberator, but he met with opposition from the people. When the war against France broke out, Muhammed Ali, dispatched to Egypt by the Ottoman Empire, established a dynasty in Egypt. Britain became involved in the internal politics of Muhammed Ali's dynasty, and the country became a British protectorate. Nasser launched a revolution in 1952, overthrew the monarchy, and led a war against Israel.

background above, we proposed a query expansion method which uses the book of histories by country as a knowledge source for the question-answering system about world history complex essay type questions from the University of Tokyo's entrance examination. We define the knowledge sources for writing essay type answers in the paper (Sakamoto et al., 2017) as *answer knowledge sources* (AKS), and the knowledge source for the query expansion in this paper as *query expansion knowledge source* (QEKS).

2 The Structure of the Complex Essay Type Question

Figure 1 shows an example of the complex essay type question of world history, which is an English translation from the original Japanese version. The question contains additional text besides the main essay topic. The first paragraph gives background information, and the texts below the essay topic are the constraints for writing the essay. The constraints include a length limitation of *"18 lines (540 Japanese characters) or less"*, a condition of keywords that have to be included in an answer, which are *"Battle of Actium, Islam, Ottoman Empire, Saladin, Nile River, Nasser, Napoleon, Muhammed Ali"*, a geographical condition of *"Egypt"*, a chronological condition of *"since the birth of its civilization"* and a theme condition that the answer has to be written while focusing on *"1. the interests of those arriving in Egypt and the reasons for their advances into Egypt, and 2. the policies and actions taken by Egypt in response to these advances"*. Note that the keywords are not concentrated in textbooks or glossary and there is no passage including all or many keywords, so there is a need to find descriptions including the keywords from multiple passages in textbooks and glossary.

3 Related Work

3.1 Non-Factoid Type Question-Answering

The questions of previous works about non-factoid type question-answering (Cohen et al., 2018; Agichtein et al., 2015; Mitamura et al., 2010) are mostly brief such as "Why doesn't U.S. ratify the Kyoto Protocol?". However, the complex essay type question is much longer and has many constraints such as a length limitation, keywords that have to be included in an answer, a geographical condition, a chronological condition, a theme condition and so on.

3.2 Extractive Multi-Document Summarization

Our approach to solve the questions is an extractive multi-document summarization. Extractive multi-document summarizations (Erkan and Radev, 2004; Ng et al., 2012; Donghong and Yu, 2008) generally generate summaries consist of only significant events retrieved from AKS and reduce redundancy as far as possible. By contrast, the proposed method has two steps: to uncover the events which are the main points of the answer by using QEKS, and subsequently to retrieve the corresponding descriptions from AKS.

3.3 Query Expansion

The proposed method has a query expansion. Query expansion generally aims to cause the new query to match other semantically similar terms, with a thesaurus or WordNet, via automatic thesaurus generation, or techniques like spelling correction (Manning et al., 2009; Wollersheim et al., 2005), by contrast, the proposed method's query expansion aims to gain chronological and descriptive before-and-after relations by QEKS.

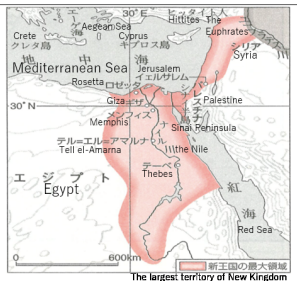
4 The Book of Histories by Country

We digitalized and used the book of histories by country. Textbooks and reference books which can be used as AKS describe many events, including relatively less important events, which are also described in detail as part of the narration. On the other hand, in the book of histories by country, the number of described events is smaller and only important events are outlined. Also, descriptions as they are in the book of histories by country are useless for generating essay type answers because they end with a sentence with a noun or noun phrase, and use symbol notation and so on as shown in Figure 3.

Since an essay type answer has a maximum number of characters, all the less important events can not be included in the answer, which instead has to cover important events in an extensive manner and also describe the details of each the important event.

(1) **National unity of Egypt**

Old Kingdom	Period and the capital – – – the 27th～the 22nd century B.C.E. : The capital is **Memphis** Formation – ┌ –「**Egypt was the gift of the Nile**」 (**Herodotus** who is a Greek historian) ├ –Agriculture which uses rise and fall of water level of **the Nile** → need for joint work and control to manage the river └ –Dynasty who integrates **Lower Egypt** as a delta area and **Nome** (a village) in **Upper Egypt** as a river valley appears The period f **Pyramids** – – – in the period of Fourth Dynasty, autocrat **Pharaoh** (king) reaches the peak prosperity └ – Pyramids of **Khufu**, Khafre and Menkaure in **Giza** Downfall – – – decentralization by weakening royal authority
Middle Kingdom	Period – – – the 21st～18th century B.C.E. The capital – – – transferred to **Thebes** → guardian deity **Ammon** (**Amun**) of Thebes becomes the chief deity After the fall – – – failing into turmoil because of invasion by Asian nomads **Hyksos** who have horses and tanks
New Kingdom	Period – – – 1567 B.C.E.～1085 B.C.E. The capital – – – **Thebes** Formation – – – to banish Hyksos and innovate their military technologies to invade Syria and conflict with Mitanni and Hittites **Akhenaten IV** (throne : about 1351 B.C.E.～about 1334 B.C.E.) ┌ – To force people to have faith in the only one God **Aton** and │ rename himself to Akhenaten └ – Transfer the capital to **Tell el-Amarna**



New Kingdom of Egypt

Figure 3: A translation of an example of histories by country (history of Egypt)

Therefore, our approach is to uncover the phrases which are the main points of the answer by using the book of histories by country, and subsequently to retrieve the corresponding descriptions from AKS.

5 Retrieval from QEKS

The number of events described in the book of histories by country is relatively small, but there are still too many events to include them all in an answer. Somehow they have to be narrowed down. As shown in the Figure 1, the question focuses on Egypt and requires to write an answer on the theme of war as written in *"1. the interests of those arriving in Egypt and the reasons for their advances into Egypt"* and politics as written in *"2. the policies and actions taken by Egypt in response to these advances"*. We narrow down relevant phrases in the book of histories by country by using the above themes as basis. The paper (Kawazoe et al., 2014) proposes an ontology of world history which broadly covers historical events in detail. Considering a set of events in the ontology, we annotated each sentence in the book of histories by country with 51 tags[5] which become themes. Table 1 shows the annotated tags. In this way, we can find a set of sentences in the book of histories by country, which are associated with themes. We define the above as query expansion by themes.

Another method to narrow down phrases is to find connected phrases, for example "Napoleon" is associated with "French campaign in Egypt" and "French campaign in Egypt" is associated with "Rosetta stone". Like the example above, we listed associated phrases one by one. Usually, keywords as shown in Figure 1 are named entities. Also, the relation between "Napoleon" and "French campaign in Egypt" is judged by co-occurrence in the book of histories by country as written in "French campaign in Egypt by Napoleon (1798-99) ··· with this as a trigger, the nationalism was exalted" Therefore, in this paper, we extract named entities which are included in the book of histories by country as nodes and we link the nodes which are included in the same description. Figure 3. A network structure as shown in Figure 4 is built. By following the links, we can expand a named entity to other named entities. However, as expected in a book of histories by country, country names appear much more frequently than other named entities, so most nodes link to nodes of country names. If we use a network structure like the above, needless phrases are retrieved. Hence,

[5]50 tags refer to the set of events in the ontology of world history and 1 tag was created by us

command by a weak agent	arrival of the peak period
extinguishing of family or dynasty	separation of an organization
transfer of capital	invasion
turmoil	banishment from a place
military advance	battle among countries
discovery	success of a weak agent
use of a technology	domination
independence of a country	expedition
war	defeat
loss of territory	trade
military victory	establishment of organization
support	to collect people
annexation	admission into an organization
opposition	dispatch of troops
expansion of territory	opening of war
international conference	approval
conclusion of treaty or agreement	management of an organization
opening a road	against law, system and/or establishment
military intervention	national independence movement
beginning	cooperation
political revolution	enforcement of laws, systems and policies
military defeat	to represent an organization
signature of a contract	attack on a person
democratic movement	ethnic migration
aggression	religious movement
establishment of dynasty or country*	(* Only this theme was created by us.)

Table 1: Themes/tags annotated to the book of histories by country

we decided not to use country names as nodes of the named entities' network. We define the above as query expansion by named entities' network.

The keywords in a question are important clues to write an answer, and they suggest chronological boundaries of events that should be written. For example, with regards to "Napoleon", he was alive from the latter half of 18 century to the former half of 19 century, so descriptions in the book of histories by country, which include "Napoleon", are written about events around the same period. Since descriptions of history by country are written in chronological order as shown in Figure 3, the first and the last descriptions including keywords are chronologically the first and the last events respectively. Therefore, descriptions can be further narrowed down by removing descriptions which are out of the boundaries. This idea can be adopted to both query expansion by themes and query expansion by named entities' network.

6 Proposed Method

Figure 5 shows the proposed method's pipeline. The proposed method consists of two steps. In the first step, phrases associated with the question are extracted from QEKS. In the second step, phrases extracted in the first step are used as retrieval query, and descriptions which should be included in the answer are extracted from AKS. In the extraction of themes, themes such as "war" or "politics" are extracted from the question. This process should be automated in the future but the precision is too low at this time, therefore the second author manually did it. In the query expansion by themes, we find a set of sentences associated with the themes using annotated tags in Section 5, and gain a set of named en-

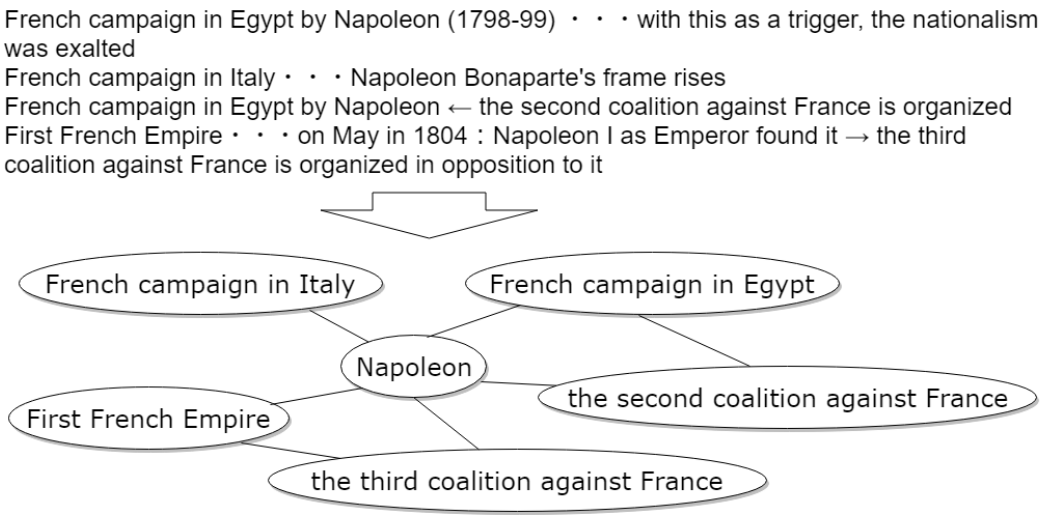


Figure 4: A translation of an example of how a named entities' network is built

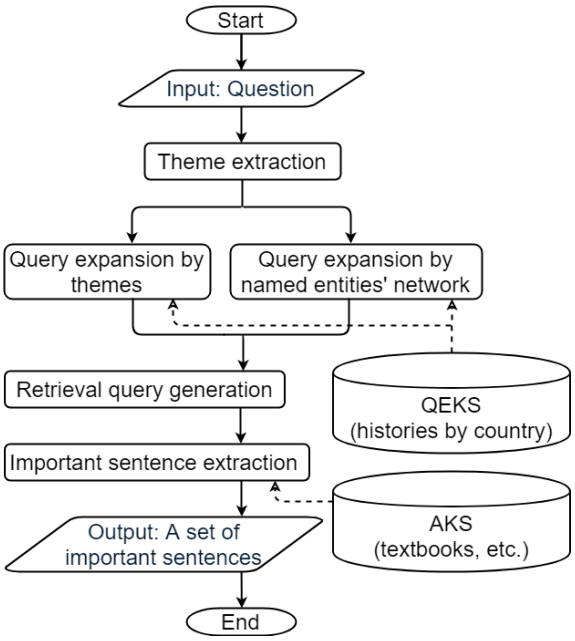


Figure 5: The proposed method's pipeline

tities from the sentences. In the query expansion by named entities' network, keywords in the question are used as starting nodes. By following the links of the named entities' network, we gain a set of named entities corresponding to nodes until n nodes ahead from the starting nodes. When named entities' network are being built, named entities judged as "POS subcategory 3" or "country" in analysis results of Japanese morpheme analyzer MeCab[6] are removed. In the generation of retrieval query, we unite the two sets of named entities collected from both query expansions. In the extraction of important sentences, as with the paper (Sakamoto et al., 2017), all the passages[7] in which named entities as retrieval queries are included are retrieved from AKS, and sentences are extracted from the passages including the named entities. Note that, in the previous method in the paper (Sakamoto et al., 2017), all the passages in which the keywords, not the named entities, retrieval queries are included are retrieved from AKS, and sentences are extracted from the passages including the keywords, not the named entities.

7 Experiment

We investigated if there is an improvement with the proposed method compared to the previous method. In the experiment, we used two questions from the

[6]https://taku910.github.io/mecab/

[7]a passage is a paragraph in textbooks or a description of a term in the glossary

	Sentence				Passage			
	Recall		Precision		Recall		Precision	
	2001	1999	2001	1999	2001	1999	2001	1999
Previous	0.133	0.123	0.024	0.118	0.656	0.695	0.080	0.394
Proposed	0.315	0.301	0.020	0.017	0.918	0.966	0.046	0.036

Table 2: Recall and precision of retrieving important sentences and passages

	Recall		Precision	
	2001	1999	2001	1999
Query expansion by themes	0.846	0.647	0.647	0.647
Query expansion by named entities' network	0.615	1.000	0.485	0.654
Intersection of both the results	0.462	0.647	0.632	0.846
Union of both the results	1.000	1.000	0.510	0.567

Table 3: Recall and precision of query expansion (results of extractions from QEKS)

University of Tokyo's entrance examination of 2001 and 1999. We set the n value of the named entities to 1, based on the result of a preliminary experiment. As evaluation index, we used a sentence-based recall R_s and precision P_s as follows.

$$R_s = \frac{\text{\# of retrieved important sentences}}{\text{\# of important sentences}} \quad (1)$$

$$P_s = \frac{\text{\# of retrieved important sentences}}{\text{\# of retrieved sentences}} \quad (2)$$

An important sentence is defined as a sentence of AKS which can be traced to gold standards. We also calculate a passage-based recall R_p and precision P_p as follows.

$$R_p = \frac{\text{\# of retrieved important passages}}{\text{\# of important passages}} \quad (3)$$

$$P_p = \frac{\text{\# of retrieved important passages}}{\text{\# of important passages}} \quad (4)$$

An important passage is defined as a passage including an important sentence.

8 Discussion

The results of the experiment are shown in Table 2. Although the results of the questions in 2001 and 1999 show little differences, on the whole they are both almost equal to each other. We reckon that there is no effect on the results from questions. Comparing the previous method and the proposed method, one can notice that the recall distinctly went up, but the precision went down instead. Therefore, there was an improvement from the viewpoint of summarization's preprocessing. Considering that the proposed method is the retrieval part of a complex essay type question-answering system, it is possible that descriptions which are not needed in the answer are removed by the part of the system devoted to summarization. In that sense, in the retrieval part, it is important to retrieve descriptions which are needed to write the answer comprehensively, so we put weight on recall rather than precision. Therefore, the fact that the passage-based recall increased remarkably is an important improvement, as indicated for example in the rise from 0.656 to 0.918 in the question in 2001 and from 0.695 to 0.966 in the question in 1999. Next, we discuss the effects of the query expansion by themes and the query expansion by named entities' network. In order to evaluate the result of the two query expansions, we define DQEKS as *descriptions in QEKS*, and important DQEKS as DQEKS that can be traced to gold standards. Also, we define output DQEKS as outputs of both query expansions that are not named entities but descriptions instead, and output important DQEKS as output DQEKS that can be traced to gold standards. We calculated the recall R_e and the precision P_e as follows.

$$R_e = \frac{\text{\# of output important DQEKS}}{\text{\# of important DQEKS}} \quad (5)$$

$$P_e = \frac{\text{\# of output important DQEKS}}{\text{\# of output DQEKS}} \quad (6)$$

In the retrieval by themes, DQEKS are sentences associated with themes right before extracting named entities. In the retrieval by named entities' network, DQEKS are sentences including the named entities of both source and target nodes. Also, we calculate recall and precision of the intersection and the union of DQEKS which are the output of both the query expansions. Table 3 shows the results. The precision of query expansion by themes is higher than the precision of query expansion by named entities' network. In terms of recall, we can not say which one is higher because of the difference between the two questions. However, the recall of the union adds up to 1.000. Therefore, they are complementary to each other. In the result of the intersection, the precision is higher than both the precisions of the single results. This may be able to improve the precision of the retrieval from AKS.

9 Conclusion and Future Work

In this paper, with regards to a question-answering system to generate an answer of a world history's complex essay type question of the University of Tokyo's entrance examination, we proposed a query expansion method which uses the book of histories by country. The proposed method has two steps: to uncover the phrases which are the main points of the answer by using the book of histories by country, and subsequently to retrieve the corresponding descriptions from AKS. As the result of the experiment with two questions of the University of Tokyo's entrance examination, the recall increased distinctly, while the precision decreased instead. Therefore, there was an improvement from the viewpoint of summarization's preprocessing. In the future we want to make the recall increase to 100%, while still considering the precision.

Acknowledgments

This work was partially supported by
JSPS KAKENHI Grant Number JP16K00296.

References

Gnes Erkan and Dragomir R. Radev. 2004. *LexRank: Graphbased Lexical Centrality As Salience in Text Summarization.* J. Artificial Intelligence Research 22, pp. 457–479.

Jun-Ping Ng, Praveen Bysani, Ziheng Lin, Min − Yen Kan, Chew − Lim Tan. 2012. *Exploiting Category-Specific Information for Multi-Document Summarization.* In Proceedings of COLING 2012, the 24th International Conference on Computational Linguistics: Technical Papers, pp. 2093–2108.

Ji Donghong and Nie Yu. 2008. *Sentence Ordering based on Cluster Adjacency in Multi-Document Summarization.* In Proceedings of IJCNLP 2008, the 3rd International Joint Conference on Natural Language Processing.

Daniel Cohen, Liu Yang and W. Bruce Croft. 2018. *WikiPassageQA: A Benchmark Collection for Research on Non-factoid Answer Passage Retrieval* In Proceedings of SIGIR'18, the 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, pp. 1165–1168.

Eugene Agichtein, David Carmel, Dan Pelleg, Yuval Pinter, and Donna Harman. 2015. *Overview of the trec 2015 liveqa track.* In Proceedings of TREC'15, the twenty-fourth Text REtrieval Conference. Available via https://trec.nist.gov/pubs/trec24/trec2015.html.

Teruko Mitamura, Hideki Shima, Tetsuya Sakai, Noriko Kando, Tatsunori Mori, Koichi Takeda, Chin-Yew Lin, Ruihua Song, Chuan-Jie Lin, and Cheng-Wei Lee. 2010. *Overview of the NTCIR-8 ACLIA Tasks: Advanced Cross-Lingual Information Access.* In Proceedings of NTCIR-8 Workshop Meeting, pp. 15–24.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Sch¨ utze. 2008. *Introduction to Information Retrieval.* Cambridge University Press, 1 edition, July 2008. ISBN 0521865719.

D. Wollersheim and J. W. Rahayu. 2005 Ontology based query expansion framework for use in medical information systems. International Journal of Web Information Systems, 1(2), ISSN 1744-0084.

Hideyuki Shibuki, Kotaro Sakamoto, Madoka Ishioroshi, Akira Fujita, Yoshinobu Kano, Teruko Mitamura, Tatsunori Mori, and Noriko Kando. 2017. *Overview of the NTCIR-13 QA Lab-3 Task*. In Proceedings of the 13th NTCIR Conference on Evaluation of Information Access Technologies, pp. 112–128.

Hideyuki Shibuki, Kotaro Sakamoto, Madoka Ishioroshi, Akira Fujita, Yoshinobu Kano, Teruko Mitamura, Tatsunori Mori, and Noriko Kando. 2016. *Overview of the NTCIR-12 QA Lab-2 Task.* In Proceedings of the 12th NTCIR Conference on Evaluation of Information Access Technologies, pp. 392–408.

Hideyuki Shibuki, Kotaro Sakamoto, Yoshinobu Kano, Teruko Mitamura, Madoka Ishioroshi, Kelly Y. Itakura, Di Wang, Tatsunori Mori, and Noriko Kando. 2014. *Overview of the NTCIR-11 QA-Lab Task.* In Proceedings of the 11th NTCIR Conference on Evaluation of Information Access Technologies, pp. 518–529.

Akira Fujita, Akihiro Kameda, Ai Kawazoe and Yusuke Miyao. 2014 *Overview of Todai Robot Project and Evaluation Framework of its NLP-based Problem Solving* In Proceedings of LREC2014, the 9th edition of the Language Resources and Evaluation Conference, pp. 2590-2597

Kotaro Sakamoto, Takaaki Matsumoto, Madoka Ishioroshi, Hideyuki Shibuki, Tatsunori Mori, Noriko Kando, and Teruko Mitamura. 2017. *FelisCatusZero: A world history essay question answering for the University of Tokyo ' s entrance exam.* In Proceedings of Open Knowledge Base and Question Answering Workshop at SIGIR (OKBQA 2017), pp. 45–46.

Yuta Fukuhara, Kotaro Sakamoto, Hideyuki Shibuki and Tatsunori Mori. 2017. *A study of an annotation to trace between an essay type question of world history and the answer ("Mondai ni okeru mohankaitouchishikigen no taiou wo arawasu anoteishon no kentou" in Japanese).* In Proceedings of twenty-third Annual Meeting of the Association for Natural Language Processing (domestic conference in Japan).

Mitsugi Sato. 2016. *The University of Tokyo's entrance examinations of world history, 25 years [The fifth version] ("Todai no sekaishi 25 ka nen [dai 5 ban]" in Japanese).* Kyogakusha, ISBN-10: 4325206175, ISBN-13: 978-4325206170.

Hiroshi Imaizumi, Iwao Kariyazono and Motoko Kondo. 2017. *Easy to understand the courses: sorting out the world history B by country or area ("Nagare ga wakaru kakkokubetu chiikibetsu sekaishi B no seiri" in Japanese).* Yamakawa Shuppansha Ltd, ISBN-10: 4634030314, ISBN-13: 978-4634030312.

Ai Kawazoe, Yusuke Miyao, Takuya Matsuzaki, Hikaru Yokono and Noriko Arai. 2014. *World History Ontology for Reasoning Truth/Falsehood of Sentences: Event Classification to Fill in the Gaps Between Knowledge Resources and Natural Language Texts.* Nakano, Yukiko, Satoh, Ken, Bekki, Daisuke (Eds.), New Frontiers in Artificial Intelligence (JSAI-isAI 2013 Workshops), Lecture Notes in Computer Science 8417 42-50.

Over-Sampling Methods for Polarity Classification of Imbalanced Microblog Texts

Kiyoaki Shirai **Yunmin Xiang**
Japan Advanced Institute of Science and Technology
1-1, Asahidai, Nomi, Ishikawa, 923-1292, Japan
{kshirai,s1710250}@jaist.ac.jp

Abstract

Polarity classification is the task of classifying sentiments or opinions shown in a given text into positive, negative or neutral. Most previous studies have developed and evaluated their methods on balanced datasets. However, in Twitter, the polarity distribution is highly imbalanced since most tweets are neutral. This paper proposes novel methods to train an accurate classifier from imbalanced data for polarity classification of tweets. They are kinds of synthesizing over-sampling methods that newly generate minority samples to balance the polarity distribution. In our approach, since sentiment words are effective features, minority samples are synthesized more from a sample that includes sentiment words. Furthermore, the number of synthesized minority samples is carefully determined by measuring the performance on development data. According to our experiments using an imbalanced dataset of tweets, the F1-measure of the polarity classification is much improved when our proposed methods are combined with two existing over-sampling methods SMOTE and ADASYN.

1 Introduction

Sentiment analysis is process of analyzing the emotions or opinions in texts. Polarity classification is one of the fundamental techniques in sentiment analysis. It is the task of classifying a given text into polarity classes, such as positive, negative or neutral. In particular, the polarity classification of texts in a microblog such as Twitter received much research attention. Since users actively express their opinions on social media, microblog texts are valuable resources for sentiment analysis and opinion mining.

Supervised machine learning is a major approach for polarity classification. In past studies, polarity classifiers have usually been trained and evaluated on balanced datasets, i.e. those in which the number of samples of each polarity class is almost the same. However, in real social media, the distribution of the polarity of texts is actually imbalanced, since there are many more neutral samples than positive or negative ones. Machine learning usually performs poorly on imbalanced data, since a classifier tends to judge a sample of a minority class as belonging to a majority class. On the other hand, the detection of minority samples (i.e. positive and negative samples) is important because they provide useful information in sentiment analysis.

This paper proposes several methods to train an accurate classifier to determine the polarity of texts in Twitter from an imbalanced dataset. Our methods are an extension of existing over-sampling methods. Over-sampling is a technique to increase the number of minority samples artificially to make a balanced dataset. In our approach, sentiment words are taken into account in the generation of the minority samples, since sentiment words are obviously useful features for polarity classification.

2 Related work

2.1 Sentiment analysis

Supervised machine learning has been widely applied to sentiment analysis and polarity classification. Pang et al. (2002) used Naive Bayes, Maximum Entropy and Support Vector Machine (SVM) to classify the polarity of a given movie review. In their experiment on the classification of positive or negative classes, SVM outperformed Naive Bayes and Maximum Entropy. The accuracy was relatively high, 82.9%. However, the classifiers were trained

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 228-236, Hakodate, Japan, September 13-15, 2019

and evaluated on a balanced dataset consisting of 700 positive and 700 negative movie reviews.

Early work on polarity classification of three classes (positive, negative or neutral) has been done by Koppel and Schler (2006). They claimed that a precise three-class classifier could not be trained from positive and negative samples, since the neutral samples would not simply be located at somewhere near a boundary between the positive and negative classes. Thus neutral samples were necessary for training. A stack of three kinds of binary classifiers (positive vs negative, positive vs neutral, and negative vs neutral) achieved 74.1% accuracy on a TV domain and 85.5% on a shopping domain. The datasets of both domains were completely balanced.

Recently, deep neural networks have been introduced to polarity classification. CharSCNN (Character to Sentence Convolutional Neural Network) employed two convolutional neural layers (Dos Santos and Gatti, 2014). One was to obtain abstract representations of the words from character embedding, which enabled the model to use character-level features. The other was to obtain abstract representations of sentences from word vectors, which were concatenations of word embedding including word-level features and the output of the first network including character-level features. Finally, two feed forward networks were used to obtain a score for each polarity class. CharSCNN achieved 85.7% accuracy on the Stanford Sentiment Treebank (Socher et al., 2013) and 86.4% on the Stanford Twitter Sentiment corpus (Go et al., 2009). The distributions of polarity classes in these two datasets were balanced.

Similar to the above papers, most of the past studies of polarity classification evaluated methods using balanced data that consists of almost equal numbers of samples for all polarity classes. However, as we will report in Section 3, in social media, the number of neutral texts is much greater than the numbers of positive and negative texts. Thus the distribution of the polarity is higly skewed. The datasets of SemEval 2016 task 4 (Nakov et al., 2016) and SemEval 2017 task 4 (Rosenthal et al., 2017) are widely used for research on sentiment analysis in Twitter. However, neutral samples are not overwhelmingly dominant in these datasets. Considering a real application to a microblog, this paper focuses on polarity classification in imbalanced data where there are many more neutral samples than the others.

2.2 Over-sampling methods

Over-sampling and under-sampling are commonly used to improve the performance of supervised machine learning on an imbalanced dataset. The core idea of these methods is to increase the number of minority samples or to decrease the number of majority samples so that the number of samples of each class becomes balanced. This research has focused on the over-sampling approach.

Chawla et al. (2002) proposed Synthetic Minority Oversampling TEchnique (SMOTE), an over-sampling method that synthesizes new minority samples from existing ones. Figure 1 shows its pseudocode. Let us suppose that an imbalanced dataset consists of a large number of majority samples and a small number of minority samples, and each sample is represented as a feature vector in a vector space. For each minority sample $\vec{x_i}$, its k nearest neighbours of minority samples are chosen at line 7. Then, one minority sample $\vec{n}$ is chosen randomly at line 9. A random point somewhere on the line between $\vec{x_i}$ and $\vec{n}$ is chosen as indicated in lines 10-12. It is the newly synthesized minority sample $\overrightarrow{syn}$ and is added to the dataset at line 13. bal is a balance parameter to control the number of synthesized samples. It is defined as the proportion of the minority samples to the majority samples in the new (over-sampled) dataset. For example, $bal = 1$ means that the new training data contains equal numbers of majority and minority samples, whereas $bal = 0.5$ means that number of minority samples becomes 50% of the number of majority samples. g_{all} at line 3 denotes the total number of samples to be synthesized, while g at line 4 denotes the number of samples to be synthesized from one minority sample. The synthesis of the minority samples from $\vec{x_i}$ is repeated g times, as indicated at line 8.

ADAptive SYNthetic sampling (ADASYN) (He et al., 2008) is another over-sampling method. The key idea is to synthesize more samples from minority samples that are located near a borderline between minority and majority classes. It can enable a trained classifier to more easily discriminate between minority and majority samples. Figure 2 shows the pseudocode of ADASYN. At line 6, $r[i]$ is the ratio of the majority samples in the k nearest

Input: X(original training data), bal(balance parameter), k(number of nearest neighbours)
Output: X' (new training data)

1: $S_{min} \leftarrow$ a set of minority samples in X
2: $S_{maj} \leftarrow$ a set of majority samples in X
3: $g_{all} \leftarrow |S_{maj}| \times bal - |S_{min}|$
4: $g \leftarrow int(g_{all}/|S_{min}|)$
5: $Syn \leftarrow \phi$
/* a set of synthesized minority samples */
6: **for each** $\vec{x_i} \in S_{min}$ **do**
7: $K_i \leftarrow k$ nearest neighbours of $\vec{x_i}$ in S_{min}
8: **for** $j = 1$ to g **do**
9: $\vec{n} \leftarrow$ a sample randomly chosen from K_i
10: $\overrightarrow{diff} \leftarrow \vec{n} - \vec{x_i}$
11: $gap \leftarrow$ random value between $[0, 1]$
12: $\overrightarrow{syn} \leftarrow \vec{x_i} + gap \times \overrightarrow{diff}$
13: $Syn \leftarrow Syn \cup \{\overrightarrow{syn}\}$
14: **end for**
15: **end for**
16: **return** $X' = X \cup Syn$

Figure 1: Pseudocode of SMOTE

neighbours of a minority sample $\vec{x_i}$. It evaluates how likely $\vec{x_i}$ is to be close to the borderline. It is normalized in line 9 to calculate the density distribution $\hat{r}$, then $g[i]$, the number of samples to be synthesized from $\vec{x_i}$, is calculated in line 10. The following procedures are almost the same as SMOTE. An important difference between SMOTE and ADASYN is that equal numbers of synthetic samples are generated for each minority sample in SMOTE, whereas in ADASYN, more samples are generated from minority samples near a borderline.

This paper extends SMOTE and ADASYN to improve the accuracy of polarity classification of imbalanced data.

3 Survey of the polarity distribution in Twitter

A preliminary survey was conducted to briefly investigate the distribution of the polarity of texts in Twitter. It is expected that neutral tweets are the overwhelming majority in the real world and thus polarity distribution is highly imbalanced.

Tweets are collected by searching a keyword with Twitter API. Eight topics including electronic products, celebrities, movies and so on, are chosen as keywords, which are shown in Table 1. One hundred tweets are retrieved for each topic. Thus 800 tweets are retrieved in total. Note that advertisement tweets are excluded. These tweets are manually classified in terms of their polarity toward a topic.

Table 1 shows the number of positive, negative and neutral tweets about 8 topics as well as the total numbers. It shows that the ratio of neutral tweets is quite high, 86%. It is found that users usually write a fact or statement about a topic and do not express their emotions or opinions.

Input: X(original training data), bal(balance parameter), k(number of nearest neighbours)
Output: X' (new training data)

1: $S_{min} \leftarrow$ a set of minority samples in X
2: $S_{maj} \leftarrow$ a set of majority samples in X
3: $g_{all} \leftarrow |S_{maj}| \times bal - |S_{min}|$
4: **for each** $\vec{x_i} \in S_{min}$ **do**
5: $NN_i \leftarrow k$ nearest neighbours of $\vec{x_i}$ in X
6: $r[i] \leftarrow \frac{|NN_i \cap S_{maj}|}{k}$
7: **end for**
8: **for each** $\vec{x_i} \in S_{min}$ **do**
9: $\hat{r}[i] \leftarrow \frac{r[i]}{\sum_i r[i]}$
10: $g[i] \leftarrow int(\hat{r}[i] \times g_{all})$
11: **end for**
12: $Syn \leftarrow \phi$
13: **for each** $\vec{x_i} \in S_{min}$ **do**
14: $K_i \leftarrow k$ nearest neighbours of $\vec{x_i}$ in S_{min}
15: **for** $j = 1$ to $g[i]$ **do**
16: $\vec{n} \leftarrow$ a sample randomly chosen from K_i
17: $\overrightarrow{diff} \leftarrow \vec{n} - \vec{x_i}$
18: $gap \leftarrow$ random value between $[0, 1]$
19: $\overrightarrow{syn} \leftarrow \vec{x_i} + gap \times \overrightarrow{diff}$
20: $Syn \leftarrow Syn \cup \{\overrightarrow{syn}\}$
21: **end for**
22: **end for**
23: **return** $X' = X \cup Syn$

Figure 2: Pseudocode of ADASYN

4 Proposed method

This section first explains how to train a classifier for polarity classification, then describes our proposed over-sampling methods.

Table 1: Distribution of classes for each topic

	pos.	neg.	neu.
iPhone X	20	12	68
HUAWEI	10	7	83
SAMSUNG	3	1	96
Morgan Freeman	3	2	95
Gabe Newell	5	3	92
Star Wars	6	3	91
Harry Potter	7	4	89
Monster Hunter : World	11	2	87
Total	65	34	701

4.1 Polarity classifier

Each tweet is represented as a feature vector as follows. After preprocessing, including conversion from upper to lower case, removal of stopwords, and replacement of URL and @+user_id with special tokens, the vector of a tweet is obtained by Equation (1).

$$\text{tweet vector} = \frac{1}{\sum_i w_i^2} \sum_i w_i \times \vec{v_i} \qquad (1)$$

where $\vec{v_i}$ is the vector representation of the ith word, and w_i is the weight of TF-IDF for the ith word. Word vectors are word embedding, which was pretrained by a skip-gram model (Mikolov et al., 2013) from the English Wikipedia corpus. The dimension of the word embedding was set to 250.

The SVM was trained using sklearn[1]. The square of the hinge loss function was chosen as the loss function for the training. The penalty parameter C of the error term was set to 0.5. The kernel of the SVM is the linear kernel.

4.2 Quantity Control Over-Sampling (QCO)

In a synthetic over-sampling strategy, minority samples are newly synthesized. The quality of such synthesized samples is questionable, since they are not real samples at all. The more samples are synthesized to balance the distribution of the classes, the more unreliable samples are likely to be added in the dataset. The generation of too many samples may cause a decline in the classification performance. However, in the papers of SMOTE (Chawla et al., 2002) and ADASYN (He et al., 2008), the number of synthesized samples was given by the user, and there was no discussion how to determine it appropriately.

The number of synthesized samples can be empirically determined. More specifically, the balance parameter bal can be optimized using the development data.[2] First, we prepare the training data and development data. We also prepare a set of balance parameters B. Next, for each balance parameter $bal_i \in B$, we train a classifier from the training data balanced by SMOTE or ADASYN, and apply it for the development data. Finally, the optimized balance parameter is chosen so that the F1-measure on the development data becomes the highest.

In this paper, we call this method Quantity Control Over-Sampling (QCO). It is not a novel method as it is common to optimize parameters using the development data. However, we will demonstrate that the optimization of the number of synthesized samples is crucial in the experiment in Section 5.

4.3 Over-sampling methods considering sentiment words

4.3.1 SMOTE with Sentiment Oriented Over-Sampling (SOO)

We propose an over-sampling method that takes sentiment words into account in the synthesis of minority samples. It is well known that sentiment words are important and effective features for polarity classification. We expect that a superior classifier could be trained from a dataset including many sentiment words. The key idea of our method is to generate more samples including sentiment words.

We introduce a sentiment weight parameter, sen. It is defined as the weight of the samples including sentiment words. Minority samples from a minority sample including a sentiment word are synthesized sen times more often than a sample that does not include a sentiment word. Note that sen is supposed to be greater than 1.

Figures 3 and 4 show the pseudocode. $y =$

[1] https://scikit-learn.org/

[2] The number of synthesized samples is chosen in different ways in SMOTE (Chawla et al., 2002) and ADASYN (He et al., 2008). In the pseudocodes in Figure 1 and 2, SMOTE and ADASYN are slightly modified so that the number of synthesized samples is defined in the same way, i.e. controlled by bal. Note that these pseudocodes are completely equivalent to the original algorithms.

Input: X(original training data), bal(balance parameter), k(number of nearest neighbours)
Output: X' (new training data)

1: $S_{min} \leftarrow$ a set of minority samples in X
2: $S_{maj} \leftarrow$ a set of minority samples in X
3: $g_{all} \leftarrow |S_{maj}| \times bal - |S_{min}|$
4: $y \leftarrow \text{SYNTHESISWEIGHTS}(S_{min})$
5: **for each** $\vec{x_i} \in S_{min}$ **do**
6: $\quad \hat{y}[i] \leftarrow \frac{y[i]}{\sum_i y[i]}$
7: $\quad g[i] \leftarrow int(\hat{y}[i] \times g_{all})$
8: **end for**
9: $Syn \leftarrow \phi$
10: **for each** $\vec{x_i} \in S_{min}$ **do**
11: $\quad K_i \leftarrow k$ nearest neighbours of $\vec{x_i}$ in S_{min}
12: $\quad$ **for** $j = 1$ to $g[i]$ **do**
13: $\quad\quad \vec{n} \leftarrow$ a sample randomly chosen from K_i
14: $\quad\quad \overrightarrow{diff} \leftarrow \vec{n} - \vec{x_i}$
15: $\quad\quad gap \leftarrow$ random value between $[0, 1]$
16: $\quad\quad \overrightarrow{syn} \leftarrow \vec{x_i} + gap \times \overrightarrow{diff}$
17: $\quad\quad Syn \leftarrow Syn \cup \{\overrightarrow{syn}\}$
18: $\quad$ **end for**
19: **end for**
20: **return** $X' = X \cup Syn$

Figure 3: Pseudocode of our proposed over-sampling algorithm

$\{y[0], \cdots, y[n]\}$ is a list of synthesis weights. Each $y[i]$ controls how many minority samples are newly synthesized from the ith minority sample $\vec{x_i}$. SYNTHESISWEIGHTS is a function to determine y, which is described in Figure 4. $y[i]$ is set to be sen if t_i (a tweet of the ith minority sample) contains a sentiment word, otherwise 1. SentiWordNet (Baccianella et al., 2010) is used to judge whether a word in a tweet is a sentiment word or not. Note that the algorithm of Figure 3 is the same as SMOTE when $y[i]$ is set to 1 for all i. In SMOTE, all $\vec{x_i}$ receive the same number of synthesized samples, whereas in our method, sen times as many samples are generated from $\vec{x_i}$ with a sentiment word.

After y is determined at line 4 in Figure 3, the procedures are almost the same as ADASYN. $y[i]$ is normalized to be $\hat{y}[i]$ at line 6, similar to $\hat{r}[i]$ in ADASYN in Figure 2. Then $g[i]$ is calculated in line 7. The minority samples are generated $g[i]$ times from $\vec{x_i}$ in lines 12-18.

1: **function** SYNTHESISWEIGHTS(S_{min})
2: $\quad$ **for each** $\vec{x_i} \in S_{min}$ **do**
3: $\quad\quad$ **if** t_i includes a sentiment word **then**
4: $\quad\quad\quad y[i] \leftarrow sen$
5: $\quad\quad$ **else**
6: $\quad\quad\quad y[i] \leftarrow 1$
7: $\quad\quad$ **end if**
8: $\quad$ **end for**
9: $\quad$ **return** y
10: **end function**

Figure 4: SYNTHESISWEIGHTS of SMOTE+SOO

Finally, the sentiment weight parameter sen is optimized on the development data. Hereafter, 'Sentiment Oriented Over-Sampling' (SOO) refers to the proposed technique that synthesizes more samples from samples including sentiment words. SMOTE combined with SOO is referred to as SMOTE+SOO.

4.3.2 ADASYN with Sentiment Oriented Over-Sampling (SOO)

SOO can be combined with ADASYN. The pseudocode of ADASYN+SOO is presented in Figure 3 where the function SYNTHESISWEIGHTS is defined as in Figure 5. The only difference between this algorithm and the original ADASYN is lines 5-9 in Figure 5: $y[i]$ is always set to $r[i]$ in ADASYN, but in ADASYN+SOO, it is multiplied by sen if t_i contains a sentiment word. Thus ADASYN+SOO is able to not only generate more synthetic samples near a borderline but also create more samples including sentiment words. Similar to SMOTE+SOO, the parameter sen is optimized using the development data.

4.3.3 Sentiment Intensity Oriented Over-Sampling (SIOO)

Another extension of ADASYN made in the present paper is ADASYN with Sentiment Intensity Oriented Over-Sampling (SIOO). In SOO, although more samples are generated from a sample including a sentiment word, the number of synthesized samples is the same for all samples with a sentiment word. However, it is supposed that samples showing intense emotion heavily contribute to polarity classification. In SIOO, more samples are generated from

```
1: function SYNTHESISWEIGHTS(S_min)
2:     for each x_i ∈ S_min do
3:         NN_i ← k nearest neighbours of x_i in X
4:         r[i] ← |NN_i ∩ S_maj| / k
5:         if t_i includes a sentiment word then
6:             y[i] ← sen × r[i]
7:         else
8:             y[i] ← r[i]
9:         end if
10:    end for
11:    return y
12: end function
```

Figure 5: SYNTHESISWEIGHTS of ADASYN+SOO

```
1: function SYNTHESISWEIGHTS(S_min)
2:     for each x_i ∈ S_min do
3:         NN_i ← k nearest neighbours of x_i in X
4:         r[i] ← |NN_i ∩ S_maj| / k
5:         if t_i includes a sentiment word then
6:             s[i] = (Σ_{w_i ∈ SW(t_i)} score(w_i)) / |SW(t_i)|
7:             y[i] ← s[i] × r[i]
8:         else
9:             y[i] ← r[i]
10:        end if
11:    end for
12:    return y
13: end function
```

Figure 6: SYNTHESISWEIGHTS of ADASYN+SIOO

a minority sample that expresses strong sentiment.

The sentiment intensity score $s[i]$ of the ith tweet t_i is defined by

$$s[i] = \frac{\sum_{w_i \in SW(t_i)} score(w_i)}{|SW(t_i)|} \quad (2)$$

$$score(w_i) = max(s_{pos}(w_i), s_{neg}(w_i)) + 1 \quad (3)$$

where $SW(t_i)$ is the set of sentiment words in t_i. $score(w_i)$ is the sentiment score of w_i defined by Equation (3), and s_{pos} and s_{neg} are the averages of the positive and negative scores of w_i in SentiWordNet.[3] Thus $s[i]$ evaluates the intensity of the sentiment of the ith sample regardless of its polarity orientation.

The pseudocode of ADASYN+SIOO is presented in Figure 3, where the function SYNTHESISWEIGHTS is defined as in Figure 6. The only difference between SOO and SIOO is that the sentiment weight parameter sen is replaced with $s[i]$ in SIOO, as indicated in lines 6 and 7. Note that $s[i]$ should be greater than one to give importance to samples with polarity words in the generation of the minority samples. That is the reason why we add 1 in Equation (3). Note that the positive and negative scores in SentiWordNet are between 0 and 1, so $s[i]$ can be less than 1 if we do not add 1 in $score(w_i)$.

Another advantage of SIOO is its lesser computational cost. SOO requires trial and error in its training and applying classifiers for the optimization of the parameter sen, but SIOO can easily calculate $s[i]$ using the sentiment lexicon.

[3] In SentiWordNet, positive and negative scores are given for each sense of a word. Therefore, polysemous words have several positive and negative scores. We take their average.

5 Evaluation

5.1 Experimental setting

The SemEval 2017 task 4 (Rosenthal et al., 2017) dataset was used for the experiment. It is a collection of tweets about several topics with manually annotated polarity labels. The polarity of each tweet is represented by 5-point labels from 1 (very negative) to 5 (very positive). In this experiment, we defined three polarity classes by converting 1 and 2 to "negative", 3 to "neutral" and 4 and 5 to "positive". Our preliminary survey showed that in Twitter, 86% of tweets are neutral. To make the distribution of the polarity labels of the dataset closer to the actual distribution, we added neutral tweets to the dataset by the following procedures.

1. Retrieve tweets via Twitter API by searching the keywords of the topics in the SemEval 2017 dataset.

2. Classify the retrieved tweets by AYLIEN[4], which is a web toolkit for polarity classification. Only tweets classified as neutral are kept, the other are discarded.

3. Add neutral tweets to the dataset until the proportion of neutral tweets reaches 86%.

[4] https://aylien.com/text-api/sentiment-analysis/

Table 2: Statistics of training, development and test data

	positive	negative	neutral
training	5,748	1,637	46,534
development	1,642	468	13,298
test	822	234	6,649
total	8,212 (11%)	2,339 (3%)	66,491 (86%)

Table 3: Optimized balance parameter bal

	positive	negative
SMOTE+QCO	0.6	0.4
ADASYN+QCO	0.6	0.5

Finally, the dataset was divided into 70% training data, 20% development data, and 10% test data. The numbers of samples in the three classes are shown in Table 2.

Two binary classifiers have been evaluated. The first classifier judges whether a tweet is positive or not. To train and evaluate it, the negative and neutral tweets were merged into "not positive" tweets as majority samples, while the positive tweets remained as minority samples. The second classifier judges whether a tweet is negative or not. Similarly, the positive and neutral tweets were merged into "not negative" tweets as majority samples. Precision, recall, and F1-measure have been used as the evaluation criteria. Throughout these experiments, the parameter of the number of nearest neighbour k was set to 7.

5.2 Results of QCO

The balance parameter bal was changed from 0.2 to 1 in steps of 0.1. Figure 7 shows the F1-measure of the positive and negative classification on the development data by SMOTE+QCO and ADASYN+QCO with different bal. It is confirmed that the F1-measure drastically changes with bal. This indicates that the optimization of the number of synthesized samples is important. The optimized parameters bal are summarized in Table 3.

Next, the performance of the methods with QCO as well as the baselines was measured on the test data. Table 4 presents the precision (P), recall (R), and F1-measure (F) of the positive and negative classification on the test data. SMOTE and ADASYN are the original algorithm with $bal = 1$. The baseline is a classifier trained from the original imbalanced dataset. All over-sampling methods outperform the baseline. Comparing SMOTE+QCO or ADASYN+QCO with methods without QCO, QCO greatly improves the precision with a little deterioration of the recall. The F1-measures of SMOTE+QCO and ADASYN+QCO are better than those of SMOTE and ADASYN in both positive and negative classification, respectively.

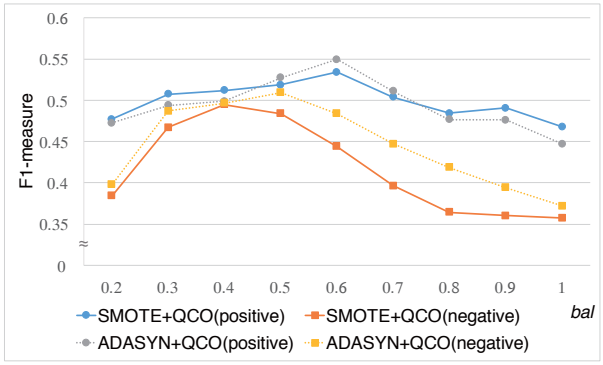


Figure 7: F1-measure of SMOTE+QCO and ADASYN+QCO on development data

Table 4: Results of methods with QCO on test data

	positive			negative		
	P	R	F	P	R	F
Baseline	.3064	.7037	.4271	.1886	.7012	.2973
SMOTE	.3437	.7155	.4652	.2529	.7170	.3739
SMOTE+QCO	.4297	.7098	.5279	.3763	.6917	.4874
ADASYN	.3136	.7335	.4374	.2788	.7315	.4037
ADASYN+QCO	.4461	.6778	.5378	.4003	.7016	.5144

5.3 Evaluation of SOO and SIOO

We conducted experiments to evaluate the proposed over-sampling methods considering sentiment words. Thoughout these experiments, bal was set to the optimized value in Table 3. As for SOO, the sentiment weight parameter sen was changed from 1 to 7. Figure 8 shows the F1-measures of positive and negative classification on the development data by SMOTE+SOO and ADASYN+SOO with different sen. When sen is greater than 1, i.e. more minority samples are synthesized from a sample including sentiment words, the F1-measure is improved. However, the performance deteriorates when sen becomes too large. The optimized parameters sen

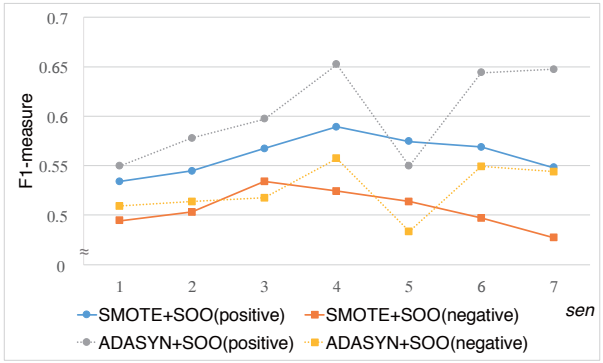


Figure 8: F1-measure of SMOTE+SOO and ADASYN+SOO on development data

Table 5: Optimized sentiment weight parameter sen

	positive	negative
SMOTE+SOO	4	3
ADASYN+SOO	4	4

are summarized in Table 5.

Table 6 presents the results of methods with SOO on the test data. Comparing SMOTE+SOO or ADASYN+SOO with SMOTE+QCO or ADASYN+QCO, the former outperforms the latter for all criteria except for the recall of negative classification by SMOTE. This proves that our method, which puts more weight on samples including sentiment words in the synthesis of the minority samples, is effective at improving the performance of the polarity classification.

We now present the evaluation of ADASYN+SIOO. The results of ADASYN+SIOO are shown in the last row of Table 6. Contrary to our expectations, ADASYN+SIOO is worse than ADASYN+SOO. This indicates that to determine the sentiment weight parameter by the sentiment scores of the words in a tweet is not as good as empirical optimization using the development data.

In the experiments as a whole, ADASYN mostly outperforms SMOTE. ADASYN+SOO achieves the best F1-measure, 0.65 and 0.55 for the positive and negative classification, respectively.

6 Conclusion

The contributions of this paper are summarized in what follows. First, the effectivness of the Quantity Control Over-Sampling (QCO) was empirically

Table 6: Results of methods with SOO and SIOO on test data

	positive			negative		
	P	R	F	P	R	F
SMOTE+QCO	.4297	.7098	.5279	.3763	.6917	.4874
SMOTE+SOO	.4752	.7421	.5794	.4466	.6851	.5407
ADASYN+QCO	.4461	.6778	.5378	.4003	.7016	.5144
ADASYN+SOO	.6037	.7047	.6503	.4314	.7559	.5493
ADASYN+SIOO	.5676	.7255	.6369	.4096	.7443	.5284

investigated. It was found that QCO could improve the F1-measure drastically. It indicates that the optimization of the number of synthesized minority samples is quite important. QCO is general and applicable to any classification task on imbalanced data. Second, we proposed the Sentiment Oriented Over-Sampling (SOO) method that synthesizes more minority samples containing sentiment words. SOO is a method only for polarity classification, but could be combined with any supervised machine learning algorithm. We also proposed the Sentiment Intensity Oriented Over-Sampling (SIOO) method that considers the intensity of the sentiment in the generation of the minority samples. The results of experiments showed that SOO could greatly improve the classification performance of SMOTE and ADASYN. On the other hand, the effectiveness of SIOO was not confirmed by the experiments in this paper.

In the future, in order to improve SIOO, the way to measure the intensity of the sentiment in tweets will be carefully investigated. For example, a combination of SOO and SIOO is worth assessing. The combination of SIOO with SMOTE (i.e. SMOTE+SIOO) should also be evaluated. In addition, since only the extended SemEval 2017 dataset was used in the experiments, our proposed methods should be applied to other microblog datasets for more precise evaluation. Another important line of future research is to extend our method to multiclass classification, since the current methods are only applicable to binary classification. This would enable us to classify a tweet into positive, negative, or neutral by a single system.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. SentiWordNet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, pages 2200–2204.

Nitesh V. Chawla, Kevin W. Bowyer, Lawrence O. Hall, and W. Philip Kegelmeyer. 2002. SMOTE: Synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Cicero Dos Santos and Maira Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of the 25th International Conference on Computational Linguistics: Technical Papers*, pages 69–78.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. Technical report, Stanford University.

Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning. In *Proceedings of 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence)*, pages 1322–1328.

Moshe Koppel and Jonathan Schler. 2006. The importance of neutral examples for learning sentiment. *Computational Intelligence*, 22(2):100–109.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 3111–3119.

Preslav Nakov, Alan Ritter, Sara Rosenthal, Fabrizio Sebastiani, and Veselin Stoyanov. 2016. SemEval-2016 task 4: Sentiment analysis in Twitter. In *Proceedings of the 10th International Workshop on Semantic Evaluation*, pages 1–18. Association for Computational Linguistics.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 79–86. Association for Computational Linguistics.

Sara Rosenthal, Noura Farra, and Preslav Nakov. 2017. SemEval-2017 task 4: Sentiment analysis in Twitter. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, pages 502–518. Association for Computational Linguistics.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642.

Thai Learners of English are Sensitive to Number-Agreement Violations

Teeranoot Siriwittayakorn
English Department, Faculty of Humanities
Chiang Mai University, 239 Huay Kaew Rd.,
Muang, Chiang Mai 50200 Thailand
teeranoot.s@cmu.ac.th

Edson T. Miyamoto
Center for Meta-Learning
Future University Hakodate, 116-2 Kameda
Nakano-cho, Hakodate, Japan
miyamoto@alum.mit.edu

Abstract

We report a reading-time experiment investigating how native Thai speakers process sentences with subject-verb number agreement in English as a second language. Participants were slower to read sentences containing agreement violations, in a manner similar to what has been reported for native English speakers. The results add to a growing literature according to which learners can acquire knowledge of number agreement even if their native language lacks it. This suggests that learners are not constrained by the features available in their native languages, and are able to acquire new features and put this knowledge to use when reading sentences for content.

1 Introduction

Following work on programming languages (Aho et al., 2007; and references therein), human sentence comprehension, or *parsing* broadly speaking, is often assumed to involve two components: a knowledge base (the grammar) and an algorithm that uses the knowledge base to process sentences (the parser). Moreover, it is commonly assumed that there is a single parser for all human languages; therefore, a child only needs to learn the grammar to be able to process sentences in a language (see Fodor, 1998, for detailed discussion, and on the impossibility for a child to learn the grammar and a language-specific parser at the same time).

A natural extension is that adults learning a second language (L2) only need to learn the grammar of the new language. The algorithm to use that knowledge is assumed to be the same as the parser for the learners' native language (L1). Therefore, behavioral differences between native speakers and L2 learners when processing sentences should be the result of differences in the knowledge base.

Within this framework, it is usually not enough to show that learners know some features of the L2 grammar. For example, in a traditional L2 task, we could ask learners of English to judge whether a sentence (e.g., *The key<u>s</u> <u>is</u> near the pencil*) is grammatical to determine whether they know that the subject and the verb have to agree in number in English. In this paper, we assume that most college students who have studied English know the basic rules of its number agreement system.

The more crucial question in this framework is to determine whether learners have acquired that knowledge and incorporated it to their L2 grammar, so that they can rapidly access it to process sentences in a manner that approaches L1 speakers' behavior. For this type of question, we can collect reading times to determine if learners slow down in situations in which native English speakers are known to be slow. For example, native English speakers are slow to read number-agreement violations as in (1) (Wagers et al., 2009; and references therein).

(1) The key to the cabinet are on the table.

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 237-243, Hakodate, Japan, September 13-15, 2019

We report data on native Thai speakers to investigate how they process number agreement when reading L2 English. Thai does not have number markers or subject-verb agreement in general (see Iwasaki and Ingkaphirom, 2009, on Thai grammar). According to some early studies, learners are unable to keep track of subject-verb agreement in L2 when reading for comprehension if their L1 does not have that kind of relation (Hawkins and Chan, 1997; Jian, 2004; *inter alia*). Contrary to those claims, we suggest that speakers whose L1 does not have number agreement can display sensitivity to number-agreement violations in L2, extending previous results (Wen et al., 2010; Wilson and Miyamoto, 2015; *inter alia*).

2 Previous studies

According to previous literature, learners cannot acquire and rapidly manipulate features of L2 that are not available in their L1 (Hawkins and Chan, 1997; *inter alia*). For example, native speakers of languages that have number agreement (e.g., Russian) have been shown to be slow to read agreement violations in L2 English, similar to what has been reported for native English speakers; whereas native speakers of languages that do not have number agreement (e.g., Chinese, Japanese) do not show sensitivity to such violations (Jiang, 2004; Jiang et al., 2011).

However, there is an alternative way of interpreting these results. If we assume that L2 learning involves modifying L1 knowledge to approximate L2 (Schwartz and Sprouse, 1996), it may take longer for learners to acquire number agreement in L2 when their L1 lacks such feature. Moreover, the effects of individual variation (e.g., learners' proficiency) on language acquisition may be greater when learners acquire a feature from scratch, adding noise to experimental results. Therefore, it is conceivable that some previous studies (Jiang, 2004; Jiang et al., 2011; *inter alia*) although carefully conducted, failed to detect sensitivity to agreement violations because they did not take proficiency into consideration during the analyses.

In studies in which proficiency was included as a factor in the analyses, native speakers of Chinese and Japanese were shown to be sensitive to agreement violations while reading L2 English. One study used simple constructions involving agreement inside noun phrases (Wen et al., 2010; also, Yamada and Hirose, 2012, for data on more complex constructions). Another study used constructions as in (1) in which a prepositional phrase (*to the cabinet*) intervenes between the head noun and verb (Wilson and Miyamoto, 2015).

Given those previous results indicating the influence of proficiency, we included English proficiency scores when analyzing the Thai speakers' reading time data.

3 Experiment

3.1 Participants

Thirty-three native Thai speakers, undergraduate students at Chiang Mai University, volunteered to participate in the experiment. One participant's data were excluded because the participant did not follow the instructions. Results for the remaining 32 participants are reported.

All participants started learning English at the age of six or later, had never lived abroad for six months or longer, and were all majoring in English. Previous studies that did not detect sensitivity to agreement violations (Jiang, 2004; Jiang et al., 2011; *inter alia*) recruited L2 learners living in the United States. It is unlikely that English majors living in Thailand had more exposure to English than learners living in the United States, but this possibility is being addressed in on-going work.

3.2 Method

Stimuli: There were 16 pairs of test items, in which grammaticality was manipulated by modifying the number of the head noun in subject position, so that the head noun was plural in the grammatical condition and singular in the ungrammatical condition. The verb was always *were* (see (2) for an example pair; all stimuli were from Wilson and Miyamoto, 2015, with mistakes such as spelling corrected).

(2)

(a) Grammatical condition
The chickens in the oven were completely burned.

(b) Ungrammatical condition
The chicken in the oven were completely burned.

If learners are sensitive to number-agreement violations, they should be slow to read the verb *were* (or the word immediately thereafter) in the ungrammatical condition compared to the same word in the grammatical condition, as has been reported for native English speakers (Wagers et al., 2009, and references therein).

There are reasons to predict that no such a difference would be observed. First, because Thai does not have number agreement, native Thai speakers may not be able to acquire number agreement in L2 English (Hawkins and Chan, 1997) as has been reported for Chinese and Japanese speakers reading L2 English (Jiang, 2004; Jiang et al., 2011). Second, the intervening prepositional phrase (PP; e.g., *in the oven*) may make it too difficult for learners to keep track of the agreement relation between head noun and verb. In particular, learners may be unable to build the hierarchical structure in which the PP modifies the head noun, and instead build a shallower structure in which *oven* is the sister of *chicken(s)* (such a simplified structure would be compatible with Clahsen and Felser, 2006). In this case the verb may be associated with *oven*, instead of *chicken(s)*, therefore making both conditions equally acceptable.

However, in the same way as native Japanese speakers reading L2 English in a more recent study (Wilson and Miyamoto, 2015), it is conceivable that native Thai speakers are sensitive to number-agreement violations as well. In which case, the ungrammatical condition in (1b) should be read more slowly at the verb or later. This would suggest that Thai speakers acquire knowledge of number morphology and are able to use it in a manner that resembles native English speakers.

There were 48 filler sentences and 32 sentences from another experiment whose structure was similar to the test items to distract participants' attention away from the point of the experiment. All of these sentences were grammatical.

Procedure: Doug Rohde's Linger program was used to present sentences in a word-by-word non-cumulative self-paced reading procedure. The critical region (the verb *were*) was always region 6. Each participant saw eight grammatical sentences and eight ungrammatical sentences, and only one version of each pair of items. The test items were interspersed with 48 fillers and 32 items from another experiment in pseudo-random order so that two test items did not follow in succession. Each sentence was followed by a yes/no comprehension question. Feedback was provided when participants' answer was incorrect.

After the reading-time experiment participants answered a c-test questionnaire, in which they had to complete the second half of every other word in five texts (from Babaii and Shahri, 2010). Such questionnaires have been used in the past as an effective measure of proficiency in the analyses of reading times (Wen et al., 2010; Wilson and Miyamoto, 2015), and the scores have been reported to correlate well with more traditional measures such as the TOEFL-ITP (Wilson and Miyamoto, 2015).

Analysis: Analyses were performed on R version 3.5.0 (R Core Team, 2018). Only reading times from trials for which the comprehension question was answered correctly were included in the analyses. Initial trimming eliminated reading times below 100 ms and those above 5000 ms as they were unlikely to reflect reading-related latencies (Baayen, 2008, pp. 243-244, for discussion).

First, we report results from analysis of variance (ANOVA) using untransformed reading times to provide a comparison with previous studies (e.g., Jiang, 2004). Moreover, like in these earlier studies we did not include learners' proficiency as a factor in this initial analysis. Similar trends were observed when log-transformed reading times were used.

Second, we report results from mixed-effects models using log-transformed reading times and including proficiency (i.e., c-test scores) as a factor. Log-transformed reading times are usually used to decrease the influence of extreme values, and are appropriate for learners in this experiment as there may be some extremely long reading times (e.g., for unknown words). Similar trends were observed with untransformed reading times. After the initial 100 - 5000 ms trimming step, model-based trimming was conducted to eliminate data points beyond three standard-deviations, and the model was refit with the remaining data (Baayen, 2008; pp. 243-244). For each region, the trimming procedure eliminated no more than 3% of the data.

For all models, by-participants and by-items random intercepts were included. Whether a term was included as by-participants or by-items random slope was determined through backward selection (Bates et al., 2015). Numerical factors were centered to facilitate interpretation and improve convergence of the models.

Results not reported were not reliable (*p*s > .1).

3.3 Results

Proficiency (c-test scores): The average for the c-test scores was 77.63% (range 29 to 93, SD 13.96).

Question-Response Accuracy: For the test items and fillers, participants' comprehension performance was 77.23% or higher (mean 90%). For the test items, participants scored 81.25% or higher (mean 95.31%). There was no difference between the grammatical (94.53%) and the ungrammatical conditions (96.09%; $p = .385$).

Reading Times: We report results from ANOVA and mixed-effects models separately.

ANOVA: In region 2 (the head noun *chicken*), there was a trend for a grammaticality effect as the ungrammatical condition was faster than the grammatical condition, marginally in the by-subjects analysis and reliably in the by-items analysis (F_1 (1, 31) = 3.56, $p = .069$; F_2 (1, 15) = 5.99, $p = .027$). This replicates previous studies (for native speakers, see Lee and Cochran, 2000; Wager et al., 2009; and for learners, see Jiang, 2004; Wilson and Miyamoto, 2015). One possible reason for this difference is that the head noun in the grammatical condition was plural (*chickens*), therefore it was one character longer than the head noun in the ungrammatical condition.

In the critical region (region 6, the verb *were*), where the difference between the two conditions was predicted, there was no effect of grammaticality (F_1 (1, 31) = 2.35, $p = .135$; F_2 (1, 15) = 0.72, $p = .409$).

Previous studies often reported a reliable difference in the next region (for native English speakers: Pearlmutter et al., 1999; Jiang, 2004; Wagers et al., 2009). In our experiment, there was a trend for the ungrammatical condition to be slower than the grammatical condition in the by-subjects analysis but not in the by-items analysis (F_1 (1, 31) = 3.36, $p = .076$; F_2 (1, 15) = 1.41, $p = .254$).

Mixed-effects Models 1: Two types of analyses were conducted with mixed effects models. In the first type of analysis, log-transformed reading times to each region were analyzed as a function of *grammaticality* so as the results can be compared to those from the ANOVAs.

In region 2, there was an effect of grammaticality such that the ungrammatical condition was read faster than the grammatical condition ($\beta = -0.15$, $p < .001$).

At the critical region, there was no effect of grammaticality, but the numerical trend was for the ungrammatical condition to be slower than the grammatical condition ($\beta = 0.026$, $p = .412$).

In region 7, the ungrammatical condition was read significantly more slowly than the grammatical condition ($\beta = 0.06$, $p = .049$).

Mixed-effects Models 2: Proficiency is likely to be an important factor when analyzing reading times (Wen et al., 2010; Yamada and Hirose, 2012; Wilson and Miyamoto, 2015); therefore, the second type of mixed-effects models included grammaticality, c-test score, and their interaction as factors.

In region 1, there was an effect of c-test score ($\beta = -0.01$, $p < .001$) such that the higher their score was, the faster participants read. There was also an interaction between c-test and grammaticality ($\beta = 0.005$, $p = .007$). This interaction was unexpected because the word in this region was always the same (the article *the*). Participants may sometimes pause at random at the beginning of a sentence, or may be affected by the previous trial (e.g., they tend to slow down when they make a mistake answering the question in the previous trial).

Because of this spurious effect, for the remaining regions log-transformed reading times to region 1 were added as a covariate (analyses without the covariate revealed similar trends).

In region 2, the ungrammatical condition was faster than the grammatical condition ($\beta = -1.25$, $p < .001$). There was also an effect of c-test score as reading times were faster as the c-test score increased ($\beta = -7.56$, $p = .009$). Moreover, the covariate was reliable as reading times to region 1 were associated with slow reading times to region 2 (as indicated by the positive estimate, $\beta = 4.89$, $p < .001$).

In regions 3 and 4, there was a main effect of covariate (region 3, $\beta = 0.33$, $p < .001$; region 4, $\beta = 0.29$, $p < .001$).

In region 5, there was a main effect of c-test score such that the reading times were faster as the c-test score got higher (β = -1.13, p = .012). The effect of covariate was also reliable (β = 2.32, p < .001).

In region 6 (the critical region), there was an effect of covariate (β = 0.21, p < .001).

In region 7, the ungrammatical condition was reliably slower than the grammatical condition (β = 0.06, p = .043). There was a marginal effect of c-test score (β = -0.006, p = .062) suggesting that reading times got faster as the c-test score increased. Moreover, the covariate was reliable (β = 0.17, p < .001).

For all the later regions, there was a marginal effect of c-test score indicating faster reading times as the c-test score increased (β = -5.19, p = .053). The covariate was reliable (β = 1.54, p < .001).

4 Discussion

The results from mixed-effects models indicate that native Thai speakers are sensitive to agreement violations when reading L2 English, in line with previous results reported for native Japanese speakers (Wilson and Miyamoto, 2015). Results from ANOVAs were less clear cut as the effect of grammaticality was marginal in by-subjects and not reliable in by-items analyses. This is similar to a previous study with Chinese learners of English (Jiang, 2004), which detected no effect of grammaticality in by-subjects or in by-items ANOVAs using sentences comparable to those used in our experiment.

Mixed-effects models are increasingly common in the analyses of behavioral data as they have various advantages over traditional analyses such as ANOVAs (Baayen et al., 2008; Jaeger, 2008; and references therein). Our results (see the section titled Mixed-Effects Models 1) suggest that using mixed-effects models allow us to detect differences that were missed in previous studies, leading to rather different conclusions with respect to L2 acquisition and parsing.

More detailed mixed-effects analysis (see the section titled Mixed-Effects Models 2) indicates that proficiency (i.e., c-test score) contributes to explaining Thai learners' reading times as higher scores were associated with faster reading times. However, differences in proficiency did not affect the sensitivity to agreement violations as there was no interaction between grammaticality and proficiency. A previous study reported such an interaction suggesting that sensitivity to agreement violations were only found in native Japanese speakers with high proficiency in L2 English (Wilson and Miyamoto, 2015). However, preliminary analyses with mixed-effect models including the data for both Japanese and Thai participants revealed no 2-way interaction between L1 and grammaticality, or 3-way interaction between L1, grammaticality and proficiency, thus suggesting that the two groups' reading times are similar with respect to grammaticality.

The Thai participants' proficiency (mean c-test score: 71.16) was higher than the Japanese participants' (59.71). One possible reason for this difference is that the Thai participants were English majors. For a better comparison with the Japanese data, a new version of this study with Thai speakers not majoring in English is under way.

5 General Discussion

We reported the results of a reading-time experiment indicating that native Thai speakers are sensitive to number-agreement violations in L2 English. This suggests that not only can Thai speakers acquire an L2 feature absent in their L1, but they can also use this knowledge in a manner similar to that of native English speakers.

There is no overt morphological number marking in Thai; therefore, according to some past proposals (Hawkins and Chan, 1997), Thai learners of English should not be able to acquire agreement knowledge, contrary to our results.

According to other proposals (Clahsen and Felser, 2006), the PP intervening between the subject head noun and the verb in sentences as those in (2) may be too complex for learners to keep track of the agreement relation across it. Parsing such a complex structure may impose demands beyond learners' cognitive resources, forcing them to rely on simplified syntactic structures (as well as lexical information and world knowledge) to accomplish the task at hand. However, such a view is not easily reconciled with the agreement violation sensitivity detected in our experiment. Keeping track of agreement relations was unnecessary to interpret the sentences in our experiment. Nevertheless, that is what participants

seemed to have done without being aware of it (when asked after the experiment, participants never mentioned anything unusual or that some sentences were ungrammatical).

Our results are compatible with the view that learners start with the knowledge of their L1 and modify this knowledge to learn L2 (Schwartz and Sprouse, 1996). When the L1 does not contain agreement features, participants may take longer or may not be as consistent in acquiring the agreement system. This is not incompatible with past results in which learners displayed agreement sensitivity only if their L1 had agreement relations (Jiang et al., 2011).

Logically speaking, acquiring the knowledge is a necessary but not sufficient condition to guarantee that learners behave in a way similar to native English speakers. However, if we assume that the parser is universal to all human languages (Fodor, 1998), then acquiring the knowledge is indeed enough for us to expect L2 learners to approach native readers' behavior as was the case in our experiment.

The Thai speakers' data together with previously reported Japanese speakers' data (Wilson and Miyamoto, 2015) indicate that L2 learners can approach native speakers' performance even if their starting point (their L1) differs in crucial ways from the target L2.

This study is part of an on-going project to investigate L2 English parsing by native speakers of languages that for the most part lack number morphology, namely, Chinese, Japanese and Thai. Despite previous proposals and results that claimed otherwise (Hawkins and Chan, 1997; Jiang, 2004; Jiang et al., 2011), preliminary results indicate that speakers of these languages can display sensitivity to agreement violations. One first goal is to investigate in detail to what extent learners approach native speakers' way of processing agreement. There are detailed results on how native English speakers process number morphemes and the situations in which *intervention effects* can occur (Wagers et al., 2009). Our prediction is that similar trends would be observed with speakers of Chinese, Japanese and Thai reading L2 English.

Another goal is to investigate how L1 affects L2 learning. According to Corder (1981), similarities between L1 and L2 can facilitate acquisition by decreasing the steps in the learning process. Although all three languages lack number agreement, they differ in how similar they are to English in other respects such as word order. For example, (a) Thai and Chinese have Subject-Verb-Object (SVO) word order like English, whereas Japanese is SOV; (b) relative clauses in Thai, like in English, are postnominal (they follow the modified noun), whereas relative clauses are prenominal in Chinese and Japanese; (c) as in English, adjectives precede the head noun in Chinese and Japanese, but in Thai adjectives follow the head noun. From (a) – (c), Thai and Chinese are more similar to English than Japanese is. The question then is whether this type of similarity metric would have an impact on how learners can acquire number agreement in English. Factors such as motivation and attitude towards English are also been measured to eliminate some basic confounds.

6 Conclusion

The present study investigated the processing of English number agreement by Thai learners. The results show that similar to native English speakers, Thai learners slowed down when encountering an agreement violation. This result together with other recent studies (Wen et al., 2010; Yamada and Hirose, 2012; Wilson and Miyamoto, 2015) indicate that the ability for learners to acquire L2 knowledge is not restricted by the knowledge of their L1. Learners whose L1 lacks some crucial aspects of L2 are still capable of acquiring such missing knowledge.

Acknowledgment

This research was partly supported by MEXT/JSPS KAKENHI Grant Number 19K00609 to the second author.

References

Aho, A. V., Lam, M. S., Sethi, R., & Ullman, J. D. (2007). *Compilers: principles, techniques, and tools* (2 ed.). USA: Pearson Education.

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to statistics using R*: Cambridge University Press.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of memory and language, 59*(4), 390-412.

Babaii, E., & Shahri, S. (2010). Psychometric rivalry: The C-test and the close test interacting with test takers' characteristics. In R. Grotjahn (Ed.), *The C-Test: Contributions from Currect Research* (pp. 41-56). Frankfurt:Lang.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015). Parsimonious mixed models. *arXiv preprint arXiv:1506.04967*.

Clahsen, H., & Felser, C. (2006). Grammatical processing in language learners. *Applied Psycholinguistics, 27*, 3-42.

Corder, S.P. (1981). *Error analysis and interlanguage*. New York: Oxford University Press.

Fodor, J. D. (1998). Learning to parse? *Journal of Psycholinguistic Research, 27*(2), 285-319.

Hawkins, R., & Chan, C. Y.-h. (1997). The partial availability of Universal Grammar in second language acquisition: The "failed functional features hypothesis". *Second Language Research, 13*, 187-226.

Iwasaki, S., & Ingkaphirom, P. (2009). *A reference grammar of Thai*: Cambridge University Press.

Jaeger, T. F. (2008). Categorical data analysis: Away from ANOVAs (transformation or not) and towards logit mixed models. *Journal of memory and language, 59*(4), 434-446.

Jiang, N. (2004). Morphological insensitivity in second language processing. *Applied Psycholinguistics, 25*, 605-634.

Jiang, N., Novokshanova, E., Masuda, K., & Wang, X. (2011). Morphological Congruency and the acquisition of L2 morphemes. *Language Learning, 61*(3), 940-967.

Lee, C. H., & Cochran, M. F. (2000). Controlling two confounding variables in word length: "Vanished word-length effect". *Reading Psycholgy, 21*, 57-66.

Pearlmutter, N. J., Garnsey, S. M., & Bock, K. (1999). Agreement processes in sentence comprehension. *Journal of memory and language, 41*, 427-456.

R Core Team (2018). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Schwartz, B. D., & Sprouse, R. A. (1996). L2 cognitive states and the Full Transfer / Full Access model. *Second Language Research, 12*, 40-72.

Wagers, M. W., Lau, E. F., & Phillips, C. (2009). Agreement attraction in comprehension: Represetation and processes. *Journal of memory and language, 61*, 206-237.

Wen, Z., Miyao, M., Takeda, A., Chu, W., & Schwartz, B. D. (2010). *Proficienc effects and distance effects in nonnative processing of English number agreement.* Paper presented at the Boston University Conference on Language Development, Boston, USA.

Wilson, B. G., & Miyamoto, E. T. (2015). *Proficiency effects in L2 processing of English number agreement across structurally complex material.* Paper presented at the the 151st Meeting of the Linguistic Society of Japan, Nagoya University.

Yamada, T., & Hirose, Y. (2012). *Singular / plural asymmetry in Japanese EFL learners' sensitivity to English number dis/agreement.* Paper presented at the Japanese Cognitive Science Society, 29.

May and *Can* Constructions in Spoken Corpus: A Constructionist Approach

Tsi-Chuen Tsai
National Chengchi University
No. 64, Sec. 2, Zhi Nan Rd.,
Wenshan District, Taipei City 11605,
Taiwan, R.O.C.
102551505@nccu.edu.tw

Huei-Ling Lai
National Chengchi University
No. 64, Sec. 2, Zhi Nan Rd.,
Wenshan District, Taipei City 11605,
Taiwan, R.O.C.
hllai@nccu.edu.tw

Abstract

This study investigates partially filled *may* and *can* constructions in the Spoken British National Corpus 2014 (Spoken BNC2014). A constructionist perspective is taken to examine the structure and distribution of *may* and *can* constructions. It is assumed that associative relations between the modal verbs and the contextual elements in the constructions designate the expressions of *may* and *can*. Adopting the collostructional analytical procedure, we identified the major [it+*may*+be+*] constructions, from which we generalized its function on enhancing the informativeness of the utterance. This function is distinct from that of [it+*can*+be+*], which is used to highlight common human capability, feelings or experience. The analysis confirms the status of modal construction and successfully distinguishes *may* and *can* constructions, which exhibit distinct features and express dynamic meanings. The findings also provide empirical evidence to a theoretical perspective that sees language as a result of use.

1. Introduction

The English language features a set of modal verbs which are central to the expression of modality-the speaker's attitudes or opinions toward the proposition of the utterance (Hoye, 1997). While identification of the syntactic features of modal verbs is quite a straightforward matter, modal semantics has been subject to heated debate for decades. There has been no consensus among linguists regarding the types or number of modality and there has been no agreement on an analytical approach toward the elucidation of the notion (Nuyts, 2005). For ease of discussion, we begin with three of the most recognized; epistemic, deontic and dynamic modality. Epistemic modality involves the estimation, by the speaker, of the possibility that the state of affairs is real. On the other hand, deontic modality is related to social norms or personal ethical criteria. Finally, dynamic modality describes the capacity or needs of the controlling-participant or similar potentials determined by the local circumstances. In Quirk et al. (1985), the former is referred to as 'intrinsic' while the latter is called 'extrinsic'. The above introduction suggests that modality may vary in degree and it is subject to different interpretations and sensitive to the sources of potential where it is generated. Since a majority of modal verbs may convey epistemic, deontic or dynamic meanings simultaneously, the study of modality continues to challenge linguists to come out with a clearer description not just within individual modals but distinction among different modals.

Recently, the topic has been approached from a constructionist perspective, which examines modality in terms of a network of constructions rather than sense relationship (Boogaart, 2009). Empowered by corpus linguistics, studies taken a constructionist approach have yielded fruitful results to provide a more comprehensive account of modality (Anthonissen & Mortelmans, 2016; Cappelle & Depraetere, 2016; De Haan, 2012; Deshors & Gries, 2014; Hilpert, 2013, 2016). Nonetheless, while the foci of most corpus studies have been on examining the verb groups associated with modal verbs, less attention has been given to some other important components of modal constructions, namely the grammatical subject and copular structure. Taking the partially filled *may*

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 244-252, Hakodate, Japan, September 13-15, 2019

and *can* constructions as examples, this paper attempts to demonstrate the usefulness of a constructionist approach in combination with corpus linguistics to provide a more precise and detailed description of modality. Particularly, we compare and contrast (1) the central elements in [it+*may*+be+*] and [it+*can*+be+*] constructions in spoken corpus, and explicate (2) their generalized meanings or functions. The rest of the paper is organized as follows. Section two provides a brief review of some previous reports on *may* and *can*. Section three introduces the methodology. Sections four and five present the result, and section six concludes the study.

2. The meanings of *may* and *can*

May and *can* have been recognized as polysemy as well as near-synonyms not just by the multiple senses they each possess but by the much overlapping of their senses. The significance of the pair can be observed in the extensive literature devoted to their identification (Coates, 1983; Collins, 2007; Dirven, 1981; Duffley *et al*., 1981; Groefsema, 1995; Quirk et al., 1985). The following sections briefly summarize some of the major conceptions.

2.1 *May* and *can* as polysemy

It is common to see *may* and *can* juxtaposed in the discussion of modality. For instance, they were grouped in a category to express permission, possibility, and ability (Quirk et al., 1985, p. 221). The conception is presented in Figure 1.

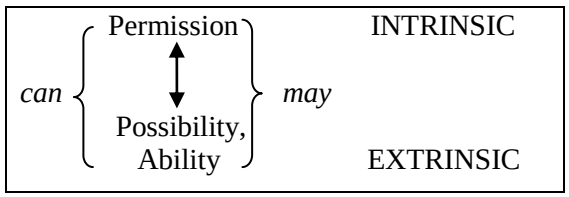


Figure 1: Meanings of *may* and *can*

The representation shows that the two modal verbs are semantically interchangeable and the variance in senses mainly results from the different enabling sources labeled 'intrinsic' and 'extrinsic'. In fact, the various senses seem to form a continuum rather than distinct categories. As Quirk et al. put it, "The ability meaning of *can* is considered extrinsic, even though ability typically involves human control over an action. Ability is best considered a special case of possibility" (p. 221). In contrast to Quirk et al., who collapsed the multiple modal meanings, Coates (1983) believed that *may* and *can* held distinct interpretations. In her investigation of approximately 200 instances of modal verbs in written and spoken corpora of British English, Coates came to the conclusion that *may* primarily denoted epistemic sense, which appeared equally frequent in both spoken and written genres and was found to co-occur with hedges like *I suppose* or adverbs such as *perhaps*. On the other hand, *can* mostly communicated non-epistemic sense which in its definition appears to correspond to Quirk et al.'s extrinsic possibility. As Coates explained, "CAN can be seen as implying a universe of possible worlds, ranging from the most restricted (where human laws and rules are in force) to the least restricted (where everything is permitted except what is contrary to so-called natural laws)" (p. 88).

Coates also used the term 'merger' to refer to instances where modal meanings became ambiguous and *may* and *can* were interchangeable. In those cases, she believed that the two modals may be distinguished in terms of degree of formality with *may* indicating a higher level of formality. Elsewhere, Wärnsby (2006) believed that the ability sense may be subsumed under weak epistemic possibility (it is possible for…) as opposed to strong epistemic sense (it is possible that…) shown in the following examples (p. 16).

(a) The window *can* be broken. (weak possibility)
    It is possible for the window to be broken.
(b) The window *may* be broken. (strong possibility)
    It is possible that the window is broken.

Wärnsby added that the two senses also differ in the way they refer to the time when the utterance may be verified. The weak sense indicates that the speaker makes reference to non-linguistic circumstances that can only be verified after the time of the utterance while the strong sense suggests that the speaker's belief can be verified at the time of the utterance. In any case, Wärnsby's argument reminisces Quirk et al.'s grossing of *may* and *can* presented in Figure 1. In sum, despite exhaustive categorization and sense analysis, issues regarding the boundary of modal senses as well as their overlap remain unresolved.

2.2 A constructionist approach toward polysemy

Boogaart (2009) pointed out the inadequacy of a notional explanation of modality, which interprets modal verbs in terms of a network of senses. He urged for a shift of attention from generating abstract meanings in isolated modals to identifying specific and concrete constructions which have modals as part of their composition. Unlike sense analysis, the constructionist approach sees human knowledge of language as a conglomeration of conventional, learned form-meaning pairings known as constructions or the building blocks of language. Goldberg (2003) provided the following definition:

... constructions which are stored pairings of form and function, including morphemes, words, idioms, partially lexically filled and fully general linguistic patterns (p. 219).

The definition highlights the major principle of the constructionist perspective in which all linguistic items however small or abstract are learned pairings of form and function. By treating constructions as symbolic units, the constructionist approach disregards the distinction traditionally made between lexicon and syntax. With its emphasis on form-function mapping, the constructionist approach is especially suitable for the analysis of polysemy like *may* and *can*. As Goldberg (2013, p. 19) put it:

... if a single phrasal pattern were truly associated with unrelated functions, then their distributional behavior is not likely to be identical. When behavior diverges, we generally decide that the syntax involved is not the same.

Conversely, any change in syntactic form may lead to a difference in meaning (Bolinger, 1968). By postulating an interconnected network of constructions, the constructionist approach regards polysemy as a result of a cognitive organizing principle shared by all areas of language, such as morphology, lexicon, and syntax. Moreover, it is believed that the meanings of polysemy are related in a systematic way to form radial categories where the more frequent and prototypical sense is related to less frequent and more peripheral ones (Kovács, 2011).

2.3 Corpus studies on *may* and *can* constructions

Supported by rich empirical data and computational power, corpus linguistics has gained prominence over the past several decades. Collin (2007) investigated *may* and *can* in three parallel English corpora based on the tripartite taxonomy of modality: deontic, epistemic, and dynamic. However, by limiting his analysis to frequency count and sense analysis, his findings were not very informative. For instance, he concluded that *may* primarily conveyed epistemic possibility whereas *can* denoted dynamic possibility with the ability sense subsumed under the category. The finding is not illuminative because it seems to reiterate the existing literature, which has already failed to distinguish *may* and *can.* In general, Collins's observation only manifests the complexity of the issue.

To better understand modality, Hilpert (2016) argued intensively for the incorporation of corpus linguistics with a constructionist perspective. He stressed that the notion of construction or form-function pairing can be better captured through the collostructional analysis, which measures the attraction or repulsion of various linguistic forms toward each other. Results from corpus analysis may highlight significant associative relations between modal verbs and other lexical elements as well as their interaction with the schematic construction, namely [NP+Modal Verb+Verb]. To demonstrate, Hilpert studied *may* construction in the Corpus of Historical American English (COHA) where he identified important verb groups that were responsible for the diachronic semantic shift of *may*. From co-occurrence frequencies, he observed that over the past two centuries *may* has come to be used more often with verbs that are abstract, stative, and unrelated to animate subjects, such as *depend*, *exist*, *involve*, or *indicate*, which are predominantly linked to informational types of text. The analysis allowed Hilpert to specify elements that have caused the change in *may* from deontic sense towards epistemic meaning. Crucially, the result explained the confounding polysemy observed in modal verbs and brought to light the reason why *may* in modern English tends to be associated with informativeness.

Encouraged by Hilpert's finding, Cappelle and Depraetere (2016) proposed that a wider scope of attention be given to associations between the modal verb and linguistic elements other than the following lexical verbs. To testify the model, Deshors and Gries (2014) conducted a multifactorial assessment to investigate the structures of *may* and *can* in written French-English interlanguage. They researched 22 morphosyntactic and semantic features as well as their interaction to identify their effects on the native and non-native use of *may* and *can*. The result showed great variation between native speakers' and learners' modal constructions. In terms of form, the learners used fewer *may* in subordinate clauses and negated clauses and they were more likely to associate *can* with animate and singular subjects. As for the verb groups, the learners preferred abstract verbs with *can* and they favored time or place verbs with *may*. Nevertheless, Deshors and Gries did not distinguish copular structure used in conjunction with *may* and *can* despite its prominent presence in both the native and non-native corpora. In general, their study attested the effectiveness of the collostructional analysis, which has shed light on the effect of the linguistic context on the use of *may* and *can*.

At present, there are few studies on modality taking a constructionist perspective and there is even less attention to modal representation in spoken data. While Hilpert highlighted the importance of entrenched patterns, he set aside such 'highly frequent' (p. 76) features as [*may*+be+*] or [*can*+be+*] to future research. Similarly, by focusing their attention on major co-occurring verb groups, Deshors and Gries left the details of the above two prevalent constructions undiscussed. On the other hand, where Collins noted ambiguous instances like 'it can/may be cold in Stockholm,' (p. 490) he simply assigned the meaning as a merger, still leaving the controversy unresolved. Following Cappelle and Depraetere's advice, this study aims at uncovering the meanings of these partially filled constructions. We believe the combination of a constructionist perspective and a corpus analytical approach may provide more detailed information and help distinguish *may* and *can*.

3. Methodology

The data for this study were collected from the free online Spoken British National Corpus 2014 (Spoken BNC2014). The corpus contains 11.5 million words of transcribed content featuring real-life, informal British English conversations (Love, Dembry, Hardie, Brezina, & McEnery, 2017). This study adopted the collostructional analysis to observe an alternating pair of partially filled modal constructions. The term collostructional (a blend of construction and collocational) refers to equal attention paid to syntactic and semantic structures where the modal verbs are found. We made use of the built-in functions provided by the annotated Spoken BNC2014 to identify the collocates of the pair constructions. The primary function used for the investigation was Loglikelihood score (Log), which measures the strength of association among collocations: the higher the score, the more significant the association. Take *may* as an example. We began by typing the target word *may* as [may_VM] in the query box in Spoken BNC2014 to extract instances of *may* used as a modal verb. The initial results showed that there were 119 instances of *may* and 3298 occurrences of *can* in per million words.

3.1 Schematic *may* and *can* constructions

We identified the schematic *may* and *can* constructions by setting the window span as R1 to R1 (to the right of the modal) and by selecting the part-of-speech tag in the collocation function. The result showed that while the most significant structure of *may* was [*may*+be] (Freq: 240/Log:1286), [*can*+be] (Freq:1163/Log:2327) was ranked fifth as *can*'s favorite collocate (Freq refers to frequency). The frequency and distribution of collocated part of speech retrieved from the corpus suggests that *can* (198 types) is a far more productive construction than *may* (139 types) and it can be predicted that the semantics of *can* construction will be more dynamic. In the next step, we conducted part-of-speech search on the L1 to L1 of [*may*+be] and [*can*+be] constructions, which produced a list of [NP+*may*/*can*+be] candidates. Tables 1 and 2 present the occurrences of the top three exemplars of schematic *may* and *can* constructions (Log score is presented in parenthesis).

May construction	Freq (Log)
There *may*+be	43 (212.46)
It *may*+be	59 (153.56)
They *may*+be	16 (31.69)

Table 1: Schematic *may* constructions

Can construction	Freq (Log)
It *can*+be	271 (676.12)
You *can*+be	177 (337.7)
They *can*+be	110 (287.39)

Table 2: Schematic *can* constructions

3.2 Partially filled *may* and *can* constructions

We administered another collocation search on the R2 to R2 of our target constructions [it+*may*/*can*+be] (the most significant construction for *may* and the second most significant for *can*) and identified the top three most frequent collocates for each construction. These items were the central members of the categories that filled in the schematic slots of *may* and *can* constructions and they represented the semantics of the categories (Bybee & Eddington, 2006). The result showed that [it+*may*+be+that+clause] (7 tokens), [it+*may*+ be+a+Noun] (7 tokens), and [it+*may*+be+Adv.+ clause/Noun] (8 tokens) were the central members of *may* construction while [it+*can*+be+Adv.+Adj.] (44 tokens) and [it+*can*+be+passive PP] (29 tokens) were important *can* constructions. To validate our findings, we queried constructions with the pronoun *it* as subject and found that while [it+lemma be] was prominent, [it+Modal Verb] was not. Meanwhile, [it+*+be] was only mildly related to modal constructions since the occurrence of modals in the slot was relatively insignificant. The result confirmed the status of *may* and *can* constructions because they are not random composition of elements but their occurrences reflect the probability of natural language use.

4. [It+*may*+be+*] construction

In this section, the functions of the three partially filled *may* constructions will be discussed in accordance with the definition of construction provided by the literature. Each construction is seen as a linguistic sign that represents a form pertaining to the phonology or morphosyntax and is equipped with its own semantic and discourse-pragmatic characteristics. Following Cappelle and Depraetere's (2016) advice, we take into consideration the linguistic context where the construction is located to give a more comprehensive understanding of its use.

4.1 The evaluating [it+*may*+be+that+clause]

This construction most frequently occurs as an evaluation to a situation. It serves as a support to the speaker's observation about an on-going event. In all the instances, the construction is always followed by a statement of fact with reference to common knowledge. That is, the construction is meant to bolster a personal claim based on a shared assumption with the other interlocutors. Examples (1) and (2) demonstrate the function.

(1) **A:** *it's funny that they're always louder though aren't they?*
B: *yes yeah*
A: *they're always loud*
B: >> ***it may be that*** *he's slightly deaf as well*
A: *yeah* (SHTW)

(2) **A:** *we don't want it back er the only things that I'll want back*
B: *or at the end we'll decide if you want it back or not*
A: *okay yeah* ***it may be that*** *--ANONnameM might want some stuff if he's moving into a house* (SAA3)

In example (1), speaker B employs the construction to introduce his appraisal of an event. He reasons that the crowd's tendency to be loud might have something to do with an unnamed individual's poor hearing. His claim is based on the common sense that people with poor hearing tend to speak louder or need to be spoken to loudly. Further evidence to the construction's evaluating role comes from the fact that all the statements following the construction are in the present tense and they primarily refer to events that are in the past or are evolving. For instance, in example (2), the speakers seem to be discussing the allocation of certain objects. After several turns of negotiation, speaker A conceded to speaker B's argument by starting his turn with 'okay' and 'yeah'. However, these positive markers appear to be mere polite recognition instead of submission to B's proposal. In fact, speaker A stands his ground by introducing [it+*may*+be+that+clause] with shared knowledge that there are other candidates to accept speaker

B's offer. This use of the construction reduces possible awkwardness caused by a conflict of opinion between the speakers. The finding is validated by hedges like *I think* or *you know*, adverbs such as *slightly* or other modals used in the clause following *that*. These devices suggest that [it+*may*+be+that+clause] concerns the speaker's evaluation about an event, a situation or a proposal he intends to comment on.

4.2 The specifying [it+*may*+be+a+Noun]

Similar to [it+*may*+be+that+clause], [it+*may*+be+a+Noun] was used to relate to the focus of a conversation. However, the latter functioned to specify an object of attention rather than an event. Moreover, little constraint was placed on the time when the event occurred. The object of focus may be located at present, in the past, or in the future. Examples (3) and (4) represent the use.

(3) **A:** *I think I think you wind it and then once the record 's finished I think you wi- I don't know I think you wind it again*
B: *because* ***it may be a*** *thing th- the right getting the right speed on that for the records -- UNCLEARWORD*
A: *yeah there is yeah yeah (.)* (S3SA)

(4) **A:** *yes*
B: *by instinct*
A: *they say a lot of it was well someone said to me once and I tend to agree with them and* ***it may be a*** *column I'm sure* (S7K2)

In example (3), the speakers appear to be working on a task at hand. They seem to encounter a technical issue where speaker B employs the construction to orient the conversation. The construction contributes to the identification of the target and facilitates the flow of exchange as well as problem solving. Likewise, example (4) shows how the construction is used to situate the item of interest in the past. By the contextual element *I am sure* that follows the construction, we learn that the speaker is searching in his memory for the source of evidence to support his claim. In some other instances, the construction is used to postulate an object of attention in the future. For example, in reference to an egg hunt, one speaker used the construction to make prediction about the item which would be used for the hunt in *or well it may be a chocolate bunny*, which attested the specifying function of the construction.

4.3 The focusing [it+*may*+be+Adv.+clause/N]

This construction occurs with adverbs that indicate degree of speaker attitude on the event or the object he or she is commenting on. These adverbs range from those that signify the speaker's affirmation of truth such as *actually* or *apparently* to those that give value judgment like *right*. This use highlights the role of the construction as a focusing device illustrated in example (5).

(5) **A:** *but it's something she likes doing a lot of the the crafting stuff*
B: *mm*
A: *so* ***it may be actually*** *she'd think oh actually I could make some- when it's a birthday we always get a nice handmade birthday card*
B: *yeah lovely* (S64H)

In example (5), speaker A utilizes the construction to reinforce a point he has made in the previous turn about an unnamed individual's desire or preference. The construction introduces a similar concern with added information related to the selection of an ideal birthday gift for the individual. The co-occurrence of the construction with an emphasizer *actually* demonstrates the speaker's confidence on the truth of his or her remark. Although downtoners such as *probably* or *partly* also appear in the construction, we found their function to be similar in drawing focus to the speaker's point.

5. [It+*can*+be+*] construction

This section discusses the three partially filled *can* constructions. The slot of the constructions was filled with a variety of items, which were first categorized before we proceeded to explain the functions of the constructions. Since [it+*can*+be+Adv.+*] has two daughter constructions, their characteristics are expounded in two separate sections.

5.1 The representing [it+*can*+be+very/quite+Adj.]

The slot which designates adverbs in the construction was primarily dominated by two adverbs which modified adjectives, namely *quite*

(15 tokens) and *very* (9 tokens). However, the adjectives that were modified by the two adverbs were comprised of miscellaneous semantic groups, which can be roughly categorized in terms of their association with perceptual (e.g., *bright, bland*), physical (e.g., *painful, hurtful*), psychological (e.g., *miserable, boring*), intellectual (e.g., *hard, tricky*) or circumstantial (e.g., *dangerous, bleak*) conditions. Meanwhile, it was noted that the primary referents of the construction often entail human experience. Examples (6) and (7) exhibit instances of the use.

(6) **A:** *it might have been a child and they weren't wearing a seat*
B: *yeah*
A: *I think we're here now*
B: *although on children in general* ***it can be quite*** *dangerous to have a seatbelt cos it can crus- crush like your ribs and stuff* (S9V8)

(7) **A:** *I suppose that it's got I'm not like really massively interested in um economics*
B: *yeah I know well it it can be ver-* ***it can be very*** *boring I understand why people find it boring because they it's quite technical and (.)* (SF2F)

Example (6) depicts a scenario where the speakers were discussing the usage of seatbelts and example (7) is a conversation about a school subject. Close examination reveals that a majority of the adjectives that filled in the slots of the construction tend to carry negative prosody and make the utterance sound distressing, annoying or alarming. That is, [it+*can*+be+very/quite+Adj.] imposes its effect by bringing out common experience or unpleasant images residing in the mind of a fellow humankind. Moreover, we found that when the adjectives describe human potential or when the construction is followed by an infinitive *to* phrase, the utterance supplies an agency sense to the subject pronoun *it* to entail the ability sense of modality.

5.2 The acknowledging [it+*can*+be+a bit+Adj.]

[It+*can*+be+a bit+Adj.] (15 tokens) constitutes a highly frequent and significant category. While *a bit* may be regarded as an adverb or modifier, we found the objects that are modified by *a bit* to comprise of a distinct semantic category. Example (8) exemplifies the use.

(8) **A:** *the daughters but I mean she wasn't mentioned anywhere so whether she had made all this up or*
B: ***it can be a bit*** *weird when people don't check this er a woman I used to work with (.)* (S68F)

In example (8), speaker A raised an issue regarding his suspicion about the authenticity of the personal information reported by an individual. In response, speaker B employs the construction to recognize speaker A's thought. The adjective that follows *a bit* is used to sum up the interlocutor's feeling or thought in a single word. That is, the construction performs an acknowledging function by resonating the interlocutor's concern. Close observation shows that the adjectives that follow *a bit* tend to portray negative or unpleasant experiences such as *delusional*, *boring*, *tough*, or *greasy*. In some cases, the construction introduces some kind of flaw about the activity under discussion. The recognition of certain exceptions to what is agreed by the interlocutors ensures that there is no misunderstanding and that the conversation can proceed without any hurdles.

5.3 The instructive [it+*can*+be+passive PP]

This construction is distinct in the way that it allows verbs with strong transitivity to enter the slot after the *be* verb and in doing so constrains the referent to the subject pronoun *it*. The semantics of these verbs is diverse to include alteration (e.g., *extend*, *repair*), allocation (e.g., *put*, *include*), perception (e.g., *see*), manipulation (e.g., *use*, *abuse*), or hindrance (e.g., *close*, *block*) etc. The following is one example.

(9)**A:** *erm (.)I swear actually pretty much never around my dad (.) I don't know I think swearing it's interesting* ***it can be used*** *for emphasis or to add colour to a*
B: *mm hm*
A: *or depth or texture to a conversation* (S94U)

By and large, most of these verbs seem to concern human adjustment to or effect on their environment. As such, the referents to the subject pronoun *it* are related to objects that may be subject to human manipulation or resistance. In example (9), the construction introduces the focus of the conversation, *swearing*, whose function is to

highlight or enrich the content of an argument. Swearing is of course a human activity and is only possible through human language. Elsewhere, the construction conveys human-only capability in managing certain objects or tasks. In sum, despite the diverse semantics depicted by the verbs, the construction primarily performs an instructive function addressing possible human effect on their environment.

6. Concluding remarks

Our analysis uncovered the central members of [it+*may*+be+*] and [it+*can*+be+*] constructions, which exhibit distinct features and express dynamic meanings. These meanings were arrived at by taking into account the constructions as a whole rather than postulating a set of abstract features that describe the modal only. These constructions were found to connect to each other in a systematical way to form a hierarchical network of constructions. The relationship between the various levels of constructions has been recognized as a process of generalization. As Goldberg (2003) explained, "Broad generalizations are captured by constructions that are inherited by many other constructions; more limited patterns are captured by positing constructions at various midpoints of the hierarchical network. Low level constructions represent exceptional patterns" (p. 221-222). Through generalizations, the meaning of partially filled [it+*may*+be+*] can be captured by studying the lower-level constructions that are related to it as shown in Figure 2.

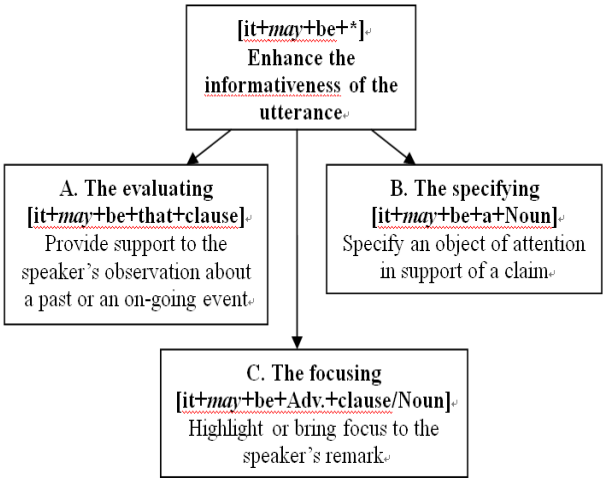


Figure 2: Network of central [it+*may*+be+*] constructions

Likewise, the meaning of [it+*can*+be+*] can be generalized in a similar manner illustrated in Figure 3.

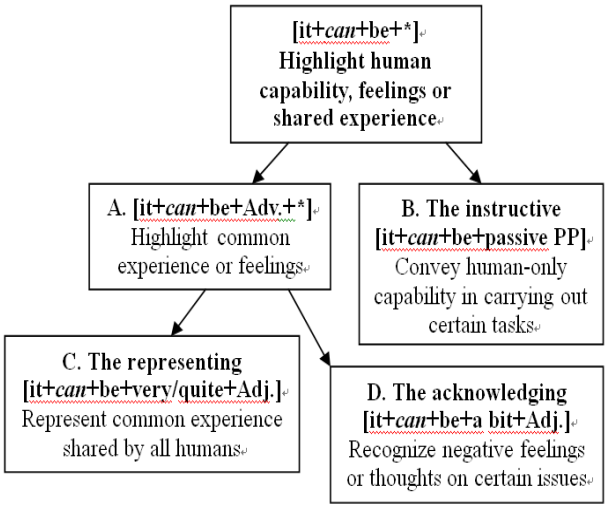


Figure 3: Network of central [it+*can*+be+*] constructions

Compared with sense analysis which focuses on individual modals, the constructionist perspective provides more detailed information and allows us to arrive at a more precise and accurate description of modality. For instance, although structurally, [it+*may*+be+*] and [it+*can*+be+*] are alike, our analysis revealed that the former was associated with statements of fact while the latter was related to human potential or experience. The result explains why there is an intuitive association of *may* with epistemic sense and *can* with non-epistemic meaning. This is because as the elements on the top of the modal hierarchy, *may* and *can* are inherited by many other constructions and have come to realize the generalized meanings of all their daughter constructions. The constructionist approach captures this dynamic relationship among related constructions and by doing so, it not only infuses analytical power to the distinction of *may* and *can* but also adds empirical evidence to untie a theoretical deadlock on modal polysemy. We believe with more research endeavor, the connectivity and systematicity of modal constructions or language construction in general can be more fully explicated. The findings also bear important implications for lexicography and language pedagogy, which rely heavily on attested data to present a more complete picture of our language.

Acknowledgements

This work was supported in part by the Ministry of Education under the Grants 107H121-08.

References

Lynn Anthonissen & Tanaj Mortelmans. 2016. German modals in second language acquisition: A constructionist approach. Yearbook of the German Cognitive Linguistics Association, 4(1), 9-30.

Dwight Bolinger. 1968. Entailment and the meaning of structures. Glossa, 2, 119–27.

Ronny Boogaart. 2009. Semantics and pragmatics in construction grammar: The case of modal verbs. In Alexander Bergs and Gabriele Diewald (Eds.), Contexts and constructions, pp. 213–241. Amsterdam: John Benjamins.

Joan L. Bybee and David Eddington. 2006. A usage-based approach to Spanish verbs of 'becoming'. Language, 82(2): 323-355.

Bert Cappelle & Ilse Depraetere. 2016. Response to Hilpert. Constructions and Frames, 8(1): 86-97.

Jennifer Coates. 1983. The Semantics of the Modal Auxiliaries. London: Croom Helm.

Peter Collins. 2007. Can/could and may/might in British, American and Australian English: a corpus-based account. World Englishes, 26(4), 474-491.

Ferdinand De Haan. 2012. The relevance of constructions for the interpretation of modal meaning: The case of must. English Studies, 93(6): 700-728.

Sandra C. Deshors & Stefan Th. Gries. 2014. A case for the multifactorial assessment of learner language: The uses of may and can in French-English interlanguage. In Dylan Glynn & Justyna A. Robinson (Eds.), Corpus Methods for Semantics: Quantitative studies in polysemy and synonymy, pp. 179-204. Amsterdam: John Benjamins.

Rene Dirven. 1981. Pragmatic forces associated with can- and may-sentences. Folia Linguistica, 13, 145-215.

Patrick Duffley, Sandra Clarke, & Walter Hirtle. 1981. May, can, and the expression of permission. Canadian Journal of Linguistics, 26, 161-202.

Adele E. Goldberg. 2003. Constructions: A new theoretical approach to language. Trends in cognitive sciences, 7(5): 219-224.

Adele E. Goldberg. 2013. Constructionist Approaches. In T. Hoffmann & G. Trousdale (Eds.), The Oxford Handbook of Construction Grammar, pp. 15-31. Oxford: Oxford University Press.

Marjolein Groefsema. 1995. Can, may, must and should: a relevance theoretic account. Journal of Linguistics, 31(1): 53-79.

Martin Hilpert. 2013. Constructional change in English: Developments in allomorphy, word formation, and syntax. Cambridge: Cambridge University Press.

Martin Hilpert. 2016. Change in modal meanings: Another look at the shifting collocates of may. Constructions and Frames, 8(1): 66-85.

Leo Hoye. 1997. Adverbs and modality in English. New York: Longman.

Éva Kovács. 2011. Polysemy in traditional vs. cognitive linguistics. Eger Journal of English Studies XI, 3-19.

Robbie Love, Claire Dembry, Andrew Hardie, Vaclav Brezina, and Tony McEnery. 2017. The Spoken BNC2014: designing and building a spoken corpus of everyday conversations. International Journal of Corpus Linguistics, 22(3): 319-344.

Jan Nuyts. 2005. The modal confusion: on terminology and the concepts behind it. In Alex Klinge & Henrik Hegel Müller (Eds.), Modality: studies in form and function, pp. 5-38. London: Equinox.

Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, & Jan Svartvik. 1985. A Comprehensive Grammar of the English Language. London: Longman.

Anna Wärnsby. 2006. (De)coding Modality. The Case of Must, May, Måste, and Kan. Lund Studies in English 113. Lund: Lund University.

On the Effectiveness of Low-Rank Matrix Factorization for LSTM Model Compression

Genta Indra Winata, Andrea Madotto, Jamin Shin, Elham J. Barezi, Pascale Fung
Center for Artificial Intelligence Research (CAiRE)
Department of Electronic and Computer Engineering
The Hong Kong University of Science and Technology, Clear Water Bay, Hong Kong
{giwinata,amadotto,jmshinaa,ejs}@connect.ust.hk
pascale@ece.ust.hk

Abstract

Despite their ubiquity in NLP tasks, Long Short-Term Memory (LSTM) networks suffer from computational inefficiencies caused by inherent unparallelizable recurrences, which further aggravates as LSTMs require more parameters for larger memory capacity. In this paper, we propose to apply low-rank matrix factorization (MF) algorithms to different recurrences in LSTMs, and explore the effectiveness on different NLP tasks and model components. We discover that additive recurrence is more important than multiplicative recurrence, and explain this by identifying meaningful correlations between matrix norms and compression performance. We compare our approach across two settings: 1) compressing core LSTM recurrences in language models, 2) compressing biLSTM layers of ELMo evaluated in three downstream NLP tasks.

1 Introduction

Long Short-Term Memory (LSTM) networks (Hochreiter and Schmidhuber, 1997; Gers et al., 2000) have become the core of many models for tasks that require temporal dependency. They have particularly shown great improvements in many different NLP tasks, such as Language Modeling (Sundermeyer et al., 2012; Mikolov, 2012), Semantic Role Labeling (He et al., 2017), Named Entity Recognition (Lee et al., 2017), Machine Translation (Bahdanau et al., 2014), and Question Answering (Seo et al., 2016). Recently, a bidirectional LSTM has been used to train deep contextualized Embeddings from Language Models (ELMo) (Peters et al., 2018), and has become a main component of state-of-the-art models in many downstream NLP tasks.

However, there is an obvious drawback of scalability that accompanies these excellent performances, not only in training time but also during inference time. This shortcoming can be attributed to two factors: the temporal dependency in the computational graph, and the large number of parameters for each weight matrix. The former problem is an intrinsic nature of RNNs that arises while modeling temporal dependency, and the latter is often deemed necessary to achieve better generalizability of the model (Hochreiter and Schmidhuber, 1997; Gers et al., 2000). On the other hand, despite such belief that the LSTM memory capacity is proportional to model size, several recent results have empirically proven the contrary, claiming that LSTMs are indeed over-parameterized (Denil et al., 2013; James Bradbury and Socher, 2017; Merity et al., 2018; Melis et al., 2018; Levy et al., 2018).

Naturally, such results motivate us to search for the most effective compression method for LSTMs in terms of performance, time, and practicality, to cope with the aforementioned issue of scalability. There have been many solutions proposed to compress such large, over-parameterized neural networks including parameter pruning and sharing (Gong et al., 2014; Huang et al., 2018), low-rank Matrix Factorization (MF) (Jaderberg et al., 2014), and knowledge distillation (Hinton et al., 2015). However, most of these approaches have been applied to Feed-forward Neural Networks and

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 253-262, Hakodate, Japan, September 13-15, 2019

Convolutional Neural Networks (CNNs), while only a small attention has been given to compressing LSTM architectures (Lu et al., 2016; Belletti et al., 2018), and even less in NLP tasks. Notably, (2016a) applied parameter pruning to standard Seq2Seq (Sutskever et al., 2014) architecture in Neural Machine Translation, which uses LSTMs for both encoder and decoder. Furthermore, in language modeling, (2017) uses Tensor-Train Decomposition (Oseledets, 2011), (2018) uses binarization techniques, and (2017) uses an architectural change to approximate low-rank factorization.

All of the above mentioned works require some form of training or retraining step. For instance, (2017) requires to be trained completely from scratch, as well as distillation based compression techniques (Hinton et al., 2015). In addition, pruning techniques (See et al., 2016a) often accompany selective retraining steps to achieve optimal performance. However, in scenarios involving large pretrained models, e.g., ELMo (Peters et al., 2018), retraining can be very expensive in terms of time and resources. Moreover, compression methods are normally applied to large and over-parameterized networks, but this is not necessarily the case in our paper. We consider strongly tuned and regularized state-of-the-art models in their respective tasks, which often already have very compact representations. These circumstances make the compression much more challenging, but more realistic and practically useful.

In this work, we advocate low-rank matrix factorization as an effective post-processing compression method for LSTMs which achieve good performance with guaranteed minimum algorithmic speed compared to other existing techniques. We summarize our contributions as the following:

- We thoroughly explore the limits of several different compression methods (matrix factorization and pruning), including fine-tuning after compression, in Language Modeling, Sentiment Analysis, Textual Entailment, and Question Answering.

- We consistently achieve an average of 1.5x (50% faster) speedup inference time while losing $\sim$1 point in evaluation metric across all datasets by compressing additive and/or multiplicative recurrences in the LSTM gates.

- In PTB, by further fine-tuning very compressed models ($\sim$98%) obtained with both matrix factorization and pruning, we can achieve $\sim$2x (200% faster) speedup inference time while even slightly improving the performance of the uncompressed baseline.

- We discover that matrix factorization performs better in general, additive recurrence is often more important than multiplicative recurrence, and we identify clear and interesting correlations between matrix norms and compression performance.

2 Related Work

The current approaches of model compression are mainly focused on matrix factorization, pruning, and quantization. The effectiveness of these approaches were shown and applied in different modalities. In speech processing, (2008; 2013; 2014; 2014) studied the effectiveness of Non-Matrix Factorization (NMF) on speech enhancement by reducing the noisy speech interference. Matrix factorization-based techniques were also applied in image captioning (Hong et al., 2016; Li et al., 2017) by exploiting the clustering intepretations of NMF. Semi-NMF, proposed by (2010), relaxed the constraints of NMF to allow mixed signs and extend the possibility to be applied in non-negative cases. (2014) proposed a variant of the Semi-NMF to learn low-dimensional representation through a multi-layer structure. (2018) proposed to replace GRUs with low-rank and diagonal weights to enable low-rank parameterization of LSTMs. (2017) modifed LSTM structure by replacing input and hidden weights with two smaller partitions to boost the training and inference time.

On the other hand, compression techniques can also be applied as post-processing steps. (2017) investigated low-rank factorization on standard LSTM model. The Tensor-Train method has been used to train end-to-end high-dimensional sequential video data with LSTM and GRU (Yang et al., 2017; Tjandra et al., 2017). In another line of work, (2016b) explored pruning in order to reduce the number of pa-

rameters in Neural Machine Translation. (2018) proposed to zero out the weights in the network learning blocks to remove insignificant weights of the RNN. Meanwhile, (2018) proposed to binarize LSTM Language Models. Finally, (2016) proposed to use all pruning, quantization, and Huffman coding to the weights on AlexNet.

3 Methodology

3.1 Long-Short Term Memory Networks

Long-Short Term Memory (LSTMs) networks are parameterized with two large matrices, $\mathbf{W}_i$, and $\mathbf{W}_h$. LSTM captures long-term dependencies in the input and avoids the exploding/vanishing gradient problems on the standard RNN. The gating layers control the information flow within the network and decide which information to keep, discard, or update in the memory. The following recurrent equations show the LSTM dynamics:

$$\begin{pmatrix} \mathbf{i}_t \\ \mathbf{f}_t \\ \mathbf{o}_t \\ \hat{\mathbf{c}}_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} \begin{pmatrix} \mathbf{W}_i & \mathbf{W}_h \end{pmatrix} \begin{pmatrix} \mathbf{x}_t \\ \mathbf{h}_{t-1} \end{pmatrix}, \quad (1)$$

$$\mathbf{W}_i = \begin{pmatrix} \mathbf{W}_i^i \\ \mathbf{W}_i^f \\ \mathbf{W}_i^o \\ \mathbf{W}_i^c \end{pmatrix}, \mathbf{W}_h = \begin{pmatrix} \mathbf{W}_h^i \\ \mathbf{W}_h^f \\ \mathbf{W}_h^o \\ \mathbf{W}_h^c \end{pmatrix}, \quad (2)$$

$$\begin{aligned} \mathbf{c}_t &= \mathbf{f}_t \odot \mathbf{c}_{t-1} + \mathbf{i}_t \odot \hat{\mathbf{c}}_t, \\ \mathbf{h}_t &= \mathbf{o}_t \odot \tanh(\mathbf{c}_t). \end{aligned} \quad (3)$$

where $\mathbf{x}_t \in \mathbb{R}^{n_{inp}}$, and $\mathbf{h}_t \in \mathbb{R}^{n_{dim}}$ at time t. Here, $\sigma(\cdot)$ and $\odot$ denote the sigmoid function and element-wise multiplication operator, respectively. The model parameters can be summarized in a compact form with: $\Theta = [\mathbf{W}_i, \mathbf{W}_h]$, where $\mathbf{W}_i \in \mathbb{R}^{4*n_{inp} \times 4*n_{dim}}$ which is the input matrix, and $\mathbf{W}_h \in \mathbb{R}^{4*n_{dim} \times 4*n_{dim}}$ which is the hidden matrix. Note that we often refer $\mathbf{W}_i$ as additive recurrence and $\mathbf{W}_h$ as multiplicative recurrence, following terminology of (2018).

3.2 Low-Rank Matrix Factorization

We consider two Low-Rank Matrix Factorization for LSTM compression: Truncated Singular Value Decomposition (SVD) and Semi Non-negative Matrix Factorization (Semi-NMF). Both methods factorize a matrix $\mathbf{W}$ into two matrices $\mathbf{U}_{m \times r}$ and $\mathbf{V}_{r \times n}$ such that $\mathbf{W} = \mathbf{UV}$ (Fazel, 2002). SVD produces a factorization by applying orthogonal constraints on the $\mathbf{U}$ and $\mathbf{V}$ factors along with an additional diagonal matrix of singular values, where instead Semi-NMF generalizes Non-negative Matrix Factorization (NMF) by relaxing some of the sign constraints on negative values for $\mathbf{U}$ and $\mathbf{W}$. The computation advantage, compared to pruning methods which require a special implementation of sparse matrix multiplication, is that the matrix $\mathbf{W}$ requires mn parameters and mn flops, while $\mathbf{U}$ and $\mathbf{V}$ require $rm+rn = r(m+n)$ parameters and $r(m+n)$ flops. If we take the rank to be very low $r << m, n$, the number of parameters in $\mathbf{U}$ and $\mathbf{V}$ is much smaller compared to $\mathbf{W}$.

As elaborated in Equation 1, a basic LSTM cell includes four gates: input, forget, output, and cell state, performing a linear combination on input at time t and hidden state at time $t-1$. We propose to replace $\mathbf{W}_i$, $\mathbf{W}_h$ pair for each gate with their low-rank decomposition, either SVD or Semi-NMF (Ding et al., 2010), leading to a significant reduction in memory and computational cost requirement. The general objective function is given as:

$$\underset{m \times n}{\mathbf{W}} = \underset{m \times r}{\mathbf{U}} \quad \underset{r \times n}{\mathbf{V}}, \quad (4)$$

$$\underset{U,V}{\text{minimize}} \quad ||\mathbf{W} - \mathbf{UV}||_F^2. \quad (5)$$

3.3 Truncated Singular Value Decomposition (SVD)

One of the constrained matrix factorization method is based on Singular Value Decomposition (SVD) which produces a factorization by applying orthogonal constraints on the $\mathbf{U}$ and $\mathbf{V}$ factors. These approaches aim to find a linear combination of the basis vectors which restrict to the orthogonal vectors in feature space that minimize reconstruction error. In the case of the SVD, there are no restrictions on the signs of $\mathbf{U}$ and $\mathbf{V}$ factors. Moreover, the data matrix $\mathbf{W}$ is also unconstrained.

$$\mathbf{W} = \mathbf{USV}, \quad (6)$$

$$\underset{\mathbf{U},\mathbf{S},\mathbf{V}}{\text{minimize}} \quad ||\mathbf{W} - \mathbf{USV}||_F^2. \quad (7)$$

s.t. $\mathbf{U}$ and $\mathbf{V}$ are orthogonal, and $\mathbf{S}$ is diagonal. The optimal values $\mathbf{U}_{m \times r}^r$, $\mathbf{S}_{r \times r}^r$, $\mathbf{V}_{r \times n}^r$ for $\mathbf{U}_{m \times n}$,

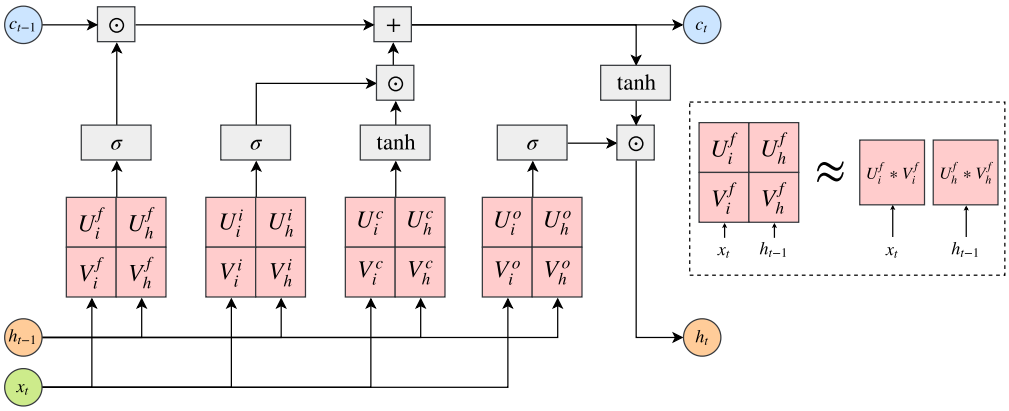


Figure 1: Factorized LSTM Cell

$\mathbf{S}_{n\times n}$, and $\mathbf{V}_{n\times n}$ are obtained by taking the top r singular values from the diagonal matrix $\mathbf{S}$ and the corresponding singular vectors from $\mathbf{U}$ and $\mathbf{V}$.

3.4 Semi-NMF

Semi-NMF generalizes Non-negative Matrix Factorization (NMF) by relaxing some of the sign constraints on negative values for $\mathbf{U}$ and $\mathbf{W}$ ($\mathbf{V}$ has to be kept positive). Semi-NMF is more preferable in application to Neural Networks because of this generic capability of having negative values. To elaborate, when the input matrix $\mathbf{W}$ is unconstrained (i.e., contains mixed signs), we consider a factorization, in which we restrict $\mathbf{V}$ to be non-negative, while having no restriction on the signs of $\mathbf{U}$. We minimize the objective function as in Equation 8.

$$\mathbf{W}_{\pm} \approx \mathbf{U}_{\pm}\mathbf{V}_{+}, \quad (8)$$

$$\underset{\mathbf{U},\mathbf{V}}{\text{minimize}} \quad ||\mathbf{W} - \mathbf{U}\mathbf{V}||_F^2 \quad \text{s.t. } \mathbf{V} \geq 0. \quad (9)$$

The optimization algorithm iteratively alternates between the update of $\mathbf{U}$ and $\mathbf{V}$ using coordinate descent (Luo and Tseng, 1992).

3.5 Pruning

We use the pruning methodology used in LSTMs from (2015) and (2016b). To elaborate, for each weight matrix $\mathbf{W}_{i,h}$, we mask the low-magnitude weights to zero, according to the compression ratio of the low-rank factorization[1].

Table 1: The table shows the total parameters, perplexity, and compression efficiency (lower is better) on PTB Language Modeling task. ‡ We reproduced the results.

PTB	Param.	w/o fine-tuning		w/ fine-tuning	
		PPL	E(r)	PPL	E(r)
AWD-LSTM	24M	58.3‡	-	57.3	-
TT-LSTM	12M	168.6	2.92†	-	-
Semi-NMF $\mathbf{W}_h$ (r=10)	9M	78.5	0.72	58.11	-0.02
SVD $\mathbf{W}_h$ (r=10)	9M	78.07	0.32	58.18	-0.02
Pruning $\mathbf{W}_h$ (r=10)	9M	83.62	0.89	57.94	-0.03
Semi-NMF $\mathbf{W}_h$ (r=400)	18M	59.7	0.05	57.84	-0.02
SVD $\mathbf{W}_h$ (r=400)	18M	**59.34**	**0.006**	57.81	-0.02
Pruning $\mathbf{W}_h$ (r=400)	18M	59.47	0.03	**57.19**	**-0.04**
Semi-NMF $\mathbf{W}_i$ (r=10)	15M	485.4	19.81	81.4	1.04
SVD $\mathbf{W}_i$ (r=10)	15M	462.19	6.83	88.12	1.35
Pruning $\mathbf{W}_i$ (r=10)	15M	676.76	28.69	82.23	1.08
Semi-NMF $\mathbf{W}_i$ (r=400)	20M	62.7	0.42	58.47	-0.01
SVD $\mathbf{W}_i$ (r=400)	20M	60.59	0.02	58.04	-0.01
Pruning $\mathbf{W}_i$ (r=400)	20M	59.62	0.10	57.65	-0.02

4 Evaluation

We evaluate using five different publicly available datasets spanning two domains: 1) Perplexity in two different Language Modeling (LM) datasets, 2) Accuracy/F1 in three downstream NLP tasks that ELMo achieved the state-of-the-art single-model performance. We also report the number of parameters, efficiency $E(r)$ (ratio of loss in performance to parameters compression), and inference time [2] in test set.

[1] We align the pruning rate with the rank with $\frac{r(m+n)}{mn}$.

[2] Using an Intel(R) Xeon(R) CPU E5-2620 v4 @2.10GHz.

Table 2: The table shows the total parameters, perplexity, and compression efficiency (lower is better) on WT-2 Language Modeling task. ‡ We reproduced the results.

WT-2	Params	PPL	E(r)
AWD-LSTM	24M	65.67‡	-
Semi-NMF $\mathbf{W}_h$ (r=10)	9M	102.17	65.14
SVD $\mathbf{W}_h$ (r=10)	9M	99.92	62.49
Pruning $\mathbf{W}_h$ (r=10)	9M	109.16	72.64
Semi-NMF $\mathbf{W}_h$ (r=400)	18M	66.5	4.33
SVD $\mathbf{W}_h$ (r=400)	18M	**66.1**	**2.28**
Pruning $\mathbf{W}_h$ (r=400)	18M	66.23	2.94
Semi-NMF $\mathbf{W}_i$ (r=10)	15M	481.61	197.57
SVD $\mathbf{W}_i$ (r=10)	15M	443.49	194.89
Pruning $\mathbf{W}_i$ (r=10)	15M	856.87	211.23
Semi-NMF $\mathbf{W}_i$ (r=400)	20M	68.41	22.68
SVD $\mathbf{W}_i$ (r=400)	20M	67.11	12.18
Pruning $\mathbf{W}_i$ (r=400)	20M	66.37	5.97

We benchmark the LM capability using Penn Treebank (Marcus et al., 1993, PTB) and WikiText-2 (Merity et al., 2017, WT2). For the downstream NLP tasks, we evaluate our method in the Stanford Question Answering Dataset (Rajpurkar et al., 2016, SQuAD) the Stanford Natural Language Inference (Bowman et al., 2015, SNLI) corpus, and the Stanford Sentiment Treebank (Socher et al., 2013, SST-5) dataset.

For all datasets, we run experiments across different levels of low-rank approximation *r* with Semi-NMF and SVD, averaged over 5 runs, and compare with Pruning with same compression ratio. We also compare the factorization efficiency when only one of $\mathbf{W}_i$ or $\mathbf{W}_h$ was factorized. This is done in order to see which recurrence type (additive or multiplicative) is more suitable for compression.

4.1 Measure

For evaluating the performance of the compression we define efficiency measure as:

$$E(r) = \frac{R(M, M^r)}{R(P, P^r)} \tag{10}$$

where M represent any evaluation metric (i.e. Accuracy, F1-score, Perplexity[3]), P represents the number of parameters[4], and $R(a, b) = \frac{a-b}{a}$ where $a = max(a, b)$, i.e. the ration. This indicator shows the ratio of loss in performance versus the loss in number of parameter. Hence, an efficient compression holds a very small E since the denominator, $P - P^r$, became large just when the number of parameter decreases, and the numerator, $M - M^r$, became small only if there is no loss in the considered measure. In some cases E became negative if there is an improvement.

4.2 Language Modeling (LM)

We train a 3-layer LSTM Language Model proposed by (Merity et al., 2018), following the same training details for both datasets, using their released code [5]. In PTB, we fine-tune the compressed model for several epochs. Table 1 reports the perplexity among different ranks in $\mathbf{W}_{i,h}$. It is clear that compressing $\mathbf{W}_h$ works notably better than $\mathbf{W}_i$. We achieve similar results for WT-2. In general, SVD has the lowest perplexity among others. This difference becomes more evident for higher compression (e.g., r=10). Moreover, all the methods perform better than the result reported by (Grachev et al., 2017) using Tensor Train (TT-LSTM). Using fine-tuning with rank 10 all the methods we achieve a small improvement compared to the baseline with a 2.13x speedup.

4.3 NLP Tasks with ELMo

To highlight the practicality of our proposed method, we also measure the factorization performances with models using pre-trained ELMo (Peters et al., 2018), as ELMo is essentially a 2-layer bidirectional LSTM Language Model that captures rich contextualized representations. Using the same publicly released pre-trained ELMo weights [6] as the input embedding layer of all three tasks, we train publicly available state-of-the-art models as in (Peters et al., 2018): BiDAF (Seo et al., 2016) for *SQuAD*, ESIM (Chen et al., 2017) for *SNLI*, and BCN (McCann et al., 2017) for *SST-5*. Similar to the Language Modeling tasks, we low-rank factorize the pre-trained ELMo layer only, and compare the accuracy and F1 scores across different levels of low-rank approxi-

[3]Note that for Perplexity, we use $R(M^r, M)$ instead, because lower is better.

[4]P^r and M^r are the parameter and the measure after semi-NMF of rank r

[5]https://github.com/salesforce/awd-lstm-lm

[6]https://allennlp.org/elmo

Table 3: The table shows the Accuracy/F1 with ELMo.

SST-5	**r=10**		**r=400**		**Best**	
	Acc.	**E(r)**	**Acc.**	**E(r)**	**Acc. (avg)**	**E(r) (avg)**
BCN	-	-	-	-	53.7‡	-
BCN + ELMo	-	-	-	-	54.5‡	-
Semi-NMF $\mathbf{W}_h$	50.18	0.29	53.93	0.21	54.16 (52.93)	0.09 (0.17)
SVD $\mathbf{W}_h$	50.4	0.27	54.11	0.13	54.11 (52.84)	0.12 (0.17)
Pruning $\mathbf{W}_h$	**50.81**	**0.25**	**54.66**	**-0.03**	**54.88 (53.59)**	**-0.07 (0.06)**
Semi-NMF $\mathbf{W}_i$	38.23	1.1	54.11	0.15	54.11 (50.56)	0.12 (0.34)
SVD $\mathbf{W}_i$	40.58	0.94	54.34	0.05	54.38 (51.19)	0.02 (0.26)
Pruning $\mathbf{W}_i$	34.57	1.35	54.61	-0.01	54.66 (50.01)	-0.02 (0.33)
SNLI	**r=10**		**r=400**		**Best**	
	Acc.	**E(r)**	**Acc.**	**E(r)**	**Acc. (avg)**	**E(r) (avg)**
ESIM	-	-	-	-	88.6	-
ESIM + ELMo	-	-	-	-	88.5‡	-
Semi-NMF $\mathbf{W}_h$	87.24	0.04	88.45	0.01	88.47 (88.18)	0.003 (0.01)
SVD $\mathbf{W}_h$	87.27	0.04	88.46	0.005	88.46 (88.18)	0.003 (0.01)
Pruning $\mathbf{W}_h$	**87.51**	**0.03**	**88.53**	**-0.003**	**88.53 (88.23)**	**-0.003 (0.01)**
Semi-NMF $\mathbf{W}_i$	77.08	0.39	88.44	0.01	88.44 (86.59)	0.01 (0.07)
SVD $\mathbf{W}_i$	78.15	0.35	88.48	0.002	88.48 (86.77)	0.007 (0.06)
Pruning $\mathbf{W}_i$	73.67	0.5	88.48	0.005	88.5 (85.8)	0.001 (0.09)
SQuAD	**r=10**		**r=400**		**Best**	
	F1	**E(r)**	**F1**	**E(r)**	**F1 (avg)**	**E(r) (avg)**
BiDAF	-	-	-	-	77.3‡	-
BiDAF + ELMo	-	-	-	-	81.75‡	-
Semi-NMF $\mathbf{W}_h$	76.59	**0.21**	81.55	0.04	81.55 (80.32)	0.03 (0.07)
SVD $\mathbf{W}_h$	**76.72**	**0.21**	81.62	0.027	81.62 (80.47)	0.02 (0.06)
Pruning $\mathbf{W}_h$	52.02	0.49	81.73	0.006	81.65 (**80.6**)	0.006 (**0.05**)
Semi-NMF $\mathbf{W}_i$	60.69	0.88	**81.78**	**-0.0003**	**81.78** (77.93)	**-0.0003** (0.17)
SVD $\mathbf{W}_i$	57.14	1.03	**81.78**	**-0.0003**	**81.78** (77.69)	**-0.0003** (0.17)
Pruning $\mathbf{W}_i$	52.02	1.24	81.73	0.004	81.73 (76.06)	0.004 (0.25)

mation. Note that although many of these models are based on RNNs, we factorize only the ELMo layer in order to show that our approach can effectively compress pre-trained transferable knowledge. As we only compress the ELMo weights, and other layers of each model also have large number of parameters, the inference time is affected less than in Language Modeling tasks. The percentage of parameters in the ELMo layer for BiDAF (*SQuAD*) is 59.7%, for ESIM (*SNLI*) 67.4%, and for BCN (*SST-5*) 55.3%.

From Table 3, for *SST-5* and *SNLI*, we can see that compressing $\mathbf{W}_h$ is in general more efficient and better performing than compressing $\mathbf{W}_i$, except for SVD in *SST-5*. On the other hand, for the results on SQuAD, Table 3 shows the opposite trend, in which compressing $\mathbf{W}_i$ constantly outperforms compressing $\mathbf{W}_h$ for all methods we experimented with. In fact, we can see that, in average, using highly compressed ELMo with BiDAF still performs better than without. Overall, we can see that for all datasets, we achieve performances that are not significantly different from the baseline results even after compressing over more than 10M parameters.

4.4 Norm Analysis

In the previous section, we observe two interesting points: 1) Matrix Factorization (MF) works consistently better in PTB and Wiki-Text 2, but Pruning works better in ELMo for $\mathbf{W}_h$, 2) Factorizing $\mathbf{W}_h$ is generally better than factorizing $\mathbf{W}_i$. To answer these questions, we collect the L1 norm and Nuclear norm statistics, defined in Figure 2, comparing among $\mathbf{W}_h$ and $\mathbf{W}_i$ for both PTB and ELMo. L1 and its standard deviation (*std*) together describe the sparsity of a matrix, and Nuclear norm approximates the matrix rank.

MF versus Pruning in $\mathbf{W}_i$ From the results, we observe that MF performs better than Pruning in compressing $\mathbf{W}_i$ for high compression ratios. Figure 2 shows rank r versus L1 norm and its standard

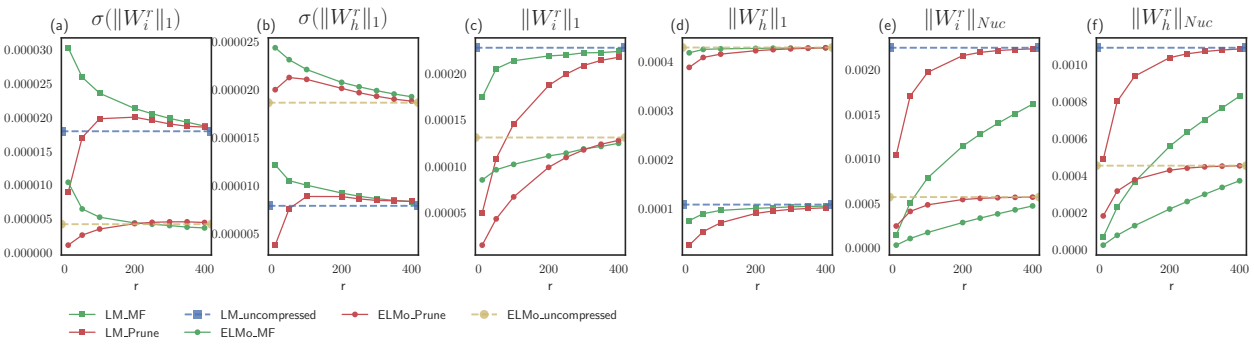


Figure 2: Norm analysis comparisons between MF and Pruning in Language Modeling (PTB) and ELMo. Rank versus (a) $\sigma(\|\mathbf{W}_i\|_1)$ (b) $\sigma(\|\mathbf{W}_h\|_1)$ (c) $\|\mathbf{W}_i\|_1$ (d) $\|\mathbf{W}_h\|_1$ (e) $\|\mathbf{W}_i\|_{Nuc}$ (f) $\|\mathbf{W}_h\|_{Nuc}$.

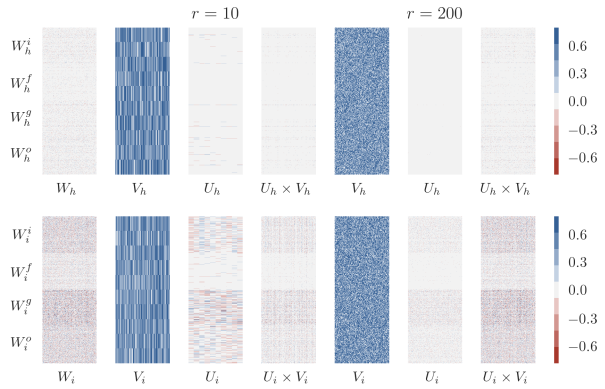


Figure 3: Heatmap LSTM weights on PTB.

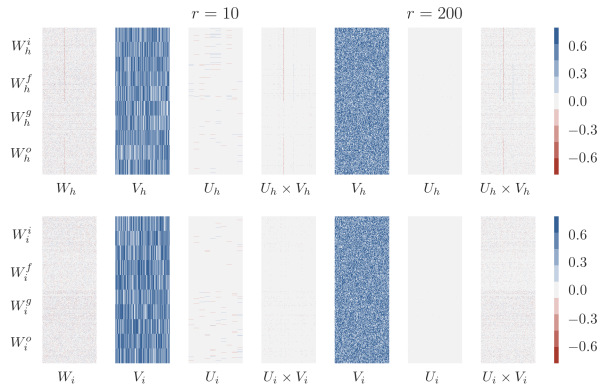


Figure 4: Heatmap of ELMo forward weights.

deviation, in both PTB and ELMo. The first notable pattern from Figure 2 Panel (a) is that MF and Pruning have diverging values from $r \leq 200$. We can see that Pruning makes the *std* of L1 lower than the uncompressed, while MF monotonically increases the *std* from uncompressed baseline. This means that as we approximate to lower ranks ($r \leq 200$), MF retains more salient information, while Pruning loses some of that salient information. This can be clearly shown from Panel (c), in which Pruning always drops significantly more in L1 than MF does.

MF versus Pruning in $\mathbf{W}_h$ The results for $\mathbf{W}_h$ are also consistent in both PTB and WT2; MF works better than Pruning for higher compression ratios. On the other hand, results from Table 3 show that Pruning works better than MF in $\mathbf{W}_h$ of ELMo even in higher compression ratios.

We can see from Panel (d) that L1 norms of MF and Pruning do not significantly deviate nor decrease much from the uncompressed baseline. Meanwhile, Panel (b) reveals an interesting pattern, in which the *std* actually increases for Pruning and is always kept above the uncompressed baseline. This means that Pruning retains salient information for $\mathbf{W}_h$, while keeping the matrix sparse.

This behavior of $\mathbf{W}_h$ can be explained by the nature of the compression and with inherent matrix sparsity. In this setting, pruning is zeroing values already close to zero, so it is able to keep the L1 stable while increasing the *std*. On the other hand, MF instead reduces noise by pushing lower values to be even lower (or zero) and keeps salient information by pushing larger values to be even larger. This pattern is more evident in Figure 3 and Figure 4, in which you can see a clear salient red line in $\mathbf{W}_h$ that gets even stronger after factorization ($\mathbf{U}_h \times \mathbf{V}_h$). Naturally, when the compression rate is low (e.g., r=300) pruning is more efficient strategy then MF.

$\mathbf{W}_i$ versus $\mathbf{W}_h$ We show the change in Nuclear norm and their corresponding starting points (i.e., uncompressed) in Figure 2 Panels (e) and (f). Notably, $\mathbf{W}_h$ has a consistently lower nuclear norm in both tasks compared to $\mathbf{W}_i$. This difference is larger for LM (PTB), in which $\|\mathbf{W}_i\|_{Nuc}$ is twice of that of $\|\mathbf{W}_h\|_{Nuc}$. By definition, having a lower nuclear norm is often an indicator of low-rank in a matrix; hence, we hypothesize that $\mathbf{W}_h$ is inherently low-rank than $\mathbf{W}_i$. We confirm this from Panel (d), in which even with a very high compression ratio (e.g., $r = 10$), the L1 norm does not decrease that much. This explains the large gap in performance between the compression of $\mathbf{W}_i$ and $\mathbf{W}_h$. On the other hand, in ELMo, this gap in norm is lower and also shows smaller differences in performance between $\mathbf{W}_i$ and $\mathbf{W}_h$, and also sometimes even the opposite in SQuAD. Hence, we believe that smaller nuclear norms lead to better performance for all compression methods.

5 Conclusion

In conclusion, we empirically verified the limits of compressing LSTM gates using low-rank matrix factorization and pruning in four different NLP tasks. Our experiment results and norm analysis show that Low-Rank Matrix Factorization works better in general than pruning, except for particularly sparse matrices. We also discover that inherent low-rankness and low nuclear norm correlate well, explaining why compressing multiplicative recurrence works better than compressing additive recurrence. In future works, we plan to factorize all LSTMs in the model, e.g. BiDAF model, and try to combine both Pruning and Matrix Factorization.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Francois Belletti, Alex Beutel, Sagar Jain, and Ed Chi. 2018. Factorized recurrent neural architectures for longer range dependence. In *International Conference on Artificial Intelligence and Statistics*, pages 1522–1530.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. 2015. A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Qian Chen, Xiaodan Zhu, Zhen-Hua Ling, Si Wei, Hui Jiang, and Diana Inkpen. 2017. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1657–1668.

Misha Denil, Babak Shakibi, Laurent Dinh, Nando De Freitas, et al. 2013. Predicting parameters in deep learning. In *Advances in neural information processing systems*, pages 2148–2156.

Chris HQ Ding, Tao Li, and Michael I Jordan. 2010. Convex and semi-nonnegative matrix factorizations. *IEEE transactions on pattern analysis and machine intelligence*, 32(1):45–55.

Hao-Teng Fan, Jeih-weih Hung, Xugang Lu, Syu-Siang Wang, and Yu Tsao. 2014. Speech enhancement using segmental nonnegative matrix factorization. In *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, pages 4483–4487. IEEE.

Maryam Fazel. 2002. *Matrix rank minimization with applications*. Ph.D. thesis, PhD thesis, Stanford University.

Jürgen T Geiger, Jort F Gemmeke, Björn Schuller, and Gerhard Rigoll. 2014. Investigating nmf speech enhancement for neural network based acoustic models. In *Proc. INTERSPEECH 2014, ISCA, Singapore, Singapore*.

Felix A. Gers, Jürgen A. Schmidhuber, and Fred A. Cummins. 2000. Learning to forget: Continual prediction with lstm. *Neural Comput.*, 12(10):2451–2471, October.

Yunchao Gong, Liu Liu, Ming Yang, and Lubomir Bourdev. 2014. Compressing deep convolutional networks using vector quantization. *arXiv preprint arXiv:1412.6115*.

Artem M Grachev, Dmitry I Ignatov, and Andrey V Savchenko. 2017. Neural networks compression for language modeling. In *International Conference on Pattern Recognition and Machine Intelligence*, pages 351–357. Springer.

Song Han, Jeff Pool, John Tran, and William Dally. 2015. Learning both weights and connections for efficient neural network. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems 28*, pages 1135–1143. Curran Associates, Inc.

Song Han, Huizi Mao, and William J Dally. 2016. Deep compression: Compressing deep neural networks with

pruning, trained quantization and huffman coding. *ICLR*.

Luheng He, Kenton Lee, Mike Lewis, and Luke Zettlemoyer. 2017. Deep semantic role labeling: What works and whats next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 473–483.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the knowledge in a neural network. *stat*, 1050:9.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Seunghoon Hong, Jonghyun Choi, Jan Feyereisl, Bohyung Han, and Larry S Davis. 2016. Joint image clustering and labeling by matrix factorization. *IEEE transactions on pattern analysis and machine intelligence*, 38(7):1411–1424.

Qiangui Huang, Kevin Zhou, Suya You, and Ulrich Neumann. 2018. Learning to prune filters in convolutional neural networks. *arXiv preprint arXiv:1801.07365*.

Max Jaderberg, Andrea Vedaldi, and Andrew Zisserman. 2014. Speeding up convolutional neural networks with low rank expansions. In *Proceedings of the British Machine Vision Conference. BMVA Press*.

Caiming Xiong James Bradbury, Stephen Merity and Richard Socher. 2017. Quasi-recurrent neural networks. In *International Conference on Learning Representations*.

Oleksii Kuchaiev and Boris Ginsburg. 2017. Factorization tricks for lstm networks. *ICLR Workshop*.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197.

Omer Levy, Kenton Lee, Nicholas FitzGerald, and Luke Zettlemoyer. 2018. Long short-term memory as a dynamically computed element-wise weighted sum. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 732–739. Association for Computational Linguistics.

Xuelong Li, Guosheng Cui, and Yongsheng Dong. 2017. Graph regularized non-negative low-rank matrix factorization for image clustering. *IEEE transactions on cybernetics*, 47(11):3840–3853.

Xuan Liu, Di Cao, and Kai Yu. 2018. Binarized lstm language model. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2113–2121. Association for Computational Linguistics.

Zhiyun Lu, Vikas Sindhwani, and Tara N Sainath. 2016. Learning compact recurrent neural networks. In *Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*, pages 5960–5964. IEEE.

Zhi-Quan Luo and Paul Tseng. 1992. On the convergence of the coordinate descent method for convex differentiable minimization. *Journal of Optimization Theory and Applications*, 72(1):7–35.

Mitchell P. Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of english: The penn treebank. *Comput. Linguist.*, 19(2):313–330, June.

Bryan McCann, James Bradbury, Caiming Xiong, and Richard Socher. 2017. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems*, pages 6294–6305.

Gbor Melis, Chris Dyer, and Phil Blunsom. 2018. On the state of the art of evaluation in neural language models. In *International Conference on Learning Representations*.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. 2017. Pointer sentinel mixture models. *ICLR*.

Stephen Merity, Nitish Shirish Keskar, and Richard Socher. 2018. Regularizing and optimizing LSTM language models. In *International Conference on Learning Representations*.

Antonio Valerio Miceli Barone. 2018. Low-rank passthrough neural networks. In *Proceedings of the Workshop on Deep Learning Approaches for Low-Resource NLP*, pages 77–86. Association for Computational Linguistics.

Tomáš Mikolov. 2012. Statistical language models based on neural networks. *Presentation at Google, Mountain View, 2nd April*.

Nasser Mohammadiha, Paris Smaragdis, and Arne Leijon. 2013. Supervised and unsupervised speech enhancement using nonnegative matrix factorization. *IEEE Transactions on Audio, Speech, and Language Processing*, 21(10):2140–2151.

Ivan V Oseledets. 2011. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for

machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392. Association for Computational Linguistics.

Abigail See, Minh-Thang Luong, and Christopher D Manning. 2016a. Compression of neural machine translation models via pruning. *CoNLL 2016*, page 291.

Abigail See, Minh-Thang Luong, and Christopher D. Manning. 2016b. Compression of neural machine translation models via pruning. In *Proceedings of The 20th SIGNLL Conference on Computational Natural Language Learning*, pages 291–301. Association for Computational Linguistics.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. 2016. Bidirectional attention flow for machine comprehension. *ICLR 2017*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642. Association for Computational Linguistics.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. Lstm neural networks for language modeling. In *Thirteenth Annual Conference of the International Speech Communication Association*.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Andros Tjandra, Sakriani Sakti, and Satoshi Nakamura. 2017. Compressing recurrent neural network with tensor train. In *Neural Networks (IJCNN), 2017 International Joint Conference on*, pages 4451–4458. IEEE.

George Trigeorgis, Konstantinos Bousmalis, Stefanos Zafeiriou, and Bjoern Schuller. 2014. A deep semi-nmf model for learning hidden representations. In *International Conference on Machine Learning*, pages 1692–1700.

Wei Wen, Yuxiong He, Samyam Rajbhandari, Minjia Zhang, Wenhan Wang, Fang Liu, Bin Hu, Yiran Chen, and Hai Li. 2018. Learning intrinsic sparse structures within long short-term memory. In *International Conference on Learning Representations*.

Kevin W Wilson, Bhiksha Raj, Paris Smaragdis, and Ajay Divakaran. 2008. Speech denoising using nonnegative matrix factorization with priors. In *Acoustics, Speech and Signal Processing, 2008. ICASSP 2008. IEEE International Conference on*, pages 4029–4032. IEEE.

Yinchong Yang, Denis Krompass, and Volker Tresp. 2017. Tensor-train recurrent neural networks for video classification. In *International Conference on Machine Learning*, pages 3891–3900.

Prospective Result of Causative Predicates: A Uniform Analysis

Yusuke Yagi
Waseda University
usk.kmksw.10.oct@toki.waseda.jp

Abstract

This study is devoted to pursuing a principled organization of the lexicon. It aims at explaining one aspect of causative predicates from semantic perspective. It has been noticed that some of the causative predicates have a modal component, making the caused eventuality "prospective" - whether or not the caused event/state holds in the actual world is undetermined. However, exactly which predicate contains modality and which does not has been a mystery, and even has not been discussed as far as I know. I will describe how the prospectivity is determined and argue that no idiosyncrasy is involved here.

1 Introduction

One remarkable fact about human language is that it takes only two or three years for children to acquire the basics of the grammar and vocabulary of their native language. That is, language learning requires only a small number of experiences. The smallness of experiences children need has inspired the work on the theory of Generative Grammar. It argues that human beings are born with Universal Grammar, the set of knowledge shared by every human language. Once UG is properly defined, it should predict a considerable number of the grammatical properties of human language. Linguists in the generative enterprise have struggled to determine exactly what constitutes UG. They pursue a theory with maximum explanatory power with minimum necessity of experiences.

However, there is an area in which necessity of experiences is irrefutable: The lexicon. As a mental dictionary, the lexicon consists of the words and morphemes of a language and their syntactic, phonetic, morphologic, and other idiosyncratic information. Since it would not make any sense to assume that, for example, an English word *dog* is a universal word to refer the certain animal, acquiring the lexicon definitely calls for experiences. In fact, since the starting point of the generative enterprise, the linguists has tended to regard the lexicon as a list of idiosyncrasies: A list of unpredictable facts which children must learn and memorize through experiences, for which no theory or even generalization is possible.

The above tendency is not totally unreasonable. Since the lexicon is finite, memorizing all the list is not logically impossible. Nevertheless, more recent study emphasizes that work on the lexicon should be "guided by the perception that there are generalizations relating apparently distinct items, which could not be simply accidental," (Reinhart, 2000) and we should "proceed from the null hypothesis that *nothing* is acquired through experience" (Pesetsky, 1995). They show that seemingly random morphological or theta-theoretic alternations are indeed predictable by their syntactic properties (Pesetsky) or a general theory of theta system (Reinhart).

Following their intuition, this study is devoted to pursuing a more principled organization of the lexicon. It aims at explaining one aspect of causative predicates from semantic perspective. It has been noticed that some of the causative predicates have a modal component, making the caused eventuality "prospective" - whether or not the caused event/state holds in the actual world is undetermined. (Koenig and Davis, 2001; Beavers,

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 263-270, Hakodate, Japan, September 13-15, 2019

2011; Martin, 2015; Martin and Schäfer, 2017; Harley and Jung, 2015, among others). However, exactly which predicate contains modality has been a mystery, and even has not been discussed as far as I know. I will describe how the prospectivity is determined and argue that no idiosyncrasy is involved here.

1.1 Theoretical Background

There are two major assumptions underlying this study. Firstly, I assume that verbs are represented with a Lexical Conceptual Structure, or LCS (Levin and Rappaport Hovav, 2011; Rappaport Hovav and Levin 2010), and this structure is formally analyzed by Neo-Davidsonian event semantics (Parsons, 1990).

An LCS is a decomposed structure of verbs. It is composed of a limited set of primitive predicates, like **act**, **become**, **have**, and **cause**. The idiosyncratic component of a verb, called *a verbal root*, may be associated with LCS in two ways, either it modifies a primitive predicate, or it becomes an argument of a primitive predicate. Typically, the former case is for an action verb like *run* represented as (1a), whereas the latter case is for a change of state verb like a transitive use of *break*, represented in (1b). (The variables *x* and *y* represent participants of the event.) Combined with the event semantics, the verbs in (1) have the denotations (2).

(1) a. *run* = [*x* **act**$_{\langle run \rangle}$]
b. *break* = [x **act**] **cause** [y **become** <break>]

(2) a. [[*run*]] = λx.λe. **act**(*e*) & **run**(*e*) & **subject**(*e*, *x*)
b. [[*break*]] = λy.λx.λe. **act**(*e*) & **subject**(*e*, *x*) & ∃e'[**cause**(*e*, *e'*) & **break**(*e'*) & **theme**(*e'*, *y*)][1]

The denotation (2b) is closely related with my second assumption. I define *a causative predicate* to be a semantically complex predicate which contains *two* eventualities: causing and caused eventualities. In short, a causative predicate is *bi-eventive*.

The bieventivity is tested with an adverb *again* (Dowty, 1979). If a verb has a bieventieve structure, it induces a scopal ambiguity in the interpretation. To see this, consider the following sentence.

(3) a. John opened the door again.
b. ∃e**[act**(*e*) & **subject**(*e*, J) & ∃e'[**cause**(*e*, *e'*) & **open**(*e'*) & **theme**(*e'*, the-door)]]
c. *again*(∃e**[act**(*e*) & **subject**(*e*, J) & ∃e'[**cause**(*e*, *e'*) & **open**(*e'*) & **theme**(*e'*, the-door)]])
d. ∃e**[act**(*e*) & **subject**(*e*, J) & *again*(∃e'[**cause**(*e*, *e'*) & **open**(*e'*) & **theme**(*e'*, the-door)])]

The transitive use of *open* is a typical instance of causative predicates. The sentence (3), ignoring *again*, has the denotation (3b). Notice that (3a) has two interpretations. It means either John caused the door open and he had opened it before, or John caused the door open and it had been open before (not necessary opened by John). *Again* modifies the whole sentence in the former interpretation, while in the latter case it modifies only the resultant state. Each interpretation has the denotation (3c) and (3d), respectively. This ambiguity is absent in a mono-eventive construction such as *John hit Mary again*. Hence, *hit* is not a causative predicate.

With the above assumptions in mind, I will pursue the theory of causative predicates which requires the minimal amount of experiences. The rest of this study is organized as follows. In the section 2 I will lay out the relevant data and detect a generalization. I will formalize it in section 3. An important implication for Manner/Result Complementarity will also be discussed there. The section 4 deals with the data which apparently poses a problem to the proposal. In section 5, I will extend the analysis to a peculiar class of verbs, namely the defeasible causative verbs. The section 6 concludes this paper.

[1] There are lot of ways to represent a change of state verb. Since the present proposal does not hinge on any specific representation, I do not commit which of them is licit.

2 Prospectivity

2.1 The Data

For some of the causative verbs, the resultant eventuality is only prospective: Whether it happens in the actual world or not is undetermined. Take ditransitive predicates, for example. Many authors have proposed that ditransitive predicates have an underlying causal relation, with a resultant state "a recipient has a theme" (Pesetsky, 1995; Harley, 2002; Harley and Jung, 2015; Beavers, 2011). As Pylkkänen (2008) and Beavers (2011) observe, whether the resultant state is entailed or not depends on the verb. Consider the following sentences and their LCS.

(4) a. John gave Mary a ball, #but she never received/got it.
b. [[John **act**] **cause** [Mary **have** a ball]]

(5) a. John threw Mary a ball, but she never received/got it.
b. [[John **act**$_{<throw>}$] **cause** [Mary **have** a ball]]

Though g*ive* and *throw* have the same LCS except for contribution of the verbal root, they show an interesting difference in an entailment pattern. G*ive* entails the resultant state *[Mary **have** a ball]* and negating that state (*but...*) leads to contradiction. On the other hand, for *throw* the resultant state is only prospective and negating it raises no contradiction. The result of *throw* may or may not happen in the actual world.

Since Koenig and Davis (2001), it is common to assume that verbs with a prospective result have a sublexical modal component. According to this analysis, the LCS of *throw* is represented as (6), with the resultant state being under the scope of a modal operator ◊ (Beavers, 2011).

(6) [[John **act**$_{<throw>}$] ◊**cause** [Mary **have** a ball]]

Although this resolution is widespread and assumed by many authors (Beavers, 2011; Martin, 2015; Martin and Schäfer, 2017; Harley and Jung, 2015), a question much less frequently addressed is which verb has a modal component. Since the lexicon is finite, it may not be totally unreasonable to conclude that the presence or absence of modality is idiosyncratically determined and we have to memorize this. However, as pointed out in the previous section, we should start a linguistic enterprise with the null hypothesis that nothing requires experiences. Below, I will argue that the prospectivity of causative verbs is actually predictable. In the rest of this section I will lay out the relevant data from various kinds of causative predicates.

2.1.1. Lexical Causative Verbs

Pylkkänen (2008) observes that lexical causative verbs in English always entail a resultant state.

(7) a. #I flew the kite over the field but it didn't fly.
b. #I broke the vase but it didn't break.
c. #I cooked the meat but it didn't cook.
(Pylkkänen, 2008: 15)

2.1.2. Periphrastic causative verbs

Kartunnen (1971) claims that *make* entails the resultant event. On the other hand, Jackendoff (1990) observes verbs like *urge, goad, pressure* do not carry the entailment. These verbs can be used with *unsuccessfully*, and negating their resultant event does not lead to contradiction.[2]

(8) a. John made Mary leave, #but Mary didn't leave.
b. Harry pressured/urged/goaded Sam to go away, but he didn't go away.
c. Harry unsuccessfully urged/pressured/goad Sam to leave.

2.1.3. Ditransitive Verbs

As observed above, *give* entails the resultant state (possession) while *throw* does not. Pylkkänen

[2] One may think that a non-manner counterpart of *urge*/*goad*/*pressure* is *force*. Indeed, Jackendoff and Kartunnen claim that *force* have the result entailment. However, literature has reported contradictory judgements on this. Koenig and Davis (2001) and Martin (2018) claim that the result of *force* can be negated (at least certain circumstances). In order to avoid complexity, I leave the analysis of *force* for another occasion.

(2008) further observes that *write* and *send* do not entail the resultant state.

(9) a. I sent Bill the letter but he never got it.
b. I wrote Sue a letter but she never got it.
(Pylkkänen, 2008: 15)

2.2 Generalization

With the set of data above, we can now detect a generalization about presence/absence of the result entailment, namely (10).

(10) If a manner of the causing event is specified in the lexicon, then the resultant event/state becomes prospective. If not, the resultant event/state is obtained in the actual world.

In order to see what (10) is supposed to mean, consider again the ditransitive verbs. *Give*, a verb with the result entailment, does not specify any manner on its causing event. Any event that causes the recipient to have the theme can be a causing event of this verb. On the other hand, *throw*, *send*, and *write* require a certain manner. In a sentence *I threw Mary a ball (and she got it)*, the causing event must be implemented in throwing manner. The same requirement exists in *send* and *write*.

The generalization (10) can be extended to the other group of verbs. The lexical causative verbs, which always entail their resultant state, have no specification on a manner of the causing event. *Make* does not require any manner on the causing event, while *urge*, *goad*, and *pressure* do. For example, *urge* requires its causing event to be implemented by verbal recommendation or persuasion. As predicted, among these verbs only *make* carries the result entailment.

3 Proposal

In this section I will formalize the generalization (10) with the proposal summarized in (11).

(11) a. All causative predicates have a modal component.
b. The generalization (10) follows from the characterization of a modal base and an ordering source of causative predicates.

(11a) states that a modal component is a universal property of causative predicates. Thus, there is no such variation that some verbs introduce modality while others do not. I state a general definition of causative predicates as in (12)[3].

(12) Let φ be a causative predicate with a resultant event (or state) ψ. Then,
[[φ]] = λx.λy.λe.λw. **act**(*e*, *y*, *w*) & ∀w'∈$\max_{g(e)}$(∩*f*(*e*)) [∃e'**cause**(*e*, *e'*, *w*') & ψ(*e'*, *x*, *w'*)], where
a. *f*(*e*) is a circumstantial modal base:
f(*e*) = {p | p is a proposition denoting the laws of nature and other relevant facts of the world where *e* happens}
b. *g(e)* is an ordering source: *g(e)* = {q| q is a proposition denoting the norms inherently associated with *e*}
c. $max_{g(e)}$ selects the most ideal world(s) from ∩*f(e)*, given the ordering source *g(e)*.

The types of the modal base and the ordering source are lexically specified, not provided from a context. I will argue that when no norms are inherently associated with *e*, the effect of the ordering source and $\max_{g(e)}$ becomes vacuous. Below, I describe how (12) works and how it derives the generalization (10).

Consider first a construction with *throw*, a ditransitive verb with a manner specification on the causing event.

(13) a. John threw Mary a ball (but she never got it).
b. ∃e **act**(*e*, John, w_0) & *throwing*(e) & ∀w'∈$\max_{g(e)}$(∩*f* (*e*)) [∃e' **cause**(*e*, *e'*, *w*') & **have**(*e'*, Mary, a ball, *w'*)]

Throw does not entail the resultant possession. This is due to the effect of the ordering source. Since the verb has a manner specification, *g(e)* contains propositions denoting the norms associated with the manner, e.g. [[*the agent throws with a proper form*]], [[*the agent put enough amount of energy*]],

[3] The definitions of *f*(e) and *g*(e) are based on Kratzer (2013). The definition of *max* operator is based on Hacquard (2011). Following Hacquard (2006, 2010), I assume that a modal base and an ordering source take an event argument.

etc. Given *g(e)*, the resultant state *e'* happens in all of the ideal worlds. Since the actual world may not be such an ideal world (i.e. w_0 may not be contained in $\max_{g(e)}(\cap f(e))$, the resultant state is not entailed in the actual world. The same reasoning applies to *send*, *write*, *urge*, *goad*, and *pressure*.

Turning to cases where the resultant state is obtained, consider a sentence with a lexical causative verb *break*.

(14) a. John broke the window.
b. ∃e **act**(*e*, John, w_0) & ∀w'∈$\max_{g(e)}$(∩*f*(*e*))
[∃e' **cause**(*e*, *e'*, *w*') & **break**(*e'*, the window, *w'*)]

Recall that lexical causative verbs always entail their resultant state. Thus, in (14b), the broken state of *the window* must be obtained in the actual world (w_0). Actually, this is exactly what (12) predicts. Notice that ∩*f*(*e*) always contains the actual world: Since *f*(*e*) is a circumstantial, realistic modal base, for all the propositions *p* in *f*(e), w_0∈[[*p*]]. Moreover, lexical causative verbs do not have any specification on a manner of the causing event (*e*), so in (14b) the effect of $\max_{g(e)}$ is vacuous. Thus, (14b) just requires that the resultant state *e'* is caused by *e* in all the worlds contained in ∩*f*(*e*). Since ∩*f*(*e*) contains the actual world, the resultant state is correctly entailed. *Give* and *make* entail the resultant state/event by the same reasoning.

Summarizing the proposal, the definition (12) derives the property of causative predicates discussed in this study. My proposal has at least two advantages. First, we can treat causative predicates uniformly by stating that they all introduce a modal component. Second, presence/absence of the result entailment is predictable from the property of the causing event. In the next subsection I will argue that as a consequence of the proposal we can derive Manner/Result Complementarity.

3.1 Manner/Result Complementarity

One of the most influential and widely shared constraints in lexical semantics is Manner/Result Complementarity (Rappaport Hovav and Levin, 1998, 2010). This constraint banns a verbal root to specify *both* manner and result. Although I believe this constraint is real and on the right track, why there is such a constraint is not frequently discussed: It is just a stipulated statement.

The present proposal offers an answer to the question. Imagine that there is a verb which specifies both manner and result. Since that verb necessarily has a causative component ("result" cannot be defined without it), the verb has the modal base and the ordering source proposed here. The manner specification makes the result prospective by the same reasoning described above, so no specific result is entailed. Thus, even if the verb specifies a result as well as a manner, that cannot be observed in the entailment pattern and the verb seems to specify only a manner. In short, Manner/Result Complementarity is an illusion caused by the modal component.

4 Discussion

In this section I will discuss two sets of data. One is about verbs *hand* and *pass*, which at first sight seems to be a counterexample of the present analysis. The other one is about a verb *force*, for which literature have reported contradictory judgements.

4.1 *hand, pass (the salt)*

Beavers (2011) points out that the resultant possession is entailed with *hand* and *pass (the salt)* [but not *pass (the ball)*. See below.]

(15) #John handed Mary the salt, but he dropped it before she got it.

Since *hand* and *pass (the salt)* clearly encode a manner of the causing event, this data seems to pose a problem to my proposal.

However, the judgement is not that clear-cut. Christopher Tancredi (p.c.) notes that a sentence like *John handed Mary a book, but she refused to take it* is acceptable. Thus, I argue here that *hand* and *pass (the salt)* basically get the same analysis as the one given to *throw*: their resultant state is prospective. The strong result implicature comes from the nature of their manner. As Beavers himself points out, these predicates "necessarily involve two people in close proximity […] in such events it is unlikely there would be a failure of

transfer." (p. 30) Thus, the successful possession in (15) comes not from logical entailment, but from pragmatic inference generated due to the nature of the manner of *hand*.

This argument is supported by the fact I briefly mentioned above: Although *pass (the salt)* seems to imply the resultant state, *pass (the ball)* does not. This is because, I argue, one is more likely to stand in close proximity to the recipient when s/he passes the salt than when s/he passes the ball. Again, the resultant sate of *pass* is prospective as my proposal predicts, but the result is strongly inferred by the nature of the manner.[4]

4.2 Speaker Variation

As noted in the footnote 2, Jackendoff (1990) observes that *force* carries the result entailment. The same observation is made in Kartunnen (1971). However, Koenig and Davis (2001) and Martin (2018) note that *force* does not entail any resultant state, which apparently pose a problem to my proposal. Why do the judgements differ like this?

Note here that in principle lexical information can vary from speaker to speaker. More specifically, some may have a different definition of *force* than other people. Of course, it is not desirable to assume lexical meaning can differ drastically – if *force* had a meaning assumed in the previous section to some speakers while to others it has a meaning of *prevent*, then the communication would be entirely impossible. However, assuming minor variation among speakers is not implausible. Thus, in order to account for the contradictory observation mentioned above, I argue that force specifies a manner of the causing event in some speakers' mind, but not in others'. For instance, one may believe that *force* must involve a direct verbal order in the causing event. This is just a minor change, but it is enough to activate the ordering source and to make the resultant state prospective.

[4] Another possible explanation is that *hand* specifies an instrument, not a manner (Akira Watanabe, p.c.). As for *pass the salt*, Ayaka Sugawara (p.c.) points out that the construction is so idiomatic that it loses the prospectivity. I leave for future research an investigation on whether these proposals are valid.

5 Extension: Defeasible Causative Verbs

5.1 General Account

In this section I will extend the present proposal to a rather peculiar group of causative predicates, called *defeasible causative verbs* (Martin, 2015; Martin and Schäfer 2017). The peculiarity of these verbs can be seen in the contrast observed in the following sentences.

(16) a. Hans schmeichelte Maria, aber sie fühlte
Hans flattered Marie, but she felt
sich überhaupt nicht geschmeichelt.
REFL absolutely NEG flattered.
'John flattered Mary, but she felt absolutely not flattered.'

b. #Dieses Detail schmeichelte Maria,
This detail flattered Marie,
aber sie fühlte sich überhaupt nicht
but she felt REFL absolutely NEG
geschmeichelt.
flattered.
'This detail flattered Mary, but she felt absolutely not flattered.'
(German, Martin and Schäfer, 2017: 88)

Notice that(16a) and (16b) differ in agentivity of the subjects. When the verb *schmeicheln* 'flatter' takes a non-agentive subject as in (16b), the sentence entails Mary got flattered. Negating this entailment leads to contradiction. On the other hand, when the verb takes an agentive subject, no entailment exists so the whole sentence (16a) is felicitous. This property is cross-linguistically observed in verbs like *teach*, *offer*, and *discourage*, among many others (see Martin and Schäfer, 2017; Kratzer, 2013).

How does the proposal so far account for this contrast? Recall that in my analysis the resultant event/state is entailed when the effect of the ordering source *g(e)* is vacuous. Then, we have to assume that defeasible causative verbs deactivate the ordering source when they take a non-agentive subject. How can we formalize this?

Martin (2015) points out that we cannot talk about a manner of a non-agentive, inanimate subject. She notes "we do not differentiate the wind that

may blow out a fire from the wind that may close a door through distinctive features: All these winds are undifferentiated for us." (p. 257)[5] I argue with her that an inanimate subject cannot manifest a manner of action. I further argue that the defeasible causative verbs are special in that they have two possible LCS representations, one with a manner specification and the other without it. Then, we can predict that a non-agentive subject will nullify the effect of an ordering source. In order to see this, consider the following contrast and the LCSs. Again, ◊ represents the modal component with the modal base and the ordering source characterized above.

(17) a. Ivan taught me Russian, but I did not learn anything.
b. Lipson's textbook taught me Russian, #but I did not learn anything.
(Martin and Schäfer, 2017: 87)

(18) a. [x **act**$_{\text{<teach>}}$] ◊**cause** [y <learn> z]
b. [x **act**] ◊**cause** [y <learn> z]

(17a) does not tell us anything new. The ordering source $g(e)$ contains, among others, propositions like [[*a teacher acts on students effectively*]], [[*a teacher's knowledge on the subject is enough*]], etc., the norms associated with the manner of teaching. Since the actual world may not be an ideal world, the sentence does not carry the entailment.

The argument that an inanimate subject cannot manifest a manner of action is crucial when we analyze (17b). Since *Lipson's textbook* cannot manifest a manner of <teach>, it is incompatible with the LCS (18a). Thus, it obligatorily enters the alternative LCS (18b). Since this entry does not have a manner specification, it correctly entails the resultant learning event.[6] The reasoning for the defeasible causative verbs is summarized in (19).

(19) a. Defeasible causative verbs are special in that they are associated with two possible LCSs. One specifies a manner of the causing event the other does not.
b. Since inanimate subjects are generally incompatible with a manner-specified LCS, they have no choice but to enter the manner-less version of LCS. As a result, the predicate entails the resultant eventuality.

In this section I showed that the peculiar behavior of the defeasible causative verbs is predicted correctly by the present proposal. The presence of the result entailment with a non-agentive subject naturally follows from the interaction between the property of the subject and manner specification.

6 Conclusion

In this study, I argued that all causative predicates introduce a modal component. This analysis enables us to treat all causative predicates uniformly and to predict which verb has the prospectivity. I also showed that the present analysis can be extended to the defeasible causative verbs. Overall, the proposal demonstrates that the prospectivety and the defeasibility is not an idiosyncratic property. Rather, they are closely related with a linguistically real notion, namely a manner specification. These properties are actually predictable based on this notion. Thus, the present study contributes to the generative enterprise which aims at minimizing unpredictable facts of human language.

Acknowledgemet

I thank Masakazu Kuno and Christopher Tancredi for insightful suggestions, discussions, and encouragements. Without them this paper would never be completed. Christopher Tancredi also helped me with grammatical judgements on the English sentences. I also thank Ayaka Sugawara, Akira Watanabe, and three PACLIC reviewers for

[5] In spite of the shared intuition, Martin takes a different path to account for the behavior of defeasible causative verbs. Comparison of the two accounts are beyond the scope of this study.

[6] In fact, verbs like *throw* and *goad* constantly resist an inanimate subject (**The heavy wind throw him a towel.* / **The situation goaded him to leave.*) Why, then, is that *teach* has an alternative manner-less LCS while *throw* does not? As Martin (2019) and Demirdache and Martin (2015) observe, there is a crosslinguistic tendency in which verbs are counted as a defeasible causative verb. Therefore, I believe there must be a linguistically plausible answer for this question. This is an interesting issue to address, but for now I have to leave it for a future study.

helpful suggestions and criticisms. All remaining errors and misinterpretations are my own.

References

Beavers, J. (2011). An Aspectual Analysis of Ditransitive Verbs of Caused Possession in English. *Journal of Semantics, 28*, 1-54.

Demirdache, H., & Martin, F. (2015). Agent control over non culminating events. In E. Barrajón López, J. Cifuentes Honrubia, & S. Rodríguez Rosique (Eds.), *Verb Classes and Aspect* (pp. 185-217). John Benjamins.

Dowty, D. (1979). *Word meaning and Montague Grammar.* Dordrecht: Reidel.

Hacquard, V. (2006). *Aspects of Modality.* Ph.D. Thesis, M.I.T.

Hacquard, V. (2010). On the event relativity of modal auxiliaries. *Natural Language Semantics, 18*, 79-114.

Hacquard, V. (2011). Modality. In C. Maienborn, K. von Hausinger, & P. Portner (Eds.), *Semantics: An International Handbook of Natural Language Meaning* (pp. 1484-1515). Mouten de Gruyter.

Harley, H. (2002). Posession and the double object construction. In *Linguistic Variation Yearbook 2* (pp. 31-70).

Harley, H., & Jung, H. K. (2015). In Support of the PHAVE analysis of the Double Object Construction. *Linguistic Inquiry, Volume 46*, 703-730.

Jackendoff, R. (1990). *Semantic Structures.* The MIT Press.

Karttunen, L. (1971). Implicative Verbs. *Language, 47*, 340-358.

Koenig, J.-P., & Davis, A. (2001). Sublexical Modality and the Structure of Lexical Semantic Representations. *Linguistics and Philosphy, 24*, 71-124.

Kratzer, A. (2013). *Creating a Family: Transfer of Posession Verbs.* Slides presenetd at the Workshop on Modality Across Categories. Barcelona (Universitat Pompeu Fabra), November 5.

Levin, B., & Rappaport Hovav, M. (2011). Lexical Conceptual Structure. In K. von Heusinger, C. Maienborn, & P. Portner (Eds.), *Semantics: An International Handbook of Natural Language Meaning I* (pp. 418-438). Mouton de Gruyter.

Martin, F. (2015). Explaining the link between agentivity and non-culminating causation. In S. D'Antonio, M. Moroney, & C. R. Little (Eds.), *Proceedings of SALT, 25* (pp. 246-266). CLC Publigations.

Martin, F. (2018). Time in probabilistic causation: direct vs. indirect uses of lexical causative verbs. *Proceedings of Sinn und Bedeutung 22*, 107-124.

Martin, F. (2019). *Aspectual differences between agentive and non-agetive uses of causative predicates.* Ms. . Retrieved 8 13, 2019

Martin, F., & Schäfer, F. (2017). Sublexical modality in defeasible causativ verbs. In A. Arregui, M. L. Rivero, & S. Andrés (Eds.), *Modality Across Syntactic Categories* (pp. 87-108). Oxford University Press.

Parsons, T. (1990). *Events in the Semantics of English: A Study in Subatomic Semantics.* The MIT Press.

Pesetsky, D. (1995). *Zero Syntax Experiencers and Cascades.* The MIT Press.

Pylkkänen, L. (2008). *Introducing Arguments.* The MIT Press.

Rappaport Hovav, M., & Levin, B. (1998). Building Verb Meanings. In M. Butt, & W. Geuder (Eds.), *The Projection of Arguments: Lexical Compositional Factors* (pp. 97-134). Stanford: CLSI Publications.

Rappaport Hovav, M., & Levin, B. (2010). Reflections on Manner/Result Complementarity. In E. Doron, M. Rappaort Hovav, & I. Sichel (Eds.), *Syntax, Lexical Semantics, and Event Structure* (pp. 21-38). Oxford University Press.

Reinhart, T. (2000). The Theta System: Syntactic realization of verbal concepts. *OTS Working Papers in Linguistics*.

Probabilistic Measures for Diffusion of Linguistic Innovation: As Seen in the Usage of Verbal "Nok" in Thai Twitter

Nozomi Yamada. Pittayawat Pittayaporn.
254 Phayathai Road, Pathumwan, Bangkok 10330 Thailand
Southeast Asian Linguistics Research Unit and
Department of Linguistics, Faculty of Arts, Chulalongkorn University
y.nozomi320@gmail.com
Pittayawat.P@chula.ac.th

Abstract

The existence of several SNS (social networking service) such as Twitter accelerates the diffusion process of language change. In this paper, we examine the diffusion of the innovative verbal usage of *nók* in Thai Twitter. We collected more than 25 millions tweets and adopted not only word frequency but also three probabilistic measures of analysis: conditional probability, PMI and cosine similarity of word embeddings. The result of these three probabilistic measures show the stability of the innovation regardless of decrease of word frequency. These facts support the idea that the innovation *nók* is lexically established in Thai language. Most importantly, it shows that the three probabilistic measures can be used to quantify diffusion of linguistic innovation regardless of its polysemy.

1 Introduction

Twitter is one of the most popular social networking services in Thailand. Though there is no official demographic profile, some online statistics websites like *we are social*[1] rank Twitter as the 3rd most popular SNS in Thailand behind Facebook and Instagram. Not only is it a large, free sources of relatively casual language used in daily communication (Crystal, 2006), but also potential space for examining early stages of language change. As networks in Twitter mainly consist of weak ties characterized by occasional contacts and lack of emotional bonding (Virk, 2011), linguistic innovations can spread quickly in Twitter, making it possible to observe complete propagation of language change in a short period of time.

An interesting and methodologically challenging case study is that of the verbal *nók* in Thai, an innovation gaining currency among Thai speakers. Originally a noun that means "bird", at present, the word is also used as a slang meaning "to fail to achieve one's expectation", especially used in the context of love or flirting. Although it is not clear when *nók* first came to be used as a verb, it was already popular to some extent among transgender women and gay men in 2014, and was commonly used among TV personalities by 2015.

This paper explores how Twitter data can be used to analyze the diffusion of an innovative lexical usage by taking the example of verbal *nók*. This innovation is chosen because it is a case of polysemy. Unlike cases in which a new variant propagates at the expense of an old one (Nevalainen and Raumolin-Brunberg, 2016), the verbal *nók* is not in competition with any other word. More specifically, the polysemy poses two challenges: how to detect and separate the innovative usage from the original usage for data processing, and how to quantify its progress through the linguistic system.

Therefore, this paper shows that changes in conditional probability, PMI, and cosine similarity of word embeddings are better measures for diffusion progress than word frequency. These measures also show that the verbal *nók* has been established as a new usage and and is broadening its meaning in Thai language.

[1] https://www.slideshare.net/DataReportal/digital-2019-thailand-january-2019-v01

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 271-278, Hakodate, Japan, September 13-15, 2019

2 Literature Review

2.1 Linguistic Innovation and Diffusion

Linguistic innovations are the ultimate origins of language changes (Milroy and Milroy, 1985), and such innovative usage of *nók* can be viewed as a case of lexical innovation. As Sornig (1981) explains, there are many kinds of lexical innovations: new word for new concept, new competitive word for existing word, new meaning for existing word, and so on. The case of our study verbal *nók* is an example of "new meaning for existing word".

Following Rogers (1962) who demonstrates that the diffusion of innovations is often represented as S-curve (logistic curve / sigmoid curve), Trudgill (1974) and Milroy and Milroy (1985) extended the theory to linguistic innovations, showing that they diffuse in the same manner as other social phenomena. More recently, Yang (2000) make probabilistic models for language change to explain how competing variations of word order in Old English and Old French were diffused and decayed. Blythe and William (2012) and Kauhanen (2017) put an emphasis on genetic replicating process in utterances, and make several mathematical models for explaining language change and S-curve. Ishii, et al. (2012) focus not only internal networks but also the relationship between external effects and utterance in order to give an account for short-term phenomenon as well as long-term propagation.

While these studies mainly focus the mechanism of diffusion, namely "how innovation is propagated", there are other studies that focus "whether innovation is diffused or not". As previous studies such as Nation and Waring (1997) and Piantadosi (2014) shows, word frequency is one of the most inteligible criteria for measuring generality of words. Metcalf (2004) and Barnhart (2007) thus introduce a scale for measuring acceptance of a new word by using word frequency. Phillips (2006) asserts that the diffusion of lexical innovation also becomes S-curve in the same way as other linguistic innovations by using word frequency. Hilpert and Gries (2008), using historical corpora, claim that diachronic change of word frequency directly indicates that the language change is in progress.

2.2 Language Change and Twitter

Social media like Twitter is an exciting domain for investigating the propagation of language change. Viewing language change as a result of diffusion of a speaker innovation (Milroy and Milroy, 1985), social networking services allow us to observe the rates at which linguistic innovation diffuses through speech communities and through linguistic systems. For example, Maybaum (2013) investigated several new words related to Twitter such as *tweeps, tweeple* by using big size of tweet data and showed that there is a tendency to be S-curve. However, this case is a type of "new word for new concept", which is not the case of *nók*.

Kershaw, Rowe and Stacey (2016) investigated innovation acceptance in twitter by using measuring scale of Metcalf (2004) and Barnhart (2007) and showed significance of word frequency. On the other hand, Yamanouchi and Komatsu (2014) focus not only word frequency, but also stochastic process of utterance and alpha-stable distribution of probability of each word. While word frequency may easily fluctuate under the influence of external events or randomness, the indices that determines distribution of probability are more stable. They proved the fact by sampling data from Twitter.

3 Data and pre-processing

The data we used is tweets written in Thai language from January 2012 to December 2018 (Table 1). First, we collected about 1000 - 2500 tweets per day (data set A) containing the word *nók*. This data is used for two analyses: conditional probability and word embeddings. Next, we collected about 10000 - 25000 random tweets per day (data set B) for three analyses: word frequency, PMI and word embeddings. In order to prevent from being biased, we collected tweets every 10 minutes.

The most crucial step in data pre-processing is to distinguish cases of the innovative verbal *nók* from cases of the original nominal usage. The most reliable heuristic is its co-occurrence with negator *mâi* or auxiliary verb. As the verbal *nók* can only have the innovative meaning 'to fail achieve one's expectations', the two occurrence patterns also distinguish between the old and new meanings. Table 2 gives examples of the most common syntactic structures

Year	A: Tweets with *nók*	B: Random Tweets
2012	118,799	0
2013	476,365	2,529,665
2014	425,421	3,732,020
2015	395,334	3,153,596
2016	778,243	3,434,185
2017	1,070,668	5,152,559
2018	891,636	3,556,596
total	4,156,466	21,558,621

Table 1: The number of collected tweets

in Thai.

Structure	Example	Gloss
S V	*phŏm pai*	I go
S Adj	*phŏm hĭw*	I hungry
S NEG V	*phŏm mâi pai*	I not go
S NEG Adj	*phŏm mâi hĭw*	I not hungry
S AUX V	*phŏm cà pai*	I will go
S Cop N	*phŏm pen nók*	I be bird
S NEG Cop N	*phŏm mâi châi nók*	I not be bird

Table 2: Most common syntactic structures in Thai

The copular verbs *pen* and *châi* are needed when the sentence is nominal sentence as shown in the last two examples. In other words, the collocations *mâi nók* or *AUX nók* will never occur as long as *nók* is occurs as nominal meaning "bird". Therefore, to make a list of co-occurrences proved essential.

In order to find these co-occurrences, we tokenized all tweets with the python toolkit `PyThaiNLP 2.0.3`, and the Maximum-Matching (MM) algorithm. MM algorithm requires a vocabulary set (dictionary) for tokenization, and we can control it. Since we wanted to locate words preceding/following *nók*, we removed all compound words containing *nók* such as *nókphirâap* ("pigeon"), from the vocabulary set beforehand.[2]

In addition, there are some tweets that include repetition of the same characters or words in order to exaggerate, such as "aaaaraaaaiiii", "nóknóknóknóknók". Since these repetitions make it more difficult for a program to tokenize, we detected them by using regular expressions, then condensed them before tokenization.

[2]then, *nókphirâap* is tokenized as *nók* and *phirâap*

4 Measures of Diffusion

4.1 Word Frequency

As mentioned above, word frequency is one of the most popular methods for measuring diffusion of innovation. We calculated word frequency (per 10000 words) by counting tokens of *nók* as well as all tokens for each month, then traced the diachronic change. However, the word *nók* is not a new word but a polyseme, so word frequency alone does not indicate how often the innovative verbal *nók* is used. We thus needed other methods to separate the two usages and normalize them.

4.2 Conditional Probability of Bigrams

Bigrams are an important methods in computational linguistics especially in building language models. For our purposes, bigrams are employed in detecting the syntactic structure of the sentence is and how often that structure occurred. We mentioned in section 3 that verbal sentences and nominal sentences take different structures. As such, identifying collocations can help to distinguish the two meanings. However, calculation of collocation must be normalized so that we could compare diachronically regardless of the size of the data. We thus defined conditional probability for preceding word $P_{pre}(w_i|nok)$ and conditional probability for following word $P_{fol}(w_i|nok)$ as follows:

$$P_{pre}(w_i|nok) = \frac{C(w_i, nok)}{\sum_w C(w, nok)} \quad (1)$$

$$P_{fol}(w_i|nok) = \frac{C(nok, w_i)}{\sum_w C(nok, w)} \quad (2)$$

where $C(w_i, nok)$ is the total number of co-occurrences of a word w_i and *nók* in all tokens[3], $\sum_w C(w, nok)$ is the total number of co-occurrences containing *nók* as the second word of the bigram. For example, the conditional probabilities for words A, B, C in the sentence "*nók A B nók B C nók A*" are given by following calculations:

$$\begin{array}{ll} P_{pre}(A|nok) = 0 & P_{fol}(A|nok) = 2/3 \\ P_{pre}(B|nok) = 1/2 & P_{fol}(B|nok) = 1/3 \\ P_{pre}(C|nok) = 1/2 & P_{fol}(C|nok) = 0 \end{array}$$

[3]not including white space and punctuation

4.3 Tweet-Level PMI

Another probabilistic measure is Pointwise Mutual Information (PMI) at tweet level. PMI is a measure of the independency of two words (Jurafsky and Martin, 2014), the degree of how often (or not) the two words co-occur. PMI is similar to conditional probability of bigrams, though the method of normalization differs slightly. Moreover, with conditional probability, we did not consider which tweet the bigram comes from. In other words, we dealt with all tweets as one text. Here, we define tweet-level PMI of bigrams as:

$$\mathrm{PMI}(w_i, w_j) = \log_2 \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \qquad (3)$$

where $p(w_i, w_j)$ is the probability that one tweet contains a co-occurrence of the bigram (w_i, w_j), in short, the proportion of the tweets that contain the target bigram. $p(w)$ is the probability of occurrence of word w in any given tweet. The normalization factor (given in the denominator) is the possibility of occurence for each word. In this case, either of the two words is *nók*. Since every tweet in data set A contains the word *nók* and therefore, $p(nok) = 1$, we used only data set B (random tweets) for this PMI calculation.

4.4 Cosine Similarity of Word Embeddings

The third probabilistic measure is cosine similarity of word embeddings. Though a word embedding itself does not provide a direct probability, the methods of obtaining word embeddings, such as SVD or word2vec, are based on distribution of words. We thus refer to it as "probabilistic" in a broad sense. Several studies such as Hamilton, Leskovec and Jurafsky (2016), Bamler and Mandt (2017), Baitong, Ying and Feicheng (2018) reveal that comparison of word embeddings derived from various periods of historical corpora can, in fact, reveal language change. Moreover, their studies also show that a word embedding of a polyseme is located between the embeddings of its two meanings. In other words, we can observe the propagation of language change by measuring whether the cosine similarity between the word *nók* and another word that means "to fail to achieve one's expectation" is rising or not.

We employed the word2vec toolkit `gensim 3.7.1` in computing 300-dimension word embeddings for each month from 2014 to 2018. In order to obtain not only the word embedding of *nók* but also the word embeddings of various words for comparison, we first combined two data sets: data set A (tweets that contain *nók*) and data set B (random tweets). For all word embeddings, we used a CBOW algorithm with symmetric windows of size 5, and iterated for 3 epochs. Though skip-gram algorithms are more popular at present, we used a CBOW algorithm as it works better with frequently occurring words (Naili, Chaibi, and Ghezala, 2017).[4]

5 Results

5.1 Word Frequency

Before addressing the frequency of *nók*, we first performed a preliminary test of a selected set of basic words -frequencies of basic words should be stable diachronically- to check that our data set was sufficiently sized and void of bias. We chose three frequent words: *mâi* "not", *pen* "be" (copula) and *tham* "do". Figure 1 gives a plot of frequencies per 10,000 tokens of each word in data set B (random tweets) from 5/2013 to 10/2018

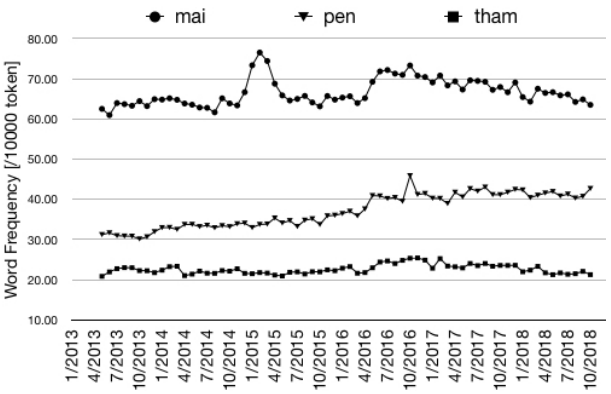


Figure 1: Word frequency of 3 words in data set B

The frequency of *mâi* fluctuates within the period examined, while *pen* is increases gradually, and *tham* is almost stable. Though these three words display slightly different patterns, none of them show an abrupt change. We can thusly conclude that our set is not biased.

[4] *bird* is 20th on the Swadesh list

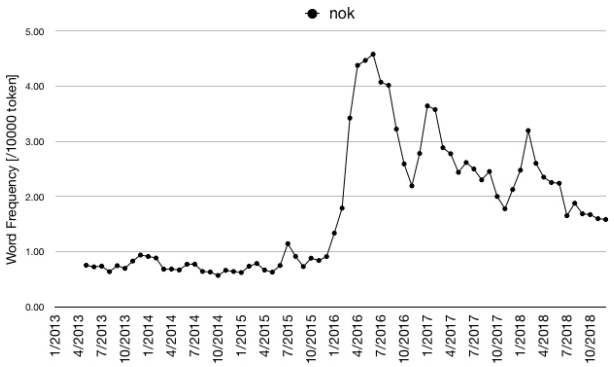


Figure 2: Word frequency of *nók* in data set B

Next, let us compare the frequency of *nók* shown in Figure 2 which, contrastively to the frequencies of the basic words, increases abruptly at the beginning of 2016 and reaches its peak in the middle of the same year at more than 4 times its original value. After that, it resembles exponential decay, falling to only 1.6 times the original value. From this result, the innovation seems to be disappearing rather than diffusing. Though this abrupt change may be evidence of linguistic innovation, we cannot make this determination based only on this data, as this word frequency is the sum of both original *nók* and verbal *nók*. Since our data is too large to check one by one, we cannot obtain how much the word frequency of verbal *nók* is.

5.2 Probabilistic Measures

The three probabilistic measures to be discussed here demonstrate a contrastive aspect to that of word frequency. First is the calculation of P_{pre} and P_{fol}, for which we selected three words capable of distinctively indicating syntactic structure. Tables 3 and 4 lists these words.

Word	POS	Meaning	Note
mâi	Adv	"not"	negation
cà	AUX	"will"	V follows
jàa	AUX	"Don't"	imperative

Table 3: 3 words selected for P_{pre} calculation

The three words for P_{fol} do not directly indicate the status of *nók* as a verb as much as the words for P_{pre}; however, they tend to follow verbs in Thai. Figure 3 and Figure 4 are transitive graphs of conditional probability P_{pre} and P_{fol}.

Word	POS	Meaning
laéw	Adv	"already"
ìik	Adv	"again", "more"
talòt	Adv	"always"

Table 4: 3 words selected for P_{fol} calculation

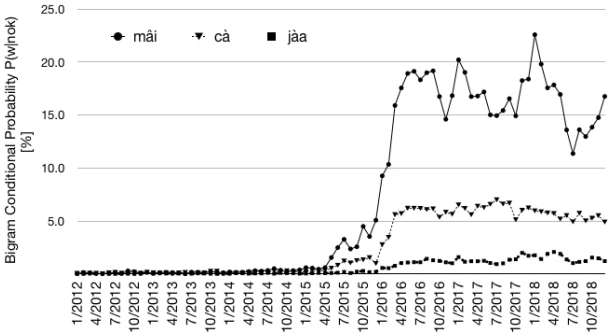


Figure 3: Conditional probability for the preceding word

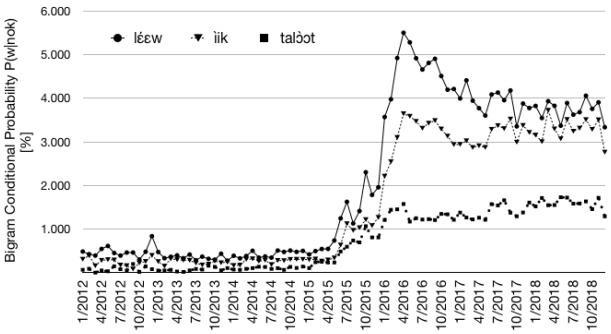


Figure 4: Conditional probability for the following word

Figure 3, of course, demonstrates an increase in each co-occurrence after 2015. Conditional probability forms an S-curve, hardly decaying after reaching its peak.

Figure 4 displays similar patterns. Though only *laéw* decays after reaching its peak, its value later becomes stable in the same way as the other two words, *ìik* and *talòt*.

A similar shape occurs, as well, with calculation of tweet-level PMI for the same two words.

Figure 5 shows PMI(*mâi*, *nók*) and PMI(*cà*, *nók*). Notably, some points are not plotted as the log of zero will equal negative infinity. Both curves

Word	POS	Meaning
mâi	Adv	"not"
cà	AUX	"will"

Table 5: 2 words selected for PMI calculation

abruptly increase in 2016 as was the case with conditional probability. Though they seem to decrease gradually, neither mirrors the decreasing pattern of word frequency and can be considered more stable.

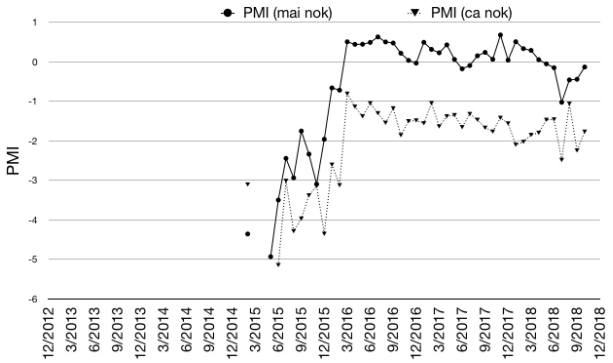


Figure 5: PMI(*mâi*, *nók*) and PMI(*cà*, *nók*)

The third measure is cosine similarity of word embeddings. We selected the two words below and measured cosine similarity between each of these and *nók*. As shown in Table 2, in Thai, verbs and adjectives will appear in the same structures. Thus, we can regard these two words as synonyms of verbal *nók*.

Word	POS	Meaning
plâat	Verb	"to miss", "to fail"
sĭacai	Adj	"(to feel) sorry"

Table 6: 2 words selected for cosine similarity calculation

Figure 6 and 7 show diachronic change of the cosine similarities for each pair. Similarity before 2015 is near zero, then after 2015, it forms an S-curve and the value hardly decays, even after reaching its peak. This means that the new meaning of the verbal *nók* still remains.

6 Discussion

There are obvious differences between measures of word frequency and the three probabilistic measures. We can surely conclude that people have been using *nók* less frequently following its boom in 2016; however, this does not simply mean that the lexical innovation is disappearing. According to Figure 2, word frequency in later 2018 is less than half of that of early 2016. only the word frequency of the new meaning of *nók* decreases while frequency of the old usage remains constant, conditional probability and PMI must decrease as well. On the contrary, the result indicates constancy of both, meaning the proportion of verbal *nók* to total use of *nók* on Twitter is unchanged, regardless of decreases in word frequency itself. Comparing conditional probability and PMI, conditional probability is more convenient for use as its measure requires only tweets containing *nók*, while, for PMI, random tweets must be gathered for calculating the probability $p(nok)$.

This constancy is true of cosine similarity as well, and since cosine similarity can indicate the meaning of the word more directly than conditional probability or PMI, the result is more convincing. The co-

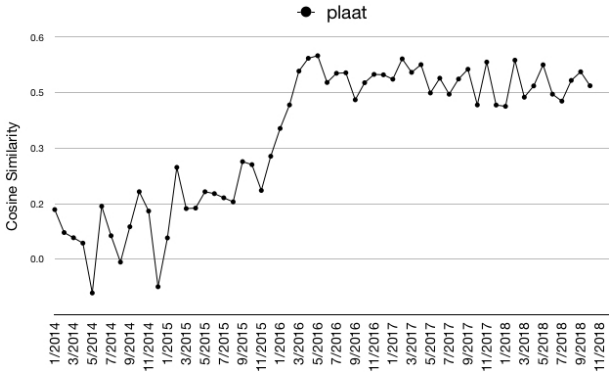


Figure 6: Cosine similarity between *nók* and *plâat*

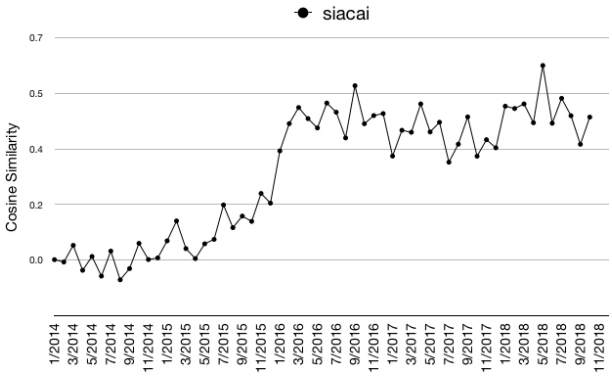


Figure 7: Cosine similarity between *nók* and *sĭacai*

sine similarity is rising throughout 2015, and finally, it becomes stable with no decrease, indicating continued use *nók* in the same context as the two compared words based on the fact that, if the usage is not already established within the linguistic system, cosine similarity would be expected to decay back to zero. According to these results, we can conclude that the lexical innovation *nók* has been established in the linguistic system of Thai.

However, our study has several limitations. First, though our results have revealed characteristics of language change on Twitter, we cannot discern the total acceptance rate. Even if a lexical innovation is already established on Twitter with a constant probability, this does not entail that every Twitter user accepts the innovation. Since this lexical innovation is of type "new meanings for existing words" and not of "new competitive words for existing words", it is impossible to measure 0-100 % acceptance rates from the beginning. We must, therefore, take an "ensemble average" by sampling a nascent language change within the linguistic system and its diffusion pattern.

Second, it is fortunate that *nók* is a polyseme of verb and noun as syntactic structure differs between nominal sentences and verbal sentences, and thusly, conditional probability and PMI containing frequently occurring grammatical words (i.e. negators or auxiliary verbs) can be used. However, if there were a polyseme that is morphosyntactically identical, differentiation may prove to be much more difficult. In that case, we would be forced to use less frequent collocation than that of grammatical words.

As well, we have not analyzed what kinds of mechanisms are present. In other words, we did not show how innovation was propagated, only whether innovation has been diffused or not. There must be at least two factors for the propagation mechanisms: internal networks and external effects. As we reviewed in section 2, there have been many previous studies accounting for diffusion mechanisms through use of various models. In the future, we plan to build from the current research and explore the mechanism of *nók* with such models.

Incidentally, we found another phenomenon occurring in the diffusion of *nók*. The innovative meaning of *nók* in the beginning is just "to fail to flirt" and applies to both men and women. After the innovation had been diffused to the masses, the meaning broadened. In other words, it came to be used as a more generalized verb. Figure 8 gives the PMI of *nók* and *bàt* "ticket". This figure shows that the first appearance and peak for this pair are delayed in comparison to other pairs, and that use is still increasing. Additionally, we found many nouns following *nók*, such as "thing", "live", "giveaway", in the data, suggesting greater favorability toward a wider array of environments and, therefore, a more general meaning. This broadening of meaning also supports the idea that innovative *nók* has been established.

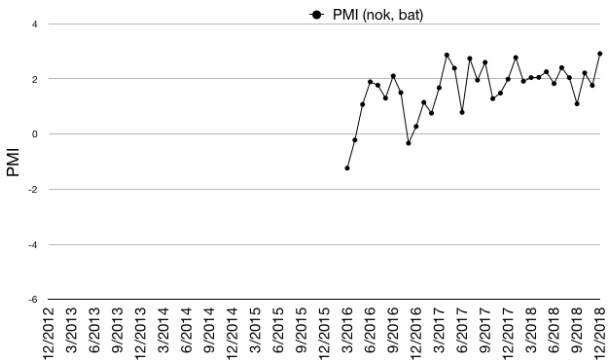


Figure 8: PMI(*nók*, *bàt*)

7 Conclusion

The results indicate that the word frequency of *nók* is now decreasing, while three probabilistic measures are stable over time. This fact supports the idea that the lexical innovation has been established in linguistic system.

The most significant point is that the three measures we adopted can deal with polysemy, unlike word frequency. These measures can be used to quantify diffusion of linguistic innovation regardless of its polymsemy. Though this study examined only one case in Thai, the methods employed here are universally applicable with potential to be extended to other languages as well.

Acknowledgement

We are deeply grateful to Dr. Attapol Thamrongrattanarit for offering good hints for analyses and our colleague Ashley Laughlin for correcting proofs.

References

Baitong Chen, Ying Ding and Feicheng Ma. 2018. Semantic word shifts in a scientific domain. *Scientometrics* 117:211–226

Bamler, R. and Mandt, S. 2017 August. Dynamic word embeddings. In *Proceedings of the 34th International Conference on Machine Learning-Volume* 70 (pp. 380-389). JMLR. org.

Barnhart, D.K. 2007. A calculus for new words. *Dictionaries: Journal of the Dictionary Society of North America*, 28(1), pp.132-138.

Blythe, Richard A., and William Croft. 2012. S-curves and the mechanisms of propagation in language change. *Language*, 269-304.

Crystal, D. 2006. *Language and the Internet*. Cambridge, Cambridge University Press.

Hamilton, W.L., Leskovec, J. and Jurafsky, D. 2016. Diachronic word embeddings reveal statistical laws of semantic change. *arXiv preprint arXiv*:1605.09096.

Hilpert, M. and Gries, S.T. 2008. Assessing frequency changes in multistage diachronic corpora: Applications for historical corpus linguistics and the study of language acquisition. *Literary and Linguistic Computing*, 24(4), pp.385-401.

Ishii, A., Arakaki, H., Matsuda, N., Umemura, S., Urushidani, T., Yamagata, N. and Yoshida, N. 2012. The 'hit'phenomenon: a mathematical model of human dynamics interactions as a stochastic process. *New journal of physics*, 14(6), p.063018.

Jurafsky, D. and Martin, J.H. 2014. *Speech and language processing* (Vol. 3). London: Pearson.

Kauhanen, H. 2017. Neutral change *Journal of Linguistics*, 53(2), 327-358.

Kershaw, D., Rowe, M. and Stacey, P. 2016, February. Towards modelling language innovation acceptance in online social networks. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining* (pp. 553-562). ACM.

Labov, W. 1966. The linguistic variable as structural unit. Wash. *Linguist* Rev. 3, 4–2*2.

Maybaum, R. 2013, December. Language change as a social process: Diffusion patterns of lexical innovations in Twitter. In *Annual Meeting of the Berkeley Linguistics Society* (Vol. 39, No. 1, pp. 152-166).

Metcalf, A. 2004. *Predicting New Words. The Secrets of Their Success* Houghton Mifflin Harcourt.

Milroy, J. and Milroy, L. 1985. Linguistic change, social network and speaker innovation. *Journal of linguistics*, 21(2), pp.339-384.

Naili, M., Chaibi, A.H. and Ghezala, H.H.B. 2017. Comparative study of word embedding methods in topic segmentation. *Procedia computer science*, 112, pp.340-349.

Nation, P. and Waring, R. 1997. Vocabulary size, text coverage and word lists. *Vocabulary: Description, acquisition and pedagogy*, 14, pp.6-19.

Nevalainen, T. and Raumolin-Brunberg, H. 2016. *Historical sociolinguistics: language change in Tudor and Stuart England*. Routledge.

Phillips, B. 2006. *Word frequency and lexical diffusion*. Springer.

Piantadosi, S.T. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21(5), pp.1112-1130.

Rogers, Everett M. 1962. *Diffusion of innovations*. New York: Free Press of Glencoe.

Sornig, K. 1981. *Lexical innovation: A study of slang, colloquialisms and casual speech*. John Benjamins Publishing.

Tamburrini, N., Cinnirella, M., Jansen, V.A. and Bryden, J. 2015. Twitter users change word usage according to conversation-partner social identity. *Social Networks* 40, pp.84-89.

Thomsen, O.N. ed. 2006. *Competing models of linguistic change: evolution and beyond* (Vol. 279). John Benjamins Publishing.

Trudgill, P. 1974. Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography. *Language in society*, 3(2), pp.215-246.

Virk, Amardeep. 2011. Twitter: The strength of weak ties. *University of Auckland Business Review*, 13(1), p.19.

Yamanouchi, Y, and Komatsu, T. 2014. Power Law in SNS Language Probability Space and Stable Distribution. Japan: In *Journal of The Infosocionomics Society* (vol. 11, No. 1)

Yang, C.D. 2000. Internal and external forces in language change. *Language variation and change*, 12(3), pp.231-250.

Thai Legal Term Correction using Random Forests with Outside-the-sentence Features

Takahiro Yamakoshi[†], **Vee Satayamas**[‡], **Hutchatai Chanlekha**[‡]
Yasuhiro Ogawa[†], **Takahiro Komamizu**[†], **Asanee Kawtrakul**[‡], **Katsuhiko Toyama**[†]
[†] Nagoya University, Japan
{yamakoshi,ogawa,komamizu,toyama}@kl.itc.nagoya-u.ac.jp
[‡] Kasetsart University, Thailand
{vee.sa,hutchatai.c,ak}@ku.th

Abstract

We propose a method for finding and correcting misused Thai legal terms in Thai statutory sentences. Our method predicts legal terms using Random Forest classifiers, each of which is optimized for each set of similar legal terms. Each classifier utilizes outside-the-sentence features, namely, promulgation year, title keywords, and section keywords of statutes, in addition to words adjacent to the targeted legal term. Our experiment shows that our method outperformed not only a Random Forest method without the outside-the-sentence features, but also BERT (Bidirectional Encoder Representations from Transformers), a powerful language representation model, in overall accuracy.

1 Introduction

Legislation drafting requires careful scrutiny. An important consideration is the appropriate use of legal terms. In Thai legislation, allowable usage of similar legal terms is described in the legislation manual from the Office of the Council of State, the bill examining authority. For example, there are two similar Thai legal terms *yang-nueng-yang-dai* (อย่างหนึ่งอย่างใด; lit. thing-one-thing-any) and *yang-dai-yang-nueng* (อย่างใดอย่างหนึ่ง; lit. thing-any-thing-one) separately used in Thai statutory sentences. Both terms are used to choose entities from a given set, like "some of the following items." However, according to the legislation manual, *yang-nueng-yang-dai* is used only when one can choose one or more entities, while *yang-dai-yang-nueng* is used only when only one entity can be chosen (Office of the Council of State, 2008). Drafters must not misuse any legal term in a bill; otherwise the bill can have unintended provisions, and thus unintentionally and incorrectly govern the people. Therefore, drafters need to scan the hundreds of pages of the bill thoroughly to locate misused legal terms and correct them; however, scanning is currently done by humans, which requires an enormous amount of time and is subject to human error.

We therefore propose a legal term correction method for Thai statutory sentences that assists drafters in finding misused legal terms in a draft and offers corrections. Inspired by Yamakoshi et al. (2018)'s idea, we handle legal term correction as a special case of the multiple-choice sentence completion test by regarding a set of similar legal terms as a set of choices. Also, we adopt Random Forest classifiers (Breiman, 2001) to score the likelihood of each candidate for the legal term. Here, we introduce additional features from outside of the statutory sentence, namely, year, title keyword, and section keyword. We expect that the year feature copes with changes in legal term usage over time, the title keyword feature captures the difference in legal term usage by statute type, and the section keyword feature adequately predicts a legal term in an item with few adjacent words.

The contributions of our paper are as follows: (1) we apply a legal term correction method that successfully completes the Japanese legal term correction task and confirm the effectiveness of this method for statutory sentences in other legislation

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 279-287, Hakodate, Japan, September 13-15, 2019

systems, (2) we design three additional features, one a temporal feature and the others topical features, and (3) we examine the extended legal term correction method with the original method and a modern method that is based on BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2018), and we demonstrate that the extended method outperforms the others.

In Section 2, we introduce several sets of Thai legal terms regulated in the legislation manual. In Section 3, we survey related work. In Section 4, we show our method. In Sections 5 and 6, we present our evaluation experiment and discuss the results, respectively. Finally, we summarize and conclude our paper in Section 7.

2 Thai Legal Terms

In this section, we explain several sets of Thai legal terms whose usage is defined in the Thai legislation manual (Office of the Council of State, 2008).

2.1 *Yang-nueng-yang-dai* and *yang-dai-yang-nueng*

Yang-nueng-yang-dai (อย่างหนึ่งอย่างใด) is literally "thing-one-thing-any," while *yang-dai-yang-nueng* (อย่างใดอย่างหนึ่ง) is "thing-any-thing-one," so they look very similar. In Thai statutory sentences, these terms are used in choosing entities from a particular set. *Yang-nueng-yang-dai* is used when one or more entities can be selected simultaneously. On the other hand, *yang-dai-yang-nueng* is used when only one entity can be selected.

Yang can be substituted for other words such as *khon* (คน; person), so we can use *khon-nueng-khon-dai* (คนหนึ่งคนใด; one or more people) or *khon-dai-khon-nueng* (คนใดคนหนึ่ง; only one person).

2.2 *Amnat-nathi*, *amnat-lae-nathi*, and *nathi-lae-amnat*

Amnat-nathi (อำนาจหน้าที่), *amnat-lae-nathi* (อำนาจและหน้าที่), and *nathi-lae-amnat* (หน้าที่และอำนาจ) consist of *amnat* (อำนาจ; power), *nathi* (หน้าที่; duty), and *lae* (และ; and). *Amnat-nathi* is now considered a compound word, while *amnat-lae-nathi* and *nathi-lae-amnat* are noun phrases. According to a Thai law dictionary, *amnat-nathi* means cognizance or competence (Tipchod and KhotchaSeni, 2013). Although still a matter of discussion, *amnat-nathi*, *amnat-lae-nathi*, and *nathi-lae-amnat* have the following usages: *amnat-nathi* means the power to perform duties; *amnat-lae-nathi* is just a combination of two words, "power" and "duty," and is used when both powers and duties are defined in the statute; and *nathi-lae-amnat* is the concept that one must have duties before having power. It is important to note that the appearance of *amnat-lae-nathi* is recent and that the constitution of Thailand has used only *amnat-nathi*.

2.3 *Panakngan-chaonathi* and *chaonathi*

Both *Panakngan-chaonathi* (พนักงานเจ้าหน้าที่; competent authority (Tipchod and KhotchaSeni, 2013)) and *chaonathi* เจ้าหน้าที่; officer) mean a person who has the power to practice a legal action. However, these terms are used for different kinds of people. The former is used for a person appointed by a minister, while the latter is used more generally.

2.4 *Kharachakan-kanmueang* and *phu-damrong-tamnaeng-thang-kanmueang*

Both *kharachakan-kanmueang* (ข้าราชการการเมือง, lit. official-politics) and *Phu-damrong-tamnaeng-thang-kanmueang* (ผู้ดำรงตำแหน่งทางการเมือง, lit. person-preserve-position-in-politics) mean a certain kind of public servant, but each has a different scope of meaning. The former is predominately used for a minister or their aide. The latter can indicate not only a person of *kharachakan-kanmueang*, but also a national assembly member, the mayor of Bangkok, a city council member, and so on.

3 Related Work

In this section, we survey related work on the legal term correction task. First, we describe the definition of the legal term correction task given by Yamakoshi et al. (2018), and then explain technologies that can be used to solve this task.

3.1 Legal Term Correction

Yamakoshi et al. defined the legal term correction task as follows (Yamakoshi et al., 2018):

Algorithm 1 Algorithm for legal term correction

Input: W, T

Output: Suggests

```
Suggests ← ∅
for all (i, j) such that w_i w_{i+1} ··· w_j = t ∈ T do
    W^ℓ ← w_1 w_2 ··· w_{i−1}
    W^r ← w_{j+1} w_{j+2} ··· w_{|W|}
    t_best ← arg max_{t'∈T} score(W^ℓ, t', W^r)
    if t ≠ t_best then
        Suggests ← Suggests ∪ {suggestion that t in position (i, j) should be replaced into t_best}
    end if
end for
```

- A statutory sentence $W = w_1\ w_2 \cdots w_{|W|}$ and a set of legal terms $T \subseteq V^+$ are given, where V^+ is the Kleene plus of vocabulary V; that is, legal term $t \in T$ can be either a word or multiple words;
- The adequacy of each legal term t found in W is judged;
- If another legal term $t_{\text{best}} \in T$ $(t_{\text{best}} \neq t)$ seems more adequate in the context, t_{best} is suggested as a replacement for t.

They also defined a general algorithm: Algorithm 1, where $\text{score}(W^\ell, t, W^r)$ is a scoring function that calculates the likelihood of term t when two word sequences W^ℓ and W^r are adjacent to the left and right of t, respectively.

This problem can be regarded as a special case of the sentence completion test by introducing the following ideas:

- W^ℓ ___ W^r is a sentence with a blank, where ___ is the blank, and W^ℓ and W^r are as defined in Algorithm 1.
- T is the choices, one of which adequately fills the blank in the sentence.

However, Yamakoshi et al. pointed out that this problem differs from the general multiple-choice sentence completion test in two ways. First, a set of choices (i.e., a legal term set) relates to many sentences with blanks. In contrast, we cannot assume that such a large number of sentences relate to a set of choices in the general multiple-choice sentence completion test, since we usually consider that each sentence with a blank has a different set of choices.

Second, we can consider only meaningful legal term sets mentioned by the legislation manuals. In contrast, we may consider any combination of choices in the general multiple-choice sentence completion test, since they are unrestricted.

3.2 Technologies for Solving the Legal Term Correction Task

In this section, we introduce some technologies for the scoring function of the legal term correction task. We use Random Forest (Breiman, 2001) as the scoring function. We describe Random Forest in Section 3.2.1. In the context of the sentence completion test, BERT (Devlin et al., 2018), a powerful language representation model, can be used as the scoring function. We briefly explain BERT in Section 3.2.2. Finally, in Section 3.2.3, we describe language models whose performance is traditionally evaluated by a sentence completion test.

3.2.1 Random Forest

Random Forest (Breiman, 2001) is a machine-learning algorithm for classification. Figure 1 explains the training and prediction processes of a Random Forest classifier.

It learns the training data by building a set of decision trees. A decision tree is conceptually a suite of if-then rules like the ones in the middle of Figure 1. After learning, the Random Forest classifier predicts the class of the given data by taking a vote on each decision tree. Here, each decision tree is constructed by randomly selected data records and features. Therefore, even if a single decision tree makes an unsophisticated decision, the ensemble of decision trees is better at predicting unseen data.

Yamakoshi et al. (2018) utilized Random Forest classifiers specialized for each legal term set as the scoring function in Algorithm 1. The following equation denotes the scoring function:

$$\begin{aligned} &\text{score}(W^\ell, t, W^r) \qquad (1)\\ &= \sum_{d \in D} P_d(t | w^\ell_{|W^\ell|-N+1}, \ldots, w^\ell_{|W^\ell|}, w^r_1, \ldots, w^r_N), \end{aligned}$$

where D is a set of decision trees, d is a decision tree, and $P_d(t|w_1, w_2, \ldots, w_N)$ is the probability (ac-

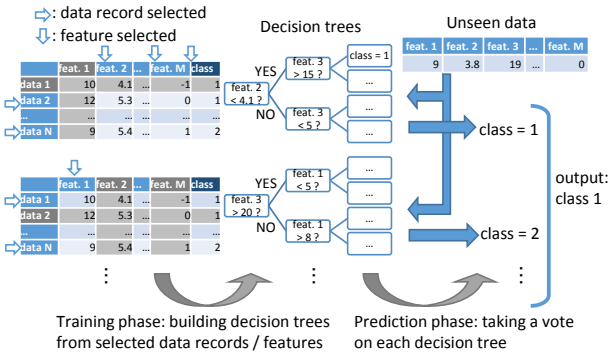


Figure 1: Processes of Random Forest

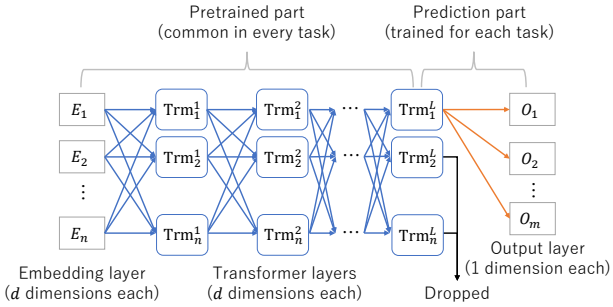


Figure 2: BERT model

tually 0 or 1) that d chooses t based on features $w_1, w_2, \ldots, w_N$. w_i^ℓ and w_i^r are the i-th word of W^ℓ and W^r, respectively. N is the window size (the number of left or right adjacent words focused on). If $|W^\ell| < N$, W^ℓ will be padded with out-of-sentence tokens (same in $|W^r|$).

After training, a Random Forest classifier outputs feature importance that indicates how much each feature contributes to a good prediction. Feature importance is calculated using out-of-bag examples (not sampled examples). The importance of the n-th feature is calculated by the following procedure:

1. Build the m-th decision tree T_m using randomly sampled examples;
2. Acquire the set of out-of-bag examples on decision tree $E_{m,0}$;
3. Make a set of examples $E_{m,n}$, where the n-th feature of each example is randomly shuffled;
4. Predict classes $C_{m,0}$ and $C_{m,n}$ of each example in $E_{m,0}$ and $E_{m,n}$, respectively, using T_m;
5. Calculate the increase of misprediction rate $C_{m,0}$ and $C_{m,n}$;
6. Calculate the total increase of misprediction rate x_n by applying 1. to 5. for every decision tree.
7. If x_n is high, the n-th feature is important because shuffling the feature brought about an inaccurate prediction.

3.2.2 BERT

Devlin et al. (2018) introduced a new language representation model called BERT (Bidirectional Encoder Representations from Transformers). This model is designed for a wide range of NLP tasks such as question answering and language inference. Figure 2 shows the construction of a BERT model. The BERT model in the figure inputs n words and outputs a probability distribution of m classes. In the legal term correction task, each output value denotes the probability of a certain legal term and will be the value of the scoring function in Algorithm 1.

A BERT model is a neural network model that consists of two parts: a pretrained part and a prediction part. The pretrained part consists of an embedding layer and Transformer layers, where each layer's unit is d-dimensional. Transformer (Vaswani et al., 2017) is a neural network model made of multi-head attention units and feedforward connections. The prediction part consists of the final Transformer layer and an output layer connected by feedforward connections. In a sentence-level classification task, only one Transformer unit connects with the output layer and other units are dropped.

When making a classification model for a particular task, we can inherit parameters of the pretrained part from a pretrained common model trained with a large-scale diversified corpus and fine-tune the whole model with a task-specific dataset. Using the pretrained common model, we can get quite high performance for various kinds of tasks with a small amount of training.

3.2.3 Language Model

A language model assigns a likelihood to each word sequence $W = w_1\ w_2 \ldots w_{|W|}$. In the legal term correction task, the model works as a scoring function in Algorithm 1 that outputs the likelihood of a sentence whose blank is filled with a legal term.

Here, each word w_i that constitutes W is chosen from a vocabulary that a language model defines. Therefore, a language model can solve any question of the sentence completion test if each word in a sentence with a blank and a set of choices is in the vocabulary.

To evaluate language models, Zweig and Burges (2011) presented a dataset of the multiple-choice sentence completion test called the MSR Sentence Completion Challenge Data.

A variety of language models are evaluated by this dataset. First, Zweig and Burges (2011) evaluated n-gram models with their dataset. The most powerful language models evaluated by this dataset have neural network architectures. For instance, Mikolov et al. (2013) proposed two neural language models: the Continuous Bag-of-Words Model (CBOW) and Continuous Skip-gram Model (Skipgram). Mnih and Kavukcuoglu (2013) proposed the vector Log-bilinear model (vLBL) and ivLBL. Mori et al. (2015) proposed vLBL(c) and vLBL+vLBL(c), which are improved models of vLBL that are sensitive to the relative positions of words adjacent to the target word.

4 Proposed Method

In this section, we show our proposed method for the legal term correction task. Our method is based on Yamakoshi et al. (2018)'s model that uses Random Forest as a scoring function. Unlike their method, our method introduces three additional features from outside of the sentence to augment prediction performance. We describe these features in Section 4.1, followed by our prediction model in Section 4.2.

4.1 Out-of-sentence Features

We introduce three additional features, namely, year, title keyword, and section keyword to our method. We describe these features and intentions below.

- **Year Feature**
  The year feature denotes the year when the statute was promulgated. We use this feature as a one-dimensional integer variable and introduce it to deal with changes in term usage over time. For example, *amnat-lae-nathi* has appeared recently; therefore, a prediction model with this feature should know that this legal term does not appear in older statutes.

1	มาตรา ๒๗ ผู้มีลักษณะอย่างใดอย่างหนึ่งดังต่อไปนี้ ต้องห้ามมิให้เป็นประธานกรรมการหรือกรรมการ คือ
2	(๑) มีส่วนได้เสียในสัญญากับการรถไฟแห่งประเทศไทย หรือในกิจการที่กระทำให้แก่การรถไฟแห่งประเทศไทย ทั้งนี้ ไม่ว่าโดยตรงหรือโดยทางอ้อม เว้นแต่จะเป็นเพียงผู้ถือหุ้นของบริษัทที่กระทำการอันมีส่วนได้เสียเช่นว่านั้น
3	(๒) เป็นพนักงานของการรถไฟแห่งประเทศไทย
4	(๓) เป็นข้าราชการการเมือง
5	(๔) ขาดคุณสมบัติหรือมีลักษณะต้องห้ามตามกฎหมายว่าด้วยคุณสมบัติมาตรฐาน สำหรับกรรมการ และ พนักงานรัฐวิสาหกิจ

Figure 3: A legal term (underlined) with few adjacent words

- **Title Keyword Feature**
  The title keyword feature denotes the keywords of the statute's title. We use this feature as a n-dimensional boolean variable, where n is the number of keywords defined, as each of its elements represents the existence of a certain keyword. We assume that the use of legal terms slightly differs by statute type. One example is that the constitution of Thailand has used only *amnat-nathi* and has not used *amnat-lae-nathi* or *nathi-lae-amnat*.

- **Section Keyword Feature**
  The section keyword feature denotes keywords of the section to which the statutory sentence belongs. As with the title keyword feature, we use this feature as a n-dimensional boolean variable. We introduce this feature to cope with legal terms having only a few adjacent words. Figure 3 demonstrates an example. In the case of Figure 3, *kharachakan-kanmueang* (ข้าราชการการเมือง) in line 4 is a legal term to be predicted. However, only the word เป็น (*pen*; being) is given as a meaningful feature if we use only adjacent words in the sentence as features. Therefore, we use the section keywords as additional features to solve this problem. In this case, the sentence in line 1 is the section (มาตรา; matra), so that keywords of this sentence are used as the section keyword feature.

Algorithm 2 Our algorithm

Input: W, y, K^t, K^s, T
Output: Suggests

```
Suggests ← ∅
for all (i, j) such that w_i w_{i+1} ··· w_j = t ∈ T do
    W^ℓ ← w_1 w_2 ··· w_{i−1}
    W^r ← w_{j+1} w_{j+2} ··· w_{|W|}
    t_best ← arg max_{t'∈T} score(W^ℓ, t', W^r, y, K^t, K^s)
    if t ≠ t_best then
        Suggests ← Suggests ∪ {suggestion that t in position (i, j) should be replaced into t_best}
    end if
end for
```

4.2 Prediction Model

Because we use the additional features to predict legal terms, we slightly modify the legal term correction task as follows:

- Statutory sentence $W = w_1\ w_2 \cdots w_{|W|}$, year feature y, title keyword feature K^t, section keyword feature K^s, and a set of legal terms $T \subseteq V^+$ are given, where K^t and K^s are a subset of vocabulary V;
- The adequacy of each legal term t found in W is judged;
- If another legal term $t_{\text{best}} \in T$ $(t_{\text{best}} \neq t)$ seems more adequate in the context, t_{best} is suggested to replace t.

Algorithm 2 is a general algorithm for this problem, where the input and scoring function are modified.

We utilize Random Forest as the scoring function $\text{score}(W^\ell, t', W^r, y, K^t, K^s)$, which is calculated by the following equation:

$$\begin{aligned}&\text{score}(W^\ell, t, W^r, y, K^t, K^s)\\ &= \sum_{d \in D} P_d(t | w^\ell_{|W^\ell|-N+1}, \ldots, w^\ell_{|W^\ell|}, w^r_1, \ldots, w^r_N,\\ &\quad y, k^t_1, \ldots, k^t_{|K^t|}, k^s_1, \ldots, k^s_{|K^s|}) = \sum_{d \in D} P_d(t|F), \quad (2)\end{aligned}$$

where D is a set of decision trees, d is a decision tree, and $P_d(t|F)$ is the probability that d chooses t based on the features F. Here, w^ℓ_i and w^r_i are the i-th words of W^ℓ and W^r, respectively. y is the year feature, and

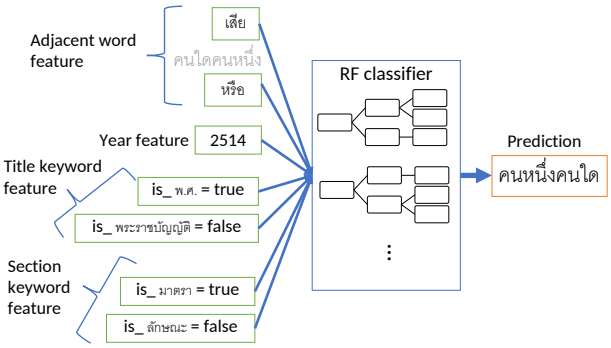


Figure 4: Our model

k^t_i and k^s_i are the existence of the i-th keyword in the title sentence and section sentence, respectively. N is the window size. Figure 4 expresses the input and output of this model.

5 Experiment

To evaluate the effectiveness of our method, we conducted an experiment on predicting legal terms in Thai statutory sentences.

5.1 Outline of Experiment

We compiled a statutory sentence corpus from the website of the Office of the Council of State [1]. We acquired 7,399 Thai statutes that include constitutions, codes, emergency decrees, royal decrees, ordinances, regulations, orders, notices, and more. There are 7,516,792 tokens and 66,671 different words in the corpus after tokenization by PyThaiNLP (v.1.7) [2]. We created the dataset using the following procedure: (1) extract all sentences where more than one legal term appears; (2) unify the sentences so that there are no identical sentences in the dataset; (3) make datasets for each legal term by grouping sentences based on the legal terms contained within; (4) split each dataset into five for five-fold cross validation; then (5) process each sentence to an example for each method.

We defined five legal term sets by referencing the Thai legislation manual (Office of the Council of State, 2008). Table 1 shows each legal term and its number of total occurrences.

[1] http://www.krisdika.go.th/
[2] https://github.com/PyThaiNLP/pythainlp

Table 1: Legal terms

Term set	Legal Term	Counts
Set1-1	*yang-nueng-yang-dai*	1,469
	yang-dai-yang-nueng	1,152
Set1-2	*khon-dai-khon-nueng*	489
	khon-nueng-khon-dai	268
Set2	*amnat-nathi*	5,631
	amnat-lae-nathi	977
	nathi-lae-amnat	519
Set3	*panakngan-chaonathi*	8,006
	chaonathi	4,579
Set4	*kharachakan-kanmueang*	595
	phu-damrong-tamnaeng -thang-kanmueang	411
Total		24,096

We compared our method (Random Forest with additional features; RF+) with Yamakoshi et al. (2018)'s Random Forest (RF) and BERT (Devlin et al., 2018). As a baseline, we also tried maximum likelihood estimation (MLE), which always selects the most frequent legal terms in the training data. For evaluation, we averaged the accuracies of each legal term set in the five datasets.

For the Random Forest methods, we set hyper-parameters as follows: the estimator number is 500; the maximum depth of a decision tree is unlimited; and the window size is 15. We tokenized each sentence by PyThaiNLP (v.1.7). Implementation, training, and testing are done by Scikit-learn (v.0.19.1).

For RF+, we used the most frequent 1,000 words in titles and sections as the keywords of the title and section, respectively. Here, we excluded some functional words using the stopword vocabulary in PyThaiNLP (v.1.7). We also excluded legal terms from the section keywords to prevent them from becoming clues to predict the legal term.

For BERT, we used the *BERT-Base, Multilingual Cased model* [3] that is offered by the authors of the paper (Devlin et al., 2018). The pretrained model has 12 Transformer layers and each layer's unit contains 768 hidden values. We replaced the target legal term in every example into a meta token "^" that is not used in the corpus, so that the model will predict the legal term based on the context around the token. The model accepts a sequence of a maximum 128 subwords and almost all subwords defined in its vocabulary consist of one character. Therefore, we truncated each example so that one example has at most 128 characters. Other hyper-parameters are as follows: the number of epochs is 20; batch size is 32; learning rate is 2e-5; and warmup proportion is 0.1. Implementation, training, and testing were done by Tensorflow on Colaboratory [4].

Table 2: Experimental results

Term set	MLE	BERT	RF	RF+
Set1-1	56.0%	85.4%	83.8%	**86.6%**
Set1-2	64.6%	**93.4%**	90.2%	91.8%
Set2	79.0%	85.5%	84.6%	**89.4%**
Set3	63.6%	**95.2%**	89.3%	94.4%
Set4	59.1%	**95.1%**	89.0%	93.4%
Average	67.2%	91.2%	87.3%	**92.0%**

5.2 Experimental Results

Table 2 shows the experimental results of each model. RF+ acheived the best accuracy in Set2, Set4-1, and overall accuracy. In every legal term set, RF+ achieved better performance than RF.

6 Discussion

In this section, we investigate the experimental results in more detail to reveal the characteristics and effectiveness of our method.

First, we decompose the experimental results per legal term in order to determine whether our method is good at predicting legal terms. Table 3 shows the accuracies of each legal term (averaged in results of five-fold cross validation). According to Table 3, RF+ achieved the best accuracy on average. It is also noteworthy that RF+ performed better than RF for almost every legal term except *kharachakan-kanmueang*, especially for *nathi-lae-amnat*. However, although RF has the same characteristic, RF+ tends to choose more frequent legal terms so that the

[3] https://github.com/google-research/bert

[4] https://colab.research.google.com/

Table 3: Accuracy per legal term

Legal term	Count	BERT	RF	RF+
yang-nueng-yang-dai	1,469	88.1%	91.8%	**95.4%**
yang-dai-yang-nueng	1,152	**81.9%**	73.4%	75.4%
khon-dai-khon-nueng	489	97.1%	97.5%	**98.4%**
khon-nueng-khon-dai	268	**86.6%**	76.9%	80.0%
amnat-nathi	5,631	93.2%	97.8%	**98.5%**
amnat-lae-nathi	977	**64.0%**	43.5%	53.1%
nathi-lae-amnat	519	41.3%	19.0%	**59.9%**
panakngan-chaonathi	8,006	96.1%	97.4%	**98.1%**
chaonathi	4,579	**93.6%**	75.1%	88.0%
kharachakan -kanmueang	595	**97.6%**	96.5%	96.2%
phu-damrong-tamnaeng -thang-kanmueang	411	**91.4%**	78.3%	89.7%
Average		84.6%	77.0%	**84.8%**

Table 4: Most important features

#	Set1-1	Set1-2	Set2	Set3	Set4
1	w-1	w+2	y	w+1	w-3
2	w-4	w-3	t-k_1	w+2	w-2
3	w-2	w-1	w-2	t-k_2	y
4	w+1	w+5	w-3	t-k_3	w-1
5	w+3	w+1	w+2	y	t-k_1
6	w-5	w-9	w-5	s-k_4	w-4
7	y	w-2	t-k_2	t-k_5	s-k_6
8	w-9	w+3	w-4	w-1	w+2
9	w-7	w-4	w-7	w+3	w+5
10	w-6	w-7	w-6	w+4	w+3

accuracies of less frequent legal terms are generally lower than those of the BERT method.

Next, we look at the feature importance of Random Forest classifiers. Table 4 shows the 10 most important features for each legal term set. In Table 4, "w+i" means the i-th right word, "w-i" means the i-th left word, "y" means the year feature, t-k indicates the existence of keyword k in the title, and s-k indicates the existence of keyword k in the section. Here, k_1, k_2, k_3, k_4, k_5, and k_6 mean รัฐธรรมนูญ (*ratthamnun*; constitution), ว่าด้วย (*waduai*; regarding), มาตรา (*matra*; section), รัฐ (*rat*; state), เจ้าหน้าที่ (*chaonathi*; officer), and ศาลฎีกา (*sandika*; supreme court), respectively.

Although most of the important features were adjacent words, the year feature and some keywords became important features. For example, the year feature was the most important one in Set2 (*amnat-nathi*, *amnat-lae-nathi*, and *nathi-lae-amnat*). This is because *amnat-lae-nathi* is a newer legal term (refer to Section 2.2). Also, รัฐธรรมนูญ (*ratthamnun*; constitution) is an important keyword in the legal term set because constitutions only use *amnat-nathi* out of the three legal terms.

The advantage of our RF+ model is not only prediction performance, but also feasibility. In terms of training cost, we need just an ordinary personal computer to train our RF+ model, while we need a TPU (Tensor Processing Unit) and at least a GPU environment to train a BERT model. In addition to that, our RF+ model is quite small compared to a BERT model. In the settings of our experiment, the total amount of RF+ models was less than 40 MB (varying from 2 MB to 20 MB per legal term set), while the total amount of BERT models was about 8 GB (1.6 GB per legal term set), which was 200 times larger than the RF+ models.

7 Summary

In this paper, we proposed a legal term correction method for Thai statutory sentences. Our method uses Random Forest classifiers to determine each legal term, to which we introduced three types of additional features from outside of the sentence: year feature, title keyword feature, and section keyword feature. Our experiment has shown that our method outperformed not only the existing Random Forest-based method, but also a method with BERT, the state-of-the-art language representation model, in overall accuracy.

Acknowledgments

This work was partly supported by JSPS KAKENHI Grant Numbers 18H03492 and the Graduate Program for Real-World Data Circulation Leaders, Nagoya University.

References

Leo Breiman. 2001. Random Forests. *Machine Learning*, 45:5–32.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv preprint, arXiv:1810.04805, 13 pages.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. *Efficient Estimation of Word Representations in Vector Space*. In *Proceedings of the International Conference on Learning Representations*, 12 pages.

Andriy Mnih and Koray Kavukcuoglu. 2013. *Learning Word Embeddings Efficiently with Noise-contrastive Estimation*. In *Proceedings of the Advances in Neural Information Processing Systems 26*, pages 2265–2273.

Koki Mori, Makoto Miwa, and Yutaka Sasaki. 2015. *Sentence Completion by Neural Language Models Using Word Order and Co-occurrences*. In *Proceedings of the 21st Annual Meeting of the Association for Natural Language Processing*, pages 760–763 (In Japanese).

Office of the Council of State. 2008. *Legislative Drafting Manual* (In Thai).

Rachata Tipchod and Viriya KhotchaSeni. 2013. *Thai Law Dictionary Thai – English*. Soutpaisal Press, Thailand.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. *Attention is All You Need*. In *Proceedings of Advances in Neural Information Processing Systems 30*, pages 6000–6010.

Takahiro Yamakoshi, Takahiro Komamizu, Yasuhiro Ogawa, and Katsuhiko Toyama. 2018. *Japanese Legal Term Correction using Random Forest*. In *Legal Knowledge and Information Systems, JURIX 2018: The Thirty-first Annual Conference*, 313:161–170, IOS Press, the Netherlands.

Geoffrey Zweig and Chris J.C. Burges. 2011. *The Microsoft Research Sentence Completion Challenge*. Technical Report of *Microsoft Research*.

A Corpus of Sentence-level Annotations of Local Acceptability with Reasons

Wonsuk Yang **Jung-Ho Kim** **Seungwon Yoon**
ChaeHun Park **Jong C. Park**[†]
School of Computing
Korea Advanced Institute of Science and Technology
{derrick0511, jhkim, swyoon, ddehun, park}@nlp.kaist.ac.kr

Abstract

News editorials are presented with arguments of different quality, which readers may or may not accept as presented. In this work, we present a corpus of news editorials with sentence-level annotations of local acceptability and a set of related attributes, where the annotators also provided detailed reasons in natural language for each attribute. The annotations were performed in both in-house and crowdsourcing environments. In total, 105 news editorials were annotated for 3,591 sentences, with an average of four annotations from different annotators per sentence, resulting in about 14K sentence-level annotations with detailed reasons in natural language (in total 121K tokens written in 1K aggregated hours). We analyze the reasons to see why given information is accepted or rejected, examine the correlation among the attributes, and compare our annotation result against argumentation strategies. This is the first corpus to provide sentence-level annotations with attributes such as local acceptability, which we argue is critical for a fine-grained and advanced analysis of argumentation quality.

1 Introduction

Argumentation is an activity aimed at promoting the acceptability of a controversial standpoint (van Eemeren et al., 1996; Stab and Gurevych, 2017). The research on argumentation quality is of much relevance to computer-assisted writing. *Local acceptability* is considered as one of the important factors that influence the quality of argumentation in writing, which is defined for the premises of an argument as to how much they are "rationally worthy of being believed to be true" (Wachsmuth et al., 2017). It also reveals the process where a particular reader accepts or rejects information delivered by each premise presented in an argument while reading it. A deeper understanding into this process is thus vital for the development of computer-assisted writing systems that can identify weak spots of an author's argument.

Development of such a system has also been considered as an important goal of many studies on computational argumentation such as parsing argumentation structures (Stab and Gurevych, 2017). Recently, there have been studies on the annotation and analysis of eloquence, evidence, and more attributes that give multiple types of feedback on persuasiveness for effective writing support systems (Carlile et al., 2018). However, persuasiveness is an indicator of how "persuasive" a particular sentence is, and takes a different dimension from local acceptability (Wachsmuth et al., 2017). For instance, it is reported that the inter-annotator agreement for persuasiveness (0.5-0.7 alpha, (Carlile et al., 2018)) is quite higher than that for local acceptability (0.22-0.45 alpha, (Wachsmuth et al., 2017)), indicating that the latter is arguably more influenced by individual subjectivity. Wachsmuth et al. (2017) proposed local acceptability as one of the 15 factors affecting argumentation quality, with which they also annotated textual debate portal arguments for two stances on several issues, such as evolution vs. creation. However, as all the attributes were scored only with a 1-3 scale and with argument-level (paragraph-

[†]Corresponding author

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 288-297, Hakodate, Japan, September 13-15, 2019

	Score	Description
Local Acceptability	7	I **strongly accept** the information given by the sentence to be true. I have sound and cogent arguments to justify my acceptance.
	6	I **accept** the information given by the sentence to be true. I have some arguments to justify my acceptance.
	5	I **weakly accept** the information given by the sentence to be true. I do not have arguments justifying my acceptance. Still, I will accept it rather than reject it.
	4	It is **hard to judge** whether I should accept or reject the information given by the sentence to be true.
	3	I **weakly reject** the information given by the sentence to be true. I do not have arguments for the rejection. Still, I will reject it rather than accept it.
	2	I **reject** the information given by the sentence to be true, and I have arguments for the rejection.
	1	I **strongly reject** the information given by the sentence to be true. I have sound and cogent arguments for the rejection.
Knowledge Awareness	3	I **already knew** the information before I read this document.
	2	I **did not know** the information before I read this document, **but came to know** it by reading the previous sentences in this document.
	1	I **did not know** the information.
Verifiability	5	I can verify it **using my knowledge**. It is a common sense. I do not need to google it to verify.
	4	I can verify it by **short-time googling**.
	3	I can verify it by **long-time googling**. I could verify it using deduction if I google it for some time for deeper understanding.
	2	I might find an **off-line way** to verify it, but it will be very hard. It needs specific witness or testimony to verify, and there may not be any evidence in written form.
	1	There is **no way to verify** it.
Disputability	4	Whether or not it is reasonable to accept the information given by the sentence as true, it is **not disputable**.
	3	Whether or not it is reasonable to accept the information given by the sentence as true, it is **weakly disputable**.
	2	Whether or not it is reasonable to accept the information given by the sentence as true, it is **disputable**.
	1	Whether or not it is reasonable to accept the information given by the sentence as true, it is **highly disputable**.

Table 1: Description of local acceptability, knowledge awareness, verifiability, and disputability.

level) granularity, it is not possible to track a specific process where the reader accepts or rejects each premise presented in the argument. This is important because knowing at which sentence a reader starts to reject the given information enables the writing support system to determine from where it needs to start editing.

In this work, we present a corpus of sentence-level annotations of local acceptability and a set of predefined, possibly related attributes for each sentence in news editorials. For a deeper understanding into the individual judgments, we also collected the reasons for the particular attribute values by each annotator on each sentence. The corpus can thus be utilized to experimentally verify to what extent it is possible to predict the very subjective reaction of the target readers accepting or rejecting the information given by each sentence.

The contributions of this paper are as follows. (1) We introduce a large corpus of 105 news editorials with 14K sentential annotations for local acceptability and three possibly related attributes, with reasons for each attribute written in natural language, in both in-house and crowdsourced settings. The gathered text for the reasons amounts to 121K tokens written in 1K aggregated hours[1]. (2) We show experimentally that the three attributes are meaningfully correlated with local acceptability. (3) We provide a detailed analysis of the reasons provided by the annotators for each of the attributes. (4) We provide key insights through the comparison against argumentation strategy, and suggest that our corpus can be utilized for future research that leads to computational argumentation.

2 Related Work

Traditionally, research on computational argumentation has focused on parsing argumentation structures, whose goal is to identify argument components of a given document such as major claims, other claims, and premises, and to analyze how they are related to one another (Stab and Gurevych, 2014a; Stab and Gurevych, 2014b; Stab and Gurevych, 2017).

Recently, research on the quality of argumentation has received much attention. Persing and Ng (2015) defined argument strength based on the expected number of readers who would be persuaded, and annotated student essays with it. For convincingness, which concerns the universal audience (Perelman et al., 1969), there has been an in-depth study to see which of two given arguments is more convincing and why (Habernal and Gurevych, 2016a; Habernal and Gurevych, 2016b). For persuasiveness, which concerns a particular audience

[1]Our corpus is available at http://credon.kaist.ac.kr.

Local Acceptability	strong accept	statement factual nature, author state facts, personally agree statement, evidence support claim, commonly know fact
	accept	author quote another, author sufficient credibility, would grind author, author focus part, reveal subjective interpretation
	weak accept	author reveal subjective, reveal subjective interpretation, author sufficient credibility, likely thing happen, not know enough
	hard to judge	not enough background, not know enough, enough background knowledge, make judgements statement, not enough knowledge
	weak reject	personally not agree, author lack credibility, not agree statement, author reveal subjective, opinion not fact
	reject	controversial author opinion, claim base controversial, claim controversial difficult, one side opinion, long search would
	strong reject	provide counter examples, evidence easily dismiss, not logical reason, highly inaccurate projections, find speculative highly

Table 2: The top five most frequent trigrams (lemmatized verbs, cf. Section 4.3.1) used for describing the reason for local acceptability. (*grind* is a lemmatization error.)

(Perelman et al., 1969), Carlile et al. (2018) annotated argument components, such as major claims, other claims, and premises, with component-specific sub-attributes such as eloquence and evidence. Ke et al. (2018) developed neural network systems to predict persuasiveness and sub-attributes. Tan et al. (2016) utilized the Reddit forum ChangeMyView, used the record of user interactions as the proxy to measure the persuasiveness of arguments in the forum, and studied multiple factors that infludence persuasiveness. El Baff et al. (2018) modeled the argumentation quality of news editorials based on whether they challenge or reinforce the stance of the reader, and gathered document-level annotations for 1,000 news editorials on the model. In addition, as part of the SemEval 2018 shared task, the Argument Reasoning Comprehension task (Habernal et al., 2018) has been offered to select the appropriate warrant for a given argument consisting of a claim and a reason. While we focus on local acceptability in this work, there is another type of acceptability of an argument or argumentation, called global acceptability. It is investigated by Cabrio and Villata (2012) who identified ground-truth debate portal arguments using textual entailments based on the formal argumentation framework of Dung (1995), and assessed the global acceptability of the arguments.

3 Data

The source data we choose to annotate is composed of 105 news editorials randomly chosen from the Webis-Editorials-16 corpus provided by Al-Khatib et al. (2016), which includes news editorials published by Al-Jazeera, FoxNews, and the Guardian. This corpus classifies argumentative discourse units (ADUs) into the following six types for the analysis of argumentation strategies: (1) *Common Ground*, (2) *Assumption*, (3) *Testimony*, (4) *Statistics*, (5) *Anecdote*, and (6) *Other*.

We used their corpus as the basis for our annotation for the following three reasons. (1) The Webis-Editorial-16 corpus is annotated with reliable quality. (2) The corpus has been used as the basis for an analytical study of topic-dependent argumentation strategies (Al-Khatib et al., 2017) and has inspired subsequent research on argumentation synthesis through rhetoric strategies (Wachsmuth et al., 2018). (3) We anticipate that the six types used for the analysis of argumentation strategies would be closely related to local acceptability. For example, the local acceptability for *Statistics* is expected to be higher than that for *Anecdote*. Therefore, we anticipate that the annotations of argumentation strategy will help the quality assessment of local acceptability annotation (see Section 4.3.5).

4 Annotation

4.1 Annotation Scheme

In this study, we annotate each of the sentences in news editorials with local acceptability and a set of predefined attributes related to it.

Our definition of local acceptability follows Wachsmuth et al. (2017), who defined local accept-

Figure 1: The word cloud for the trigram frequencies for *accept* (left) and *reject* (right) for local acceptability.

ability of a premise as: A premise is locally acceptable if it is "rationally worthy of being believed to be true." It is noted that an argument is composed of a claim and one or more premises, where a premise is a reason for justifying (or refuting) a claim and the claim is a possibly controversial statement and the central argument component (Stab and Gurevych, 2017). We define the local acceptability of a sentence based on the truth-value of the sentence following the *truth-conditional theory* (Lewis, 1970): A sentence is locally acceptable if the truth-value of the sentence is rationally worthy of being believed to be true. For complex cases where the truth-values of the phrases in a sentence are combined to generate the truth-value of a sentence, we also follow Lewis (1970).

We also annotated three possibly related attributes as follows. *Knowledge Awareness* asks whether or not an annotator already knew the information. *Verifiability* indicates how easy it is to verify the information. *Disputability* is about how controversial the information is. We chose the attributes for a deeper understanding of local acceptability, focusing on fact-checking (Wintersieck et al., 2018) and journalistic aspects (Cheruiyot and Ferrer-Conill, 2018; Aharoni and Tenenboim-Weinblatt, 2019). Table 1 shows detailed rubrics for the local acceptability and the three attributes that we used for our annotation.

4.2 Annotation Procedure

4.2.1 In-house Annotations

An in-house annotation was conducted on 105 news editorials by eight undergraduate students with native competence in English, where four of them were student journalists responsible for the school newspaper. We introduced the rubrics to the students through one seminar so that they familiarized themselves with the rubrics over one week. Then, for the following two weeks, each student had two separate meetings with the authors for further instruction on the rubrics. The whole annotation process took about 6 months.

4.2.2 Crowdsourcing Annotations

We also conducted the annotation through the Amazon Mechanical Turk (AMT) crowdsourcing platform. Each AMT annotator received specific annotation guidelines, including a detailed description of such cases that could cause confusion, identified as such during the training period of the in-house annotation. In each assignment, the workers were presented with a URL for a news editorial. At the URL, they were asked to annotate following the guidelines, and to write the reasons for choosing each attribute value. The crowdsourced annotation took

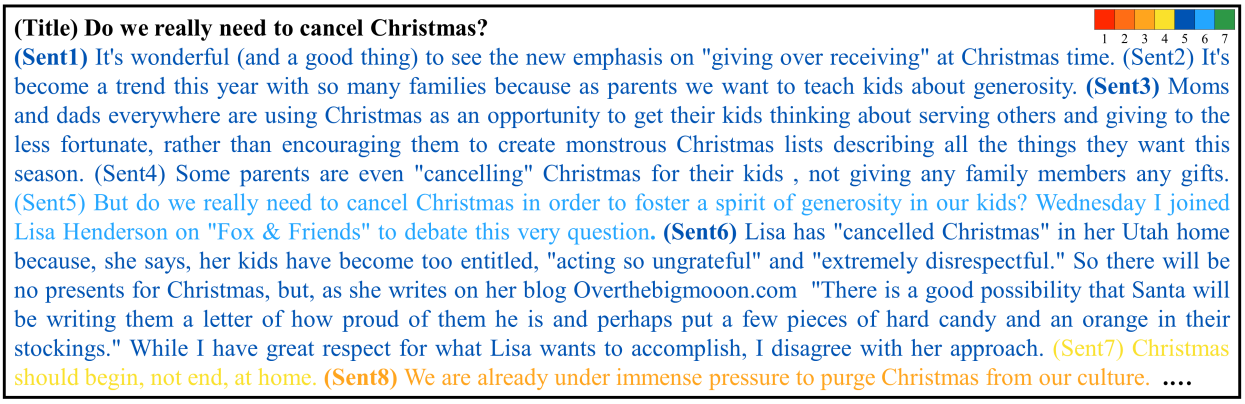


Figure 2: Example news editorial from the presented corpus. The color coding indicates the mean value of the Local Acceptability of each sentence across the three annotators. We rounded each mean value to the first decimal place. (article link: `https://www.foxnews.com/opinion/do-we-really-need-to-cancel-christmas`)

about two weeks, where it started after the in-house annotation was over. During the whole annotation process, 183 workers (with $\geq$ 95% acceptance rate for previous tasks in AMT) participated, where one annotator was allowed to annotate multiple news editorials.

4.2.3 Corpus Statistics

As a result, per news editorial, we obtained annotations from an average of 1.3 different students and from an average of 2.7 different workers (resulting in an average of 4.0 annotators per news editorial), amounting to 14,225 sentence-level annotations for 3,591 sentences in 105 news editorials in total. The students and workers spent about 4 and 5 minutes on average, respectively, to annotate a single sentence including the time to read and select the attribute values and to write down specific reasons, in total 1,108 aggregated hours for the 14,225 anntotations.

4.3 Analysis of Annotations

4.3.1 Reason

The students participating in the in-house annotation described the reasons using 4.2 words per attribute value on average, and the workers used 10.6 words on average. In calculating the average length, we filtered out all non-English words using Natural Language Toolkit (NLTK) [2] and did not count them.

Table 2 shows the top five most frequent trigrams for local acceptability, and Figure 1 shows word clouds of trigram frequencies of the reasons specified for the attribute values of *accept* and *reject*. For the trigram frequency, we preprocessed the reasons written in natural language as follows. We first removed all non-English words, stop words except for *not*, and punctuation marks from the reasons specified by the annotators. We also removed the words that are used to describe the corresponding attribute value itself in the rubrics, but did not remove those that are used to explain other attribute values. For example, in the case of *accept* of local acceptability, words such as *justify* that appear in the explanation of *accept* were removed, but words such as *cogent* that appear in the explanation of *strong accept* were not removed. We lemmatized verbs and measured trigram frequencies across all the (preprocessed) reasons. When one annotator repeatedly used a certain trigram for a specific attribute, it was counted only once.

From the trigram frequency, we find that the major reasons to accept a piece of information are factuality, related evidence, third party information source, and prominence of the information. The reasons to reject are controversy, bias, counter evidence, logical inconsistency, and non-factual nature of the information.

4.3.2 Example

In order to illustrate the overall process of how the (three) annotators accepted or rejected the sentences while reading a given news editorial, we present eight example sentences and the average of the three

[2] http://www.nltk.org

Sentences		**LA**	KA	V	D
(Sent3) Moms and dads everywhere are using Christmas as an opportunity to get their kids thinking about serving others and giving to the less fortunate , rather than encouraging them to create monstrous Christmas lists describing all the things they want this season .	Student1	(5) weak accept, because no counter evidence to reject	(1) did not know	(1) no way to verify, because general statements cannot be verified	(2) weakly disputable, because (…), as some parents may not agree
	Worker1	(3) weak reject, because it seems to be a rhetorical device to make a point.	(1) did not know, becaus e I am not aware that this is a widespread behavior.	(3) long-time googling, because it might be possi ble to find many individu al cases online (…)	(3) disputable, because this is an unsourced opinion.
	Worker2	(6) accept, because I can verify and provide supporting examples	(3) already knew, because (…) my family is trying to make [Christmas] less focused on material things	(4) short-time googling, because you can find blogs and stories of families doing this	(3) disputable, because y ou can argue that there is only a small group of pe ople doing this and that it isn't a Movement
(Sent6) Lisa has "cancelled Christmas" in her Utah home because, she says, her kids have become too entitled, "acting so ungrateful" and "extremely disrespectful." So there will be no presents for (…) While I have great respect for what Lisa wants to accomplish, I disagree with her approach.	Student1	(5) weak accept, because no reason to reject author 's personal opinion	(1) did not know	(1) no way to verify, because author's personal opinion on Lisa's approach to Christmas	(2) weakly disputable, because personal opinions on Christmas may vary
	Worker1	(6) accept, because it is what Lisa claims it is, and what the other speaker stated.	(1) did not know, because Lisa, her plan (…) are all introduced for the first time here.	(4) short-time googling, because it is from her blogoverthebigmoon.co m. The video (…)	(2) weakly disputable, because this is second hand information, and a personal opinion.
	Worker2	(5) weak accept, because I accept their points of view but it is hard to accept this statement as an absolute fact	(1) did not know, because (…) there is no way I would know this without knowing her or being one of her blog readers	(4) short-time googling, because you could read her blog article or talk with her and the interviewer to confirm her opinion	(2) weakly disputable, because (…) on the topic but you [could] if you found scientific or other evidence that proved them wrong
(Sent8) We are already under immense pressure to purge Christmas from our culture.	Student1	(4) hard to judge, because do not fully agree with the statement, but no reason to reject	(1) did not know	(1) no way to verify, because general statements cannot be verified	(3) disputable, because general statements can be controversial, especially involving the audience ("we")
	Worker1	(1) strong reject, because there is no immense pressure to purge Christmas in "our culture"	(1) did not know, because I know this to be untrue.	(4) short-time googling, because this is a common [right-wing] Christian view, but it does not represent reality.	(4) highly disputable, because this is simply not happening.
	Worker2	(3) weak reject, because I need more evidence to prove this is true without it I'm not sure if I believe the claims in this Story	(3) already knew, because I don't agree with this statement but [i] am aware there are people who believe this	(1) no way to verify, because this is an opinion and you can find articles (…)	(4) highly disputable, because some people believe there is a war on [christmas] (…)

Figure 3: The annotation results for the three sentences that we find the most interesting in the paragraph of Figure 2. Each color bar above the column name indicates the color coding for the values of the corresponding attribute in the column.

attribute values for local acceptability for each sentence (Figure 2). It is noted that the annotators start to reject the given sentence at Sent 8.

Moreover, we present the annotation results of the three sentences (Sent3, 6, 8) that we find the most interesting among the annotation results of the eight sentences (Figure 3). In the case of Sent3, we see that each annotator focused on different aspects: (1) counter-evidence (Student1), (2) rhetoric device (Worker1), and (3) verifiability and supporting examples (Worker2). Thus, they made different, subjective judgments on local acceptability for the same sentence. Sent6 was relatively more complex, and we found that each annotator focused on a different phrase in the sentence in accepting or rejecting it. In the case of Sent8, all three annotators responded negatively to local acceptability or gave the *hard to judge* response.

	Pearson Correlation Coefficient			
	LA	KA	V	D
LA		.235 (.154)	.466 (.402)	-.658 (-.481)
KA	.235 (.154)		.407 (.285)	-.118 (.040)
V	.466 (.402)	.407 (.285)		-.425 (-.397)
D	-.658 (-.481)	-.118 (.040)	-.425 (-.397)	

Table 3: Correlation between local acceptability and the related attributes. All the p-values were less than 0.001. The numbers in parentheses are for the in-house annotation only, whereas the numbers outside the parentheses are for the entire (i.e., both in-house and crowdsourced) annotation.

Overall, the example annotation results show that both the in-house students and the AMT workers produced high-quality annotations, as indicated by the overall length and specific content of the reasons. We also note that they tend to make a highly subjective, yet reasonable judgment, and that each of their judgments cannot be considered irrational simply because of such subjectivity and uniqueness.

4.3.3 Inter-Annotator Agreement

As can be seen in the examples in Section 4.3.2, the annotators responded to a sentence differently based on their (different) viewpoints. Even for the case where different annotators chose the same attribute value, the reason was not necessarily the same. Even when the annotators chose different attribute values, their reasons for the choices were apparently valid. From the fact that we can argue for the quality of the annotations in Figure 3 based on the validity of the reasons suggested by the annotators, we see that (the validity of) the reason itself is a good indicator for the quality of annotation. We do not use inter-annotator agreement (IAA) as the quality measure for our corpus in this paper. In Yang et al. (2019), we present an in-depth analysis on the IAA for local acceptability with a new method of quality control that uses the validity of the reasons.

4.3.4 Correlation Analysis

To understand the significance of how the attributes are related to local acceptability, we computed the Pearson's Correlation Coefficient (PC) between local acceptability and each of the attributes along with the corresponding p-values. Results are as shown in Table 3. For all of the calculated PC, the p-value was less than 0.001.

Overall, local acceptability and disputability show a high negative correlation. This is not surprising because it is hard to (strongly) accept highly disputable information. Local acceptability and verifiability show a positive correlation. We speculate that this is because it is less likely that an author delivers a piece of misinformation about easily verifiable information and it is more likely to be acceptable information. Local acceptability and knowledge awareness have a positive correlation. We speculate that this is because the information already known to a reader would be easier to be accepted by that reader. The results show that the three attributes that we speculated as possibly related are actually related to local acceptability.

4.3.5 Local Acceptability and Argumentation Strategy

We look into the relationship between our annotation results and previously annotated argumentation strategies for further insights. Note that during the entire annotation process, the annotators were not provided with information about the argumentation strategy pre-annotated in the corpus, and conducted annotation only on plain text.

Of a total of 3,591 sentences in the 105 news editorials, 3,150 (87.7%) sentences are found to contain only a single type of argumentation strategy (other than the type *Other*), and 441 (12.3%) sentences contain more than one type or do not contain any type. We counted the annotations for the same sentence by different annotators separately (as different annotations). As we note that the labels for argumentation strategy type were imbalanced, we compared the relative frequency (the ratio in the table) of the attribute values for each type. For example, we found that 1,802 annotations on 458 sentences were mapped to the type *Anecdote*, and that 657 (36.5%) of them had the local acceptability of *strong accept*. In this case, the relative frequency of *strong accept* for *Anecdote* is 36.5%. For local acceptability and each of the other attributes, we compared the relative frequencies across different argumentation strategy types in the same way, as shown in Table 4.

Strong accept of local acceptability occurred most frequently with *Statistics*, *Anecdote*, and *Common Ground*, whereas the relatively lower value *accept* occurred most frequently with *Assumption* and *Testimony* (not counting *N/A*). The local acceptability for

		Count (Ratio)						Total
		AS	AN	ST	TE	CO	N/A	
LA	strong accept	1461 (15.8)	657 (**36.5**)	200 (**40.9**)	234 (30.1)	70 (**36.3**)	383 (22.0)	3005 (21.1)
	accept	2341 (**25.4**)	553 (30.7)	175 (35.8)	305 (**39.2**)	66 (34.2)	498 (**28.6**)	3938 (**27.7**)
	weak accept	2232 (24.2)	373 (20.7)	73 (14.9)	127 (16.3)	28 (14.5)	343 (19.7)	3176 (22.3)
	hard to judge	1595 (17.3)	132 (7.3)	13 (2.7)	52 (6.7)	15 (7.8)	334 (19.2)	2141 (15.1)
	weak reject	719 (7.8)	51 (2.8)	12 (2.5)	38 (4.9)	7 (3.6)	87 (5.0)	914 (6.4)
	reject	573 (6.2)	23 (1.3)	11 (2.2)	12 (1.5)	4 (2.1)	64 (3.7)	687 (4.8)
	strong reject	300 (3.3)	13 (0.7)	5 (1.0)	10 (1.3)	3 (1.6)	33 (1.9)	364 (2.6)
KA	already knew	2137 (23.2)	258 (14.3)	44 (9.0)	47 (6.0)	125 (**64.8**)	255 (14.6)	2866 (20.1)
	came to know	1233 (13.4)	135 (7.5)	65 (13.3)	148 (19.0)	15 (7.8)	194 (11.1)	1790 (12.6)
	did not know	5851 (**63.5**)	1409 (**78.2**)	380 (**77.7**)	583 (**74.9**)	53 (27.5)	1293 (**74.2**)	9569 (**67.3**)
V	using my knowledge	1284 (13.9)	141 (7.8)	26 (5.3)	26 (3.3)	82 (**42.5**)	142 (8.2)	1701 (12.0)
	short-time googling	1712 (18.6)	552 (30.6)	232 (**47.4**)	296 (**38.0**)	42 (21.8)	477 (**27.4**)	3311 (23.3)
	long-time googling	1971 (21.4)	557 (**30.9**)	211 (43.1)	264 (33.9)	39 (20.2)	445 (25.5)	3487 (**24.5**)
	off-line way	763 (8.3)	275 (15.3)	12 (2.5)	125 (16.1)	5 (2.6)	170 (9.8)	1350 (9.5)
	no way to verify	2610 (**28.3**)	229 (12.7)	4 (0.8)	53 (6.8)	19 (9.8)	281 (16.1)	3196 (22.5)
	none of the above	881 (9.6)	48 (2.7)	4 (0.8)	14 (1.8)	6 (3.1)	227 (13.0)	1180 (8.3)
D	not disputable	2562 (27.8)	1318 (**73.1**)	358 (**73.2**)	496 (**63.8**)	101 (**52.3**)	882 (**50.6**)	5717 (**40.2**)
	weakly disputable	2875 (**31.2**)	298 (16.5)	82 (16.8)	139 (17.9)	56 (29.0)	425 (24.4)	3875 (27.2)
	disputable	2765 (30.0)	141 (7.8)	34 (7.0)	115 (14.8)	25 (13.0)	326 (18.7)	3406 (23.9)
	highly disputable	1019 (11.1)	45 (2.5)	15 (3.1)	28 (3.6)	11 (5.7)	109 (6.3)	1227 (8.6)
	Total	9221 (100)	1802 (100)	489 (100)	778 (100)	193 (100)	1742 (100)	14225 (100)

Table 4: Occurrence frequencies of the attribute values for each argumentation strategy type. The number in parenthesis is the relative occurrence frequency of an attribute value (row index) for an argumentation strategy type (column index). For each strategy type, the highest value of the relative frequency for each attribute is marked bold. AS, AN, ST, TE, and CO indicate *Assumption*, *Anecdote*, *Statistics*, *Testimony*, and *Common Ground*, respectively. N/A indicates the annotations for the sentences that contain more than one type or do not contain any type (other than the type *Other*).

Assumption was higher than we initially expected. We speculate that this is due to the accumulated credibility of the publishers (Al-Jazeera, FoxNews, and the Guardian) and/or their authors, which may indicate the importance of the author credibility or authority perceived by the readers.

For knowledge awareness, *did not know* occurred most frequently with all the types except for *Common Ground*, and *already knew* occurred most frequently with *Common Ground*. For verifiability, *using my knowledge* occurred most frequently with *Common Ground*, whereas *no way to verify* occurred most frequently with *Assumption*. *Short-time googling* occurred most frequently with *Statistics* and *Testimony*, and *long-time googling* occurred most frequently with *Anecdote*. For disputability, *not disputable* occurred most frequently with all the types except for *Assumption*, whereas *weakly disputable* occurred most frequently with *Assumption*. As such, we find that the relationship between the argumentation strategies and our attributes indicated by the relative frequencies precisely matches good linguistic intuition, suggesting further that our corpus is overall of remarkable quality.

5 Conclusion

In this study, we presented a corpus of 105 news editorials, in which each sentence is annotated with local acceptability and a predefined set of related attributes where reasons for each attribute are also provided in natural language by the annotators. We collected the reasons accepting or rejecting the information given by each sentence, in total 121K tokens written in 1K aggregated hours. A detailed analysis demonstrates that our corpus is overall of remarkable quality. We anticipate that our corpus, the first of its kind at sentence-level annotation with notes, may be utilized meaningfully for computer-assisted writing, as well as for a deeper understanding into the argumentation strategy. Our corpus is available at http://credon.kaist.ac.kr.

Acknowledgements

This work was supported by Institute for Information and communications Technology Promotion (IITP) grant funded by the Korea government (MSIT) (No. 2018-0-00582-002, Prediction and augmentation of the credibility distribution via linguistic analysis and automated evidence document collection).

References

Tali Aharoni and Keren Tenenboim-Weinblatt. 2019. Unpacking journalists(dis) trust: Expressions of suspicion in the narratives of journalists covering the israeli-palestinian conflict. *The International Journal of Press/Politics*.

Khalid Al-Khatib, Henning Wachsmuth, Johannes Kiesel, Matthias Hagen, and Benno Stein. 2016. A news editorial corpus for mining argumentation strategies. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3433–3443.

Khalid Al-Khatib, Henning Wachsmuth, Matthias Hagen, and Benno Stein. 2017. Patterns of argumentation strategies across topics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1351–1357.

Elena Cabrio and Serena Villata. 2012. Combining textual entailment and argumentation theory for supporting online debates interactions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL, Short Papers)*, volume 2, pages 208–212.

Winston Carlile, Nishant Gurrapadi, Zixuan Ke, and Vincent Ng. 2018. Give me more feedback: Annotating argument persuasiveness and related attributes in student essays. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL, Long Papers)*, volume 1, pages 621–631.

David Cheruiyot and Raul Ferrer-Conill. 2018. "Fact-Checking Africa" epistemologies, data and the expansion of journalistic discourse. *Digital Journalism*, 6(8):964–975.

Phan Minh Dung. 1995. On the acceptability of arguments and its fundamental role in nonmonotonic reasoning, logic programming and n-person games. *Artificial Intelligence*, 77(2):321–357.

Roxanne El Baff, Henning Wachsmuth, Khalid Al-Khatib, and Benno Stein. 2018. Challenge or empower: Revisiting argumentation quality in a news editorial corpus. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 454–464.

Ivan Habernal and Iryna Gurevych. 2016a. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1214–1223.

Ivan Habernal and Iryna Gurevych. 2016b. Which argument is more convincing? analyzing and predicting convincingness of web arguments using bidirectional lstm. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL, Long Papers)*, volume 1, pages 1589–1599.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018. SemEval-2018 Task 12: The Argument Reasoning Comprehension Task. In *Proceedings of The 12th International Workshop on Semantic Evaluation*, pages 763–772.

Zixuan Ke, Winston Carlile, Nishant Gurrapadi, and Vincent Ng. 2018. Learning to Give Feedback: Modeling Attributes Affecting Argument Persuasiveness in Student Essays. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 4130–4136.

David Lewis. 1970. General semantics. *Synthese*, 22(1):18–67.

Chaim Perelman, Lucie Olbrechts-Tyteca, John Wilkinson, and Purcell Weaver. 1969. *The new rhetoric: a treatise on argumentation*. University of Notre Dame Press.

Isaac Persing and Vincent Ng. 2015. Modeling argument strength in student essays. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP, Long Papers)*, volume 1, pages 543–552.

Christian Stab and Iryna Gurevych. 2014a. Annotating argument components and relations in persuasive essays. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pages 1501–1510.

Christian Stab and Iryna Gurevych. 2014b. Identifying argumentative discourse structures in persuasive essays. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 46–56.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion

strategies in good-faith online discussions. In *Proceedings of the 25th International Conference on World Wide Web (WWW)*, pages 613–624.

Frans H. van Eemeren, Rob Grootendorst, Francisca Snoeck Henkemans, J. Anthony Blair, Ralph H. Johnson, Erik C. W. Krabbe, Christian Plantin, Douglas N. Walton, Charles A. Willard, John Woods, and David Zarefsky. 1996. *Fundamentals of argumentation theory: A handbook of historical backgrounds and contemporary developments.* Lawrence Erlbaum Associates, Inc.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL, Long Papers)*, volume 1, pages 176–187.

Henning Wachsmuth, Manfred Stede, Roxanne El Baff, Khalid Al-Khatib, Maria Skeppstedt, and Benno Stein. 2018. Argumentation Synthesis following Rhetorical Strategies. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING)*, pages 3753–3765.

Amanda Wintersieck, Kim Fridkin, and Patrick Kenney. 2018. The message matters: The influence of fact-checking on evaluations of political messages. *Journal of Political Marketing*, pages 1–28.

Wonsuk Yang, Seungwon Yoon, Ada Carpenter, and Jong C. Park. 2019. Nonsense!: Quality control via two-step reason selection for annotating local acceptability and related attributes in news editorials. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). (*to appear).

Explicit Contextual Semantics for Text Comprehension

Zhuosheng Zhang[1,2,3,*], **Yuwei Wu**[1,2,3,4,*], **Zuchao Li**[1,2,3], **Hai Zhao**[1,2,3,†]
[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]Key Laboratory of Shanghai Education Commission for Intelligent Interaction
and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, China
[3]MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, China
[4]College of Zhiyuan, Shanghai Jiao Tong University, China
{zhangzs,will8821,charlee}@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn

Abstract

Who did what to whom is a major focus in natural language understanding, which is right the aim of semantic role labeling (SRL) task. Despite of sharing a lot of processing characteristics and even task purpose, it is surprisingly that jointly considering these two related tasks was never formally reported in previous work. Thus this paper makes the first attempt to let SRL enhance text comprehension and inference through specifying verbal predicates and their corresponding semantic roles. In terms of deep learning models, our embeddings are enhanced by explicit contextual semantic role labels for more fine-grained semantics. We show that the salient labels can be conveniently added to existing models and significantly improve deep learning models in challenging text comprehension tasks. Extensive experiments on benchmark machine reading comprehension and inference datasets verify that the proposed semantic learning helps our system reach new state-of-the-art over strong baselines which have been enhanced by well pretrained language models from the latest progress.

1 Introduction

Text comprehension is challenging for it requires computers to read and understand natural language texts to answer questions or make inference, which is indispensable for advanced context-oriented dialogue (Zhang et al., 2018d; Zhu et al., 2018) and interactive systems (Chen et al., 2015; Huang et al., 2018; Zhang et al., 2019a). This paper focuses on two core text comprehension (TC) tasks, *machine reading comprehension* (MRC) and *textual entailment* (TE).

One of the intrinsic challenges for text comprehension is semantic learning. Though deep learning has been applied to natural language processing (NLP) tasks with remarkable performance (Cai et al., 2017; Zhang et al., 2018a; Zhang and Zhao, 2018; Bai and Zhao, 2018; Zhang et al., 2019b; Xiao et al., 2019), recent studies have found deep learning models might not really understand the natural language texts (Mudrakarta et al., 2018) and vulnerably suffer from adversarial attacks (Jia and Liang, 2017). Typically, an MRC model pays great attention to non-significant words and ignores important ones. To help model better understand natural language, we are motivated to discover an effective way to distill semantics inside the input sentence explicitly, such as semantic role labeling, instead of completely relying on uncontrollable model parameter learning or manual pruning.

Semantic role labeling (SRL) is a shallow semantic parsing task aiming to discover *who* did *what* to *whom*, *when* and *why* (He et al., 2018; Li et al., 2018a, 2019), providing explicit contextual semantics, which naturally matches the task target of text comprehension. For MRC, questions are usually formed with *who*, *what*, *how*, *when* and *why*, whose predicate-argument relationship that is supposed to be from SRL is of the same importance as well. Be-

*These authors contribute equally. † Corresponding author. This paper was partially supported by National Key Research and Development Program of China (No. 2017YFB0304100) and Key Projects of National Natural Science Foundation of China (U1836222 and 61733011).

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 298-308, Hakodate, Japan, September 13-15, 2019

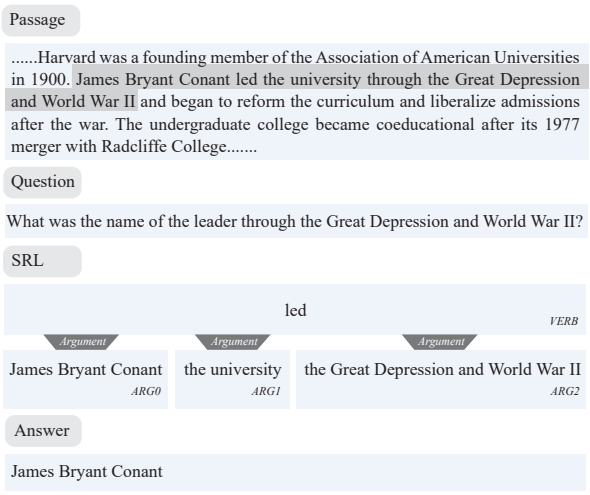


Figure 1: Semantic role labeling guides text comprehension.

sides, explicit semantics has been proved to be beneficial to a wide range of NLP tasks, including discourse relation sense classification (Mihaylov and Frank, 2016), machine translation (Shi et al., 2016) and question answering (Yih et al., 2016). All the previous successful work indicates that explicit contextual semantics may hopefully help into reading comprehension and inference tasks.

Some work studied question answering (QA) driven SRL, like QA-SRL parsing (He et al., 2015; Mccann et al., 2018; Fitzgerald et al., 2018). They focus on detecting argument spans for a predicate and generating questions to annotate the semantic relationship. However, our task is quite different. In QA-SRL, the focus is commonly simple and short factoid questions that are less related to the context, let alone making inference. Actually, text comprehension and inference are quite challenging tasks in NLP, requiring to dig the deep semantics between the document and comprehensive question which are usually raised or re-written by humans, instead of shallow argument alignment around the same predicate in QA-SRL. In this work, to alleviate such an obvious shortcoming about semantics, we make attempt to explore integrative models for finer-grained text comprehension and inference.

In this work, we propose a semantics enhancement framework for TC tasks, which boosts the strong baselines effectively. We implement an easy and feasible scheme to integrate semantic signals in downstream neural models in end-to-end manner to boost strong baselines effectively. An example about how contextual semantics helps MRC is illustrated in Figure 1. A series of detailed case studies are employed to analyze the robustness of the semantic role labeler. To our best knowledge, our work is the first attempt to apply explicit contextual semantics for text comprehension tasks, which have been ignored in previous works for a long time.

The rest of this paper is organized as follows. The next section reviews the related work. Section 3 will demonstrate our semantic learning framework and implementation. Task details and experimental results are reported in Section 4, followed by case studies and analysis in Section 5 and conclusion in Section 6.

2 Related Work

2.1 Text Comprehension

As a challenging task in NLP, text comprehension is one of the key problems in artificial intelligence, which aims to read and comprehend a given text, and then answer questions or make inference based on it. These tasks require a comprehensive understanding of natural languages and the ability to do further inference and reasoning. We focus on two types of text comprehension, document-based question-answering (Table 1) and textual entailment (Table 2). Textual entailment aims for a deep understanding of text and reasoning, which shares the similar genre of machine reading comprehension, though the task formations are slightly different.

In the last decade, the MRC tasks have evolved from the early cloze-style test (Hill et al., 2015; Hermann et al., 2015; Zhang et al., 2018c,b) to span-based answer extraction from passage (Rajpurkar et al., 2016, 2018). The former has restrictions that each answer should be a single word in the document and the original sentence without the answer part is taken as the query. For the span-based one, the query is formed as questions in natural language whose answers are spans of texts. Various attentive models have been employed for text representation and relation discovery, including Attention Sum Reader (Kadlec et al., 2016), Gated attention Reader (Dhingra et al., 2017) and Self-matching Network

Passage	There are three major types of rock: igneous, sedimentary, and metamorphic. The rock cycle is an important concept in geology which illustrates the relationships between these three types of rock, and magma. When a rock crystallizes from melt (magma and/or lava), it is an igneous rock. This rock can be weathered and eroded, and then redeposited and lithified into a sedimentary rock, or be turned into a metamorphic rock due to heat and pressure that change the mineral content of the rock which gives it a characteristic fabric. The sedimentary rock can then be subsequently turned into a metamorphic rock due to heat and pressure and is then weathered, eroded, deposited, and lithified, ultimately becoming a sedimentary rock. Sedimentary rock may also be re-eroded and redeposited, and metamorphic rock may also undergo additional metamorphism. All three types of rocks may be re-melted; when this happens, a new magma is formed, from which an igneous rock may once again crystallize.
Question	What changes the mineral content of a rock?
Answer	heat and pressure.

Table 1: A machine reading comprehension example.

Premise	A man parasails in the choppy water.	Label
Hypo.	The man is competing in a competition.	Neutral
	The man parasailed in the calm water.	Contra.
	The water was choppy as the man parasailed.	Entailment

Table 2: A textual entailment example.

(Wang et al., 2017).

With the release of the large-scale span-based datasets (Rajpurkar et al., 2016; Joshi et al., 2017; Rajpurkar et al., 2018), which constrain answers to all possible text spans within the reference document, researchers are investigating the models with more logical reasoning and content understanding (Wang et al., 2018). Recently, language models also show their remarkable performance in reading comprehension (Devlin et al., 2018; Peters et al., 2018).

For the other type of text comprehension, natural language inference (NLI) is proposed to serve as a benchmark for natural language understanding and inference, which is also known as recognizing textual entailment (RTE). In this task, a model is presented with a pair of sentences and asked to judge the relationship between their meanings, including entailment, neutral and contradiction. Bowman et al. (2015) released Stanford Natural language Inference (SNLI) dataset, which is a high-quality and large-scale benchmark, thus inspiring various significant work.

Most of existing NLI models apply attention mechanism to jointly interpret and align the premise and hypothesis, while transfer learning from external knowledge is popular recently. Notably, Chen et al. (2017) proposed an enhanced sequential inference model (ESIM), which employed recursive architectures in both local inference modeling and inference composition, as well as syntactic parsing information, for a sequential inference model. ESIM is simple with satisfactory performance, and thus is widely chosen as the baseline model. Mccann et al. (2017) proposed to transfer the LSTM encoder from the neural machine translation (NMT) to the NLI task to contextualize word vectors. Pan et al. (2018) transferred the knowledge learned from the discourse marker prediction task to the NLI task to augment the semantic representation.

2.2 Semantic Role Labeling

Given a sentence, the task of semantic role labeling is dedicated to recognizing the semantic relations between the predicates and the arguments. For example, given the sentence, *Charlie sold a book to Sherry last week*, where the target verb (predicate) is *sold*, SRL system yields the following outputs,

[$_{ARG0}$ Charlie] [$_{V}$ sold] [$_{ARG1}$ a book]
[$_{ARG2}$ to Sherry] [$_{AM-TMP}$ last week].

where $ARG0$ represents the seller (agent), $ARG1$ represents the thing sold (theme), $ARG2$ represents the buyer (recipient), $AM-TMP$ is an adjunct indicating the timing of the action and V represents the predicate.

Recently, SRL has aroused much attention from researchers and has been applied in many NLP tasks (Mihaylov and Frank, 2016; Shi et al., 2016; Yih et al., 2016). SRL task is generally formulated as multi-step classification subtasks in pipeline systems, consisting of predicate identification, predicate disambiguation, argument identification and argument classification. Most previous SRL approaches adopt a pipeline framework to handle these subtasks one after another. Notably, Gildea and Jurafsky (2002) devised the first automatic semantic role labeling system based on FrameNet. Traditional systems relied on sophisticated handcraft features or some declarative constraints, which suffer from poor efficiency and generalization ability. A recently ten-

dency for SRL is adopting neural networks methods thanks to their significant success in a wide range of applications. The pioneering work on building an end-to-end neural system was presented by (Zhou and Xu, 2015), applying an 8 layered LSTM model, which takes only original text information as input feature without using any syntactic knowledge, outperforming the previous state-of-the-art system. He et al. (2017) presented a deep highway BiLSTM architecture with constrained decoding, which is simple and effective, enabling us to select it as our basic semantic role labeler. These studies tackle argument identification and argument classification in one shot. Inspired by recent advances, we can easily integrate semantics into text comprehension.

3 Semantic Role Labeling for Text Comprehension

For both downstream text comprehension tasks, we consider an end-to-end model as well as the semantic learning model. The former may be regarded as downstream model of the latter. Thus, our semantics augmented model will be an integration of two end-to-end models through simple embedding concatenation as shown in Figure 2.

In detail, we apply semantic role labeler to annotate the semantic tags (i.e. predicate, argument) for each token in the input sequence so that explicit contextual semantics can be directly introduced, and then the input sequence along with the corresponding semantic role labels is fed to downstream models. We regard the semantic signals as SRL embeddings and employ a lookup table to map each label to vectors, similar to the implementation of word embedding. For each word x, a joint embedding $e^j(w)$ is obtained by the concatenation of word embedding $e^w(x)$ and SRL embedding $e^s(x)$,

$$e^j(w) = e^w(x) \oplus e^s(x)$$

where $\oplus$ is the concatenation operator. The downstream model is task-specific. In this work, we focus on the textual entailment and machine reading comprehension, which will be discussed latter.

3.1 Semantic Role Labeler

Our concerned SRL task includes two subtasks: predicate identification and argument labeling.

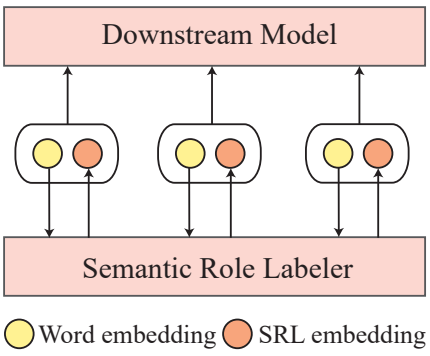


Figure 2: Overview of the semantic learning framework.

While the CoNLL-2005 shared task assumes gold predicates as input, this information is not available in many applications, which requires us to identify the predicates for a input sentence at the very beginning. Thus, our SRL module has to be end-to-end, predicting all predicates and corresponding arguments in one shot.

For predicate identification, we use spaCy[1] to tokenize the input sentence with part-of-speech (POS) tags and the verbs are marked as the binary predicate indicator for whether the word is the verb for the sentence.

Following (He et al., 2017), we model SRL as a span tagging problem[2] and use an 8-layer deep BiLSTM with forward and backward directions interleaved. Different from the baseline model, we replace the GloVe embeddings with ELMo representations[3] due to the recent success of ELMo in NLP tasks (Peters et al., 2018).

In brief, the implementation of our SRL is a series of stacked interleaved LSTMs with highway connections. The inputs are embedded sequences of words concatenated with a binary indicator containing whether a word is the verbal predicate. Additionally, during inference, Viterbi decoding is applied to accommodate valid BIO sequences. The details are

[1] https://spacy.io/

[2] Actually, the easiest way to deal with segmentation or sequence labeling problems is to transform them into raw labeling problems. A standard way to do this is the *BIO* encoding, representing a token at the beginning, interior, or outside of any span, respectively.

[3] The ELMo representation is obtained from `https://allennlp.org/elmo`. We use the original one for this work whose output size is 512.

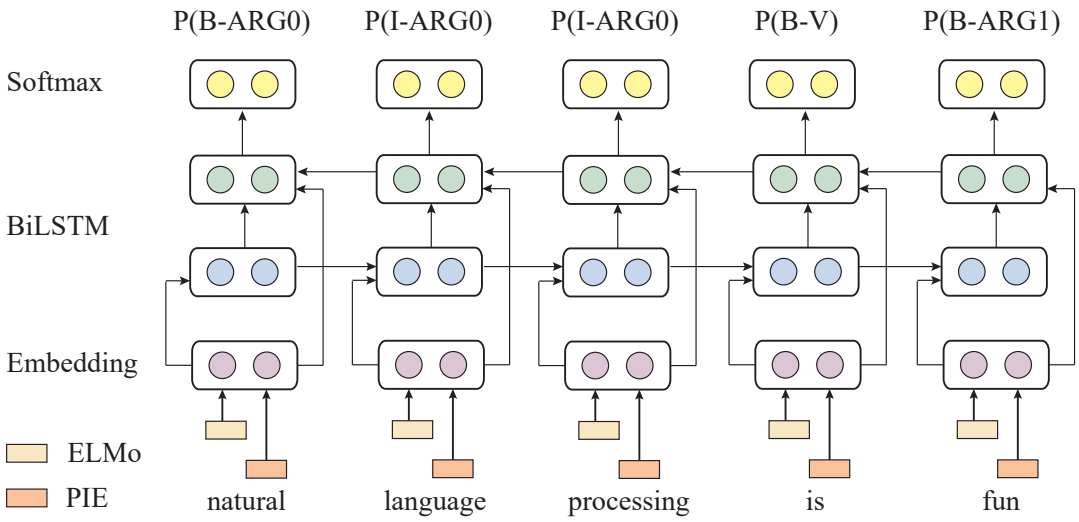


Figure 3: Semantic role labeler.

as follows.

Word Representation The word representation of our SRL model is the concatenation of two vectors: an ELMo embedding $\boldsymbol{e}^{(l)}$ and predicate indicator embedding (PIE) $\boldsymbol{e}^{(p)}$. ELMo is trained from the internal states of a deep bidirectional language model (BiLM), which is pre-trained on a large text corpus with approximately 30 million sentences (Chelba et al., 2014). Besides, following (Li et al., 2019) who shows the predicate-specific feature is helpful in promoting the role labeling, we employ a predicate indicator embedding $\boldsymbol{e}^{(p)}$ to mark whether a word is a predicate when predicting and labeling the arguments. The final word representation is given by $\boldsymbol{e} = \boldsymbol{e}^{(l)} \oplus \boldsymbol{e}^{(p)}$, where $\oplus$ is the concatenation operator. The downstream model will take such a joint embedding as input for specific task.

Encoder As commonly used to model the sequential input, BiLSTM is adopted for our sentence encoder. By incorporating a stack of distinct LSTMs, BiLSTM processes an input sequence in both forward and backward directions. In this way, the BiLSTM encoder provides the ability to incorporate the contextual information for each word.

Given a sequence of word representation $S = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_n\}$ as input, the hidden state $\boldsymbol{h} = \{\boldsymbol{h}_1, \boldsymbol{h}_2, \cdots, \boldsymbol{h}_n\}$ is encoded by BiLSTMs layer where each LSTM uses highway connections between layers and variational recurrent dropout. The encoded representation is then projected using a final dense layer followed by a softmax activation to form a distribution over all possible tags. The predicted semantic role Labels are defined in PropBank (Palmer et al., 2005) augmented with B-I-O tag set to represent argument spans.

Model Implementation The training objective is to maximize the logarithm of the likelihood of the tag sequence, and we expect the correct output sequence matches with,

$$y^* = \underset{\widetilde{y} \in C}{argmax}\, s(x, \widetilde{y}) \tag{1}$$

where $\mathbf{C}$ is candidate label set.

Our semantic role labeler is trained on English *OntoNotes v5.0* dataset (Pradhan et al., 2013) for the CoNLL-2012 shared task, achieving an F1 of 84.6%[4] on the test set. At test time, we perform Viterbi decoding to enforce valid spans using BIO constraints[5]. For the following evaluation, the default dimension of SRL embeddings is 5 and the case study concerning the dimension is shown in the subsection *dimension of SRL Embedding*.

The model is run forward for every verb in the sentence. In some cases there is more than one predicate in a sentence, resulting in various semantic role

[4]This result is comparable with the state-of-the-art (Li et al., 2019).

[5]The BIO format requires argument spans to begin with a B tag.

sets whose number is equal to the number of predicates. For convenient downstream model input, we need to ensure the word and the corresponding label are matched one-by-one, that is, only one set for a sentence. To this end, we select the corresponding BIO sets with the most non-O labels as the semantic role labels. For sentences with no predicate, we directly assign *O* labels to each word in those sentences.

3.2 Text Comprehension Model

Textual Entailment Our basic TE model is the reproduced Enhanced Sequential Inference Model (ESIM) (Chen et al., 2017) which is a widely used baseline model for textual entailment. ESIM employs a BiLSTM to encode the premise and hypothesis, followed by an attention layer, a local inference layer, an inference composition layer. Slightly different from (Chen et al., 2017), we do not include extra syntactic parsing features and directly replace the pre-trained Glove word embedding with ELMo which are completely character based. Our SRL embedding is concatenated with ELMo embeddings and the joint embeddings are then fed to the BiLSTM encoders.

Machine Reading Comprehension Our baseline MRC model is an enhanced version of Bidirectional Attention Flow (Seo et al., 2017) following (Clark and Gardner, 2018). The token embedding is the concatenation of pre-trained GloVe word vectors, a character-level embedding from a convolutional neural network with max-pooling and pre-trained ELMo embeddings (Peters et al., 2018). Our semantics enhanced model takes input of concatenating the token embedding with SRL embeddings. The embeddings of document and question are passed through a shared bi-directional GRU, followed by a BiDAF attention (Seo et al., 2017). The contextual document and question representations are then passed to a residual self-attention layer. The above model is denoted as *ELMo*. Table 5 shows the results on SQuAD MRC task[6]. The SRL embeddings give substantial performance gains over all the strong baselines, showing it is also quite effective for more complex document and question encoding.

Model	Accuracy (%)
Deep Gated Attn. BiLSTM	85.5
Gumbel TreeLSTM	86.0
Residual stacked	86.0
Distance-based SAN	86.3
BCN + CoVe + Char	88.1
DIIN	88.0
DR-BiLSTM	88.5
CAFE	88.5
MAN	88.3
KIM	88.6
DMAN	88.8
ESIM + TreeLSTM	88.6
ESIM + ELMo	88.7
DCRCN	88.9
LM-Transformer	89.9
MT-DNN†	91.1
Baseline (ELMo)	88.4
+ SRL	89.1
Baseline ($\text{BERT}_{\text{BASE}}$)	89.2
+ SRL	89.6
Baseline ($\text{BERT}_{\text{LARGE}}$)	90.4
+ SRL	**91.3**

Table 3: Accuracy on SNLI test set. Models in the first block are sentence encoding-based. The second block embodies the joint methods while the last block shows our SRL based model. All the results except ours are from the SNLI Leaderboard. Previous state-of-the-art model is marked by †. Since ensemble systems are commonly integrated with multiple heterogeneous models and resources, we only show the results of single models to save space though our single model also outperforms the ensemble models.

4 Evaluation

In this section, we evaluate the performance of SRL embeddings on two kinds of text comprehension tasks, *textual entailment* and *reading comprehension*. Both of the concerned tasks are quite challenging, and could be even more difficult considering that the latest performance improvement has been already very marginal. However, we present the semantics enhanced solution instead of heuristically stacking network design techniques to give further advances. In our experiments, we basically

[6]For BERT evaluation, we only use SQuAD training set instead of joint training with other datasets to keep the model simplicity. Since the test set of SQuAD is not publicly available, our evaluations are based on dev set.

Model	Dev	Test
Our model	**89.11**	**89.09**
-ELMo	88.51	88.42
-SRL	88.89	88.65
-ELMo -SRL	88.39	87.96

Table 4: Ablation study. Since we use ELMo as the basic word embeddings, we replace ELMO with 300D GloVe embeddings for the case *-ELMo*.

follow the same hyper-parameters for each model as the original settings from their corresponding literatures (Peters et al., 2018; Chen et al., 2017; Clark and Gardner, 2018) except those specified (e.g. SRL embedding dimension). For both of the tasks, we also report the results by using pre-trained BERT (Devlin et al., 2018) as word representation in our baseline models [7]. The hyperparameters were selected using the Dev set, and the reported Dev and Test scores are averaged over 5 random seeds using those hyper-parameters.

4.1 Textual Entailment

Textual entailment is the task of determining whether a *hypothesis* is *entailment, contradiction* and *neutral*, given a *premise*. The Stanford Natural Language Inference (SNLI) corpus (Bowman et al., 2015) provides approximately $570k$ hypothesis/premise pairs. We evaluate the model performance in terms of accuracy.

Results in Table 3 show that SRL embedding can boost the ESIM+ELMo model by +0.7% improvement. With the semantic cues, the simple sequential encoding model yields substantial gains, and our single $\text{BERT}_{\text{LARGE}}$ model also achieves a new state-of-the-art, even outperforms all the ensemble models in the leaderboard[8]. This would be owing to more accurate and fine-grained information from effective explicit semantic cues.

To evaluate the contributions of key factors in our method, a series of ablation studies are performed on the SNLI dev and test set. The results are in Table 4. We observe both SRL and ELMo embeddings contribute to the overall performance. Note that ELMo is obtained by deep bidirectional language with 4,096 hidden units on a large-scale corpus, which requires long training time with 93.6 million parameters. The output dimension of ELMo is 512. Compared with the massive computation and high dimension, SRL embedding is much more convenient for training and much easier for model integration, giving the same level of performance gains.

4.2 Machine Reading Comprehension

To investigate the effectiveness of the SRL embedding in conjunction with more complex models, we conduct experiments on machine reading comprehension tasks. The reading comprehension task can be described as a triple $< D, Q, A >$, where D is a document (context), Q is a query over the contents of D, in which a span is the right answer A.

As a widely used benchmark dataset for machine reading comprehension, the Stanford Question Answering Dataset (SQuAD) (Rajpurkar et al., 2016) contains $100k$+ crowd sourced question-answer pairs where the answer is a span in a given Wikipedia paragraph. Two official metrics are selected to evaluate the model performance: Exact Match (EM) and a softer metric F1 score, which measures the weighted average of the precision and recall rate at a character level. Our baseline includes MQAN (Mccann et al., 2018) for single task and multi-task with SRL, BiDAF+ELMo (Peters et al., 2018), R.M. Reader and BERT (Devlin et al., 2018).

Table 5 shows the results[9]. The SRL embeddings give substantial performance gains over all the strong baselines, showing it is also quite effective for more complex document and question encoding.

5 Case Studies

From the above experiments, we see our semantic learning framework works effectively and the semantic role labeler boosts model performance, verifying our hypothesis that semantic roles are critical for text understanding. Though the semantic role labeler is trained on a standard benchmark dataset,

[7] We use the last layer of BERT output. Since BERT is in subword-level while semantics role labels are in word-level, to use BERT in conjunction with our SRL embeddings, we need to keep them aligned. Therefore, we use the BERT embedding for the first subword of each word, which is slightly different from the original BERT.

[8] Since March 24th, 2019. The leaderboard is here: https://nlp.stanford.edu/projects/snli/.

[9] Since the test set of SQuAD is not publicly available, our evaluations are based on dev set.

Model	EM	F1	RERR
Published			
$\text{MQAN}_{\text{single-task}}$	-	75.5	-
$\text{MQAN}_{\text{multi-task}}$	-	74.3	-
BiDAF+ELMo	-	85.6	-
R.M. Reader	78.9	86.3	-
$\text{BERT}_{\text{BASE}}$	80.8	88.5	-
$\text{BERT}_{\text{LARGE}}$†	84.1	90.9	-
Our implementation			
Baseline (ELMo)	77.5	85.2	-
+SRL	**78.5**	**86.0**	**5.4%**
Baseline ($\text{BERT}_{\text{BASE}}$)	81.3	88.5	-
+SRL	**81.7**	**88.8**	**2.6%**
Baseline ($\text{BERT}_{\text{LARGE}}$)	84.2	90.9	-
+SRL	**84.5**	**91.2**	**3.3%**

Table 5: Exact Match (EM) and F1 scores on SQuAD dev set. RERR is short for relative error rate reduction of our model to the baseline evaluated on F1 score. Previous state-of-the-art model is marked by †.

Ontonotes, whose source ranges from news, conversational telephone speech, weblogs, etc., it turns out to be generally useful for text comprehension from probably quite different domains in both textual entailment and machine reading comprehension. To further evaluate the proposed method, we conduct several case studies as follows.

5.1 Dimension of SRL Embedding

The dimension of embedding is a critical hyper-parameter in deep learning models that may influence the performance. Too high dimension would cause severe over-fitting issues while too low dimension would also cause under-fitting results. To investigate the influence of the dimension of SRL embeddings, we change the dimension in the intervals [1, 2, 5, 10, 20, 50, 100]. Figure 4 shows the results. We see that 5-dimension SRL embedding gives the best performance on both SNLI and SQuAD datasets.

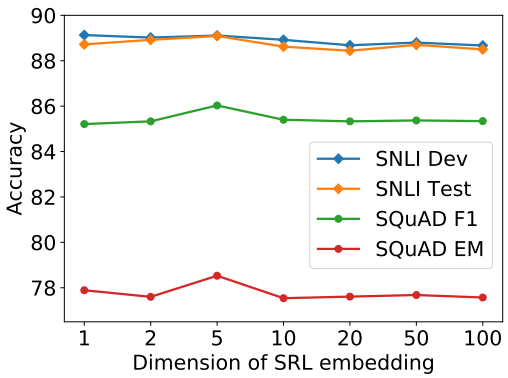


Figure 4: Results on SNLI and SQuAD with different SRL embedding dimensions.

Model	Dev	Test
Baseline	88.89	88.65
Word + SRL	**89.11**	**89.09**
Word + POS	88.90	88.68
Word + NE	89.14	88.51

Table 6: Comparison with different NLP tags.

5.2 Comparison with POS/NER Tags

The study of computational linguistics is a critical part in NLP (Zhou and Zhao, 2019; Li et al., 2018b). In particular, Part-of-speech (POS) and named entity (NE) tags have been broadly used in various tasks. To make comparisons, we conduct experiments on SNLI with modifications on label embeddings using tags of SRL, POS and NE, respectively. Results in Table 6 show that SRL gives the best result, showing semantic roles contribute to the performance, which also indicates that semantic information matches the purpose of NLI task best.

6 Conclusion

This paper presents a novel semantic learning framework for fine-grained text comprehension and inference. We show that our proposed method is simple yet powerful, which achieves a significant improvement over strong baseline models, including those which have been enhanced by the latest BERT. This work discloses the effectiveness of explicit semantics in text comprehension and inference and proposes an easy and feasible scheme to integrate explicit contextual semantics in neural models. A series of detailed case studies are employed to analyze the adopted robustness of the semantic role labeler. Different from most recent works focusing on heuristically stacking complex mechanisms for performance improvement, this work is to shed some lights on fusing accurate semantic signals for deeper comprehension and inference.

References

Hongxiao Bai and Hai Zhao. Deep enhanced representation for implicit discourse relation recognition. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 571–583, 2018.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proceedings of the 20th conference on Empirical Methods in Natural Language Processing (EMNLP 2015)*, pages 632–642, 2015.

Deng Cai, Hai Zhao, Zhisong Zhang, Yuan Xin, Yongjian Wu, and Feiyue Huang. Fast and accurate neural word segmentation for Chinese. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, pages 608–615, 2017.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Ge Qi, Thorsten Brants, Phillipp Koehn, and Tony Robinson. One billion word benchmark for measuring progress in statistical language modeling. *arXiv:1312.3005*, 2014.

Qian Chen, Xiaodan Zhu, Zhenhua Ling, Si Wei, Hui Jiang, and Diana Inkpen. Enhanced lstm for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, pages 1657–1668, 2017.

Shenyuan Chen, Hai Zhao, and Rui Wang. Neural network language model for chinese pinyin input method engine. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 455–461, 2015.

Christopher Clark and Matt Gardner. Simple and effective multi-paragraph reading comprehension. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 845–855, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Bhuwan Dhingra, Hanxiao Liu, Zhilin Yang, William W. Cohen, and Ruslan Salakhutdinov. Gated-attention readers for text comprehension. In *Proceedings of the 55th annual meeting of the Association for Computational Linguistics (ACL 2017)*, pages 1832–1846, 2017.

Nicholas Fitzgerald, Luheng He, and Luke Zettlemoyer. Large-scale QA-SRL parsing. *ACL*, 2018.

Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 512–520, 2002.

Luheng He, Mike Lewis, and Luke Zettlemoyer. Question-answer driven semantic role labeling: Using natural language to annotate natural language. In *Proceedings of the 2015 conference on empirical methods in natural language processing (EMNLP 2015)*, pages 643–653, 2015.

Luheng He, Kenton Lee, Mike Lewis, and Zettlemoyer. Deep semantic role labeling: What works and what's next. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, pages 473–483, 2017.

Shexia He, Zuchao Li, Hai Zhao, Hongxiao Bai, and Gongshen Liu. Syntax for semantic role labeling, to be, or not to be. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, 2018.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. In *Advances in Neural Information Processing Systems 28 (NIPS 2015)*, pages 1693–1701, 2015.

Felix Hill, Antoine Bordes, Sumit Chopra, and Jason Weston. The goldilocks principle: Reading children's books with explicit memory representations. *arXiv preprint arXiv:1511.02301*, 2015.

Yafang Huang, Zuchao Li, Zhuosheng Zhang, and Hai Zhao. Moon IME: neural-based chinese pinyin aided input method with customizable association. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018), System Demonstration*, pages 140–145, 2018.

Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems.

In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP 2017)*, pages 2021–2031, 2017.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, pages 1601–1611, 2017.

Rudolf Kadlec, Martin Schmid, Ondrej Bajgar, and Jan Kleindienst. Text understanding with the attention sum reader network. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 908–918, 2016.

Zuchao Li, Shexia He, Jiaxun Cai, Zhuosheng Zhang, Hai Zhao, Gongshen Liu, Linlin Li, and Luo Si. A unified syntax-aware framework for semantic role labeling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pages 2401–2411, 2018a.

Zuchao Li, Shexia He, Zhuosheng Zhang, and Hai Zhao. Joint learning of POS and dependencies for multilingual universal dependency parsing. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 65–73, 2018b.

Zuchao Li, Shexia He, Hai Zhao, Yiqing Zhang, Zhuosheng Zhang, Xi Zhou, and Xiang Zhou. Dependency or span, end-to-end uniform semantic role labeling. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, 2019.

Bryan Mccann, James Bradbury, Caiming Xiong, and Richard Socher. Learned in translation: Contextualized word vectors. In *Advances in Neural Information Processing Systems (NIPS 2017)*, pages 6294–6305, 2017.

Bryan Mccann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. The natural language decathlon: Multitask learning as question answering. *ArXiv:1806.08730*, 2018.

Todor Mihaylov and Anette Frank. Discourse relation sense classification using cross-argument semantic similarity based on word embeddings. In *Conll-16 Shared Task*, pages 100–107, 2016.

Pramod Kaushik Mudrakarta, Ankur Taly, Mukund Sundararajan, and Kedar Dhamdhere. Did the model understand the question? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1896–1906, 2018.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 2005.

Boyuan Pan, Yazheng Yang, Zhou Zhao, Yueting Zhuang, Deng Cai, and Xiaofei He. Discourse marker augmented network with reinforcement learning for natural language inference. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, pages 989–999, 2018.

M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*, 2018.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. Towards robust linguistic analysis using OntoNotes. *CoNLL*, 2013.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. Squad: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP 2016)*, pages 2383–2392, 2016.

Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for squad. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, 2018.

Minjoon Seo, Aniruddha Kembhavi, Ali Farhadi, and Hannaneh Hajishirzi. Bidirectional attention flow for machine comprehension. In *ICLR 2017 : 5th International Conference on Learning Representations*, 2017.

Chen Shi, Shujie Liu, Shuo Ren, Shi Feng, Mu Li, Ming Zhou, Xu Sun, and Houfeng Wang.

Knowledge-based semantic embedding for machine translation. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 2245–2254, 2016.

Wei Wang, Ming Yan, and Chen Wu. Multi-granularity hierarchical attention fusion networks for reading comprehension and question answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL 2018)*, volume 1, pages 1705–1714, 2018.

Wenhui Wang, Nan Yang, Furu Wei, Baobao Chang, and Ming Zhou. Gated self-matching networks for reading comprehension and question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL 2017)*, pages 189–198, 2017.

Fengshun Xiao, Jiangtong Li, Hai Zhao, Rui Wang, and Kehai Chen. Lattice-based transformer encoder for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3090–3097, 2019.

Wen Tau Yih, Matthew Richardson, Chris Meek, Ming Wei Chang, and Jina Suh. The value of semantic parse labeling for knowledge base question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, pages 201–206, 2016.

Zhisong Zhang, Rui Wang, Masao Utiyama, Eiichiro Sumita, and Hai Zhao. Exploring recombination for efficient decoding of neural machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*, pages 4785–4790, 2018a.

Zhuosheng Zhang and Hai Zhao. One-shot learning for question-answering in gaokao history challenge. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 449–461, 2018.

Zhuosheng Zhang, Yafang Huang, and Hai Zhao. Subword-augmented embedding for cloze reading comprehension. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 1802–1814, 2018b.

Zhuosheng Zhang, Yafang Huang, Pengfei Zhu, and Hai Zhao. Effective character-augmented word embedding for machine reading comprehension. In *Proceedings of the Seventh CCF International Conference on Natural Language Processing and Chinese Computing (NLPCC 2018)*, pages 27–39, 2018c.

Zhuosheng Zhang, Jiangtong Li, Pengfei Zhu, and Hai Zhao. Modeling multi-turn conversation with deep utterance aggregation. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, pages 3740–3752, 2018d.

Zhuosheng Zhang, Yafang Huang, and Hai Zhao. Open vocabulary learning for neural chinese pinyin ime. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1584–1594, 2019a.

Zhuosheng Zhang, Hai Zhao, Kangwei Ling, Jiangtong Li, Shexia He, and Guohong Fu. Effective subword segmentation for text comprehension. *IEEE/ACM Transactions on Audio, Speech, and Language Processing (TASLP)*, 27(11):1664–1674, 2019b.

Jie Zhou and Wei Xu. End-to-end learning of semantic role labeling using recurrent neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL 2015)*, pages 1127–1137, July 2015.

Junru Zhou and Hai Zhao. Head-driven phrase structure grammar parsing on penn treebank. In *Proceedings of the 57th Conference of the Association for Computational Linguistics (ACL 2019)*, pages 2396–2408, 2019.

Pengfei Zhu, Zhuosheng Zhang, Jiangtong Li, Yafang Huang, and Hai Zhao. Lingke: A fine-grained multi-turn chatbot for customer service. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018), System Demonstrations*, pages 108–112, 2018.

Chinese–Japanese Unsupervised Neural Machine Translation Using Sub-character Level Information

Longtu Zhang
Computational Linguistics Lab,
Graduate School of System Design,
Tokyo Metropolitan University
zhang-longtu@ed.tmu.ac.jp

Mamoru Komachi
Computational Linguistics Lab,
Graduate School of System Design,
Tokyo Metropolitan University
komachi@tmu.ac.jp

Abstract

Unsupervised neural machine translation (UNMT) requires only monolingual data of similar language pairs during training and can produce bidirectional translation models with relatively good performance on alphabetic languages (Lample et al., 2018). However, little research has been done on logographic language pairs. This study focuses on Chinese–Japanese UNMT trained by data containing sub-character (ideograph or stroke) level information, which is obtained by decomposing character-level data. BLEU (Papineni et al., 2002) scores of both character-level and sub-character-level systems were compared against each other. The results showed that, despite the effectiveness of UNMT on character-level data, sub-character-level data could further enhance the performance. Moreover, the stroke-level system outperformed the ideograph-level system.

1 Introduction

Although supervised neural machine translation (NMT) has achieved great success in recent years (Wu et al., 2016; Vaswani et al., 2017), the fact that it may fail without large quantities of parallel training data is a practical problem (Koehn and Knowles, 2017; Isabelle et al., 2017), particularly for low-resource domains and language pairs. Lample et al. (2018) proposed an unsupervised neural machine translation (UNMT) method that requires only monolingual training data to train bidirectional translation models on similar language pairs; it relies heavily on the shared information between source and target data. They experimented on alphabetic language pairs (English–French and English–German) and showed the effectiveness of such methods: although the BLEU score is not as high as state-of-the-art supervised models, the translation quality is highly acceptable.

Language	Word
JA-character	風 景
JA-ideograph	⿰几虫 ⿰日京
JA-stroke	⿰⿴丿乙⿰日丿⿱⿰日⿴丨フ 一⿴⿰日丨 一、 ⿰日⧧⿰⿴丨フ日一一 ...
ZH-character	风 景
ZH-ideograph	⿰几乂 ⿰日京
ZH-stroke	⿰⿴丿乙⿱丿、 ⿰日⧧⿰⿴丨フ日一一 ...
EN	landscape

Table 1: Examples of decomposition of a Japanese word "風景" and Chinese word "风景," both meaning "landscape" in English.

Chinese and Japanese are also similar language pairs, using Chinese characters in their logographic writing systems; there are no natural word boundaries and the characters are formed compositionally by sub-character level units, such as ideographs and strokes. Table 1 shows examples of how words in Chinese and Japanese are decomposed. Compared with words, the ideograph and stroke sequences have a higher proportion of shared parts ; shared parts are very useful for byte pair encoding (BPE) algorithms and shared vocabularies in ma-

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 309-315, Hakodate, Japan, September 13-15, 2019

chine translation systems. Given this significant difference, it is worth asking whether natural language processing (NLP) methods that are successful for alphabetic languages will also work for logographic languages.

The idea of integrating sub-character-level information into NLP tasks is not entirely new. For example, such information helps in training better word embeddings (Shi et al., 2015; Peng et al., 2017) and text classification systems (Toyama et al., 2017). Recently, Zhang et al. (2018) have demonstrated that sub-character level information will help Chinese–Japanese supervised NMT systems on both the encoder and decoder sides. However, there is still no study on logographic UNMT systems.

Therefore, this study attempted to answer the following questions:

1. Is UNMT effective for logographic language pairs, such as Chinese–Japanese, particularly when sub-character-level information is used?

2. What is the influence of the shared token rate on UNMT?

2 Background

2.1 Chinese Characters

Chinese and Japanese use structured strokes to form ideographs and then form characters. (Japanese also has kanas, which function as phonetic letters.) According to the UNICODE 10.0 standard, there are 36 strokes (such as "一," "丨," "丿," and "乀,") which compose hundreds of ideographs[1], and more than 90,000 different characters. Table 2 shows examples of how strokes and ideographs compose different characters.

2.2 The Structure of Transformer Units

The UNMT architecture, introduced in Section 2.3, is built based on transformer units in which there are three basic structures (Vaswani et al., 2017): *positional embedding* (PE), *multihead attention* (MA), and *position-wise feedforward network* (FFN).

[1]The number depends on the definition of ideographs (usually around 500 or more).

Character	Semantic ideograph	Phonetic ideograph	Pinyin
驰 run	马 horse	也	chí
池 pool	水(氵) water	也	chí
施 impose	方 direction	也	shī
弛 loosen	弓 bow	也	chí
地 land	土 soil	也	dì
驱 drive	马 horse	区	qū

Table 2: Examples of Chinese characters. (Pinyin is the official romanization representing a character's pronunciation.) Both semantic and phonetic ideographs can be shared across different characters for similar functions. For example, "驰" and "驱," both containing "马," have related meanings, while characters containing "也" are usually pronounced similarly.

Positional embedding. The positional embedding matrix is computed by two trigonometric functions, given the token position pos and the hidden index i, as shown in Equation 1. It is then applied to normal pretrained embeddings by simple addition:

$$\begin{aligned} PE_{(pos,2i)} &= \sin(pos/10000^{2i/d_{model}}) \\ PE_{(pos,2i+1)} &= \cos(pos/10000^{2i/d_{model}}) \end{aligned} \quad (1)$$

Functioning as an improved version of the traditional attention mechanism (Equation 2), multihead attention computes scaled attention scores on split $query$, key, and $value$ pairs according to Equation 3, and then concatenates the results. In Equation 3, QW_i^Q, KW_i^K, and VW_i^V are Q_i, K_i, and V_i, respectively, projected by FFNs.

Multihead attention. The MA that takes identical hidden states as Q, K, and V is the so-called "self attention." The MA that takes target states as Q and source states as K and V is the so-called "context attention."

$$Attention(Q, K, V) = softmax(QK^T/\sqrt{d_k})V \quad (2)$$

$$\begin{aligned} &MultiHead(Q, K, V) = Concat(h_1, ..., h_i)W^o \\ &h_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \end{aligned} \quad (3)$$

Position-wise FFN. The position-wise FFN is a combination of two FFNs with a ReLU activation function in between, as shown in Equation 4.

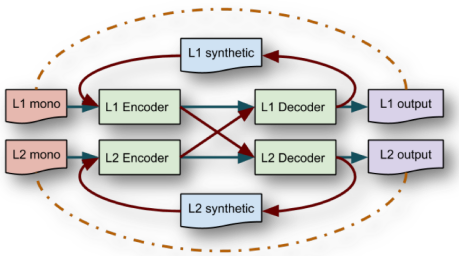


Figure 1: The architecture of the unsupervised NMT model. The green arrows indicate the direction of data flow in encoder–decoder language models, while the red arrows indicate the direction of data flow in back-translation models. The dotted lines are losses computed on the same language; therefore, no supervision is needed.

$$FFN(x) = max(0, xW_1 + b_1)W_2 + b_2 \quad (4)$$

Each encoder layer contains one "self MA" and one FFN; each decoder layer contains one "self MA," one "context MA," and one FFN. Encoders will first embed the source sequence using source PE and feed the output to stacked encoder layers to obtain the encoder hidden state. The decoders will take the encoder state and embed the target sequence using target PE, and then feed both of them to stacked decoder layers to obtain the decoder state. Like normal NMT systems, a linear layer and a softmax layer are used to project the decoder state to vocabulary scores.

2.3 The UNMT Architecture

The UNMT architecture uses two transformer encoders and two transformer decoders to form two "encoder–decoder language models" (LM) and two "back-translation models" (BT) in a crossed fashion, as shown in Figure 1:

- *L1 LM*: L1 mono ⇒ L1 encoder ⇒L1 decoder ⇒L1 output
- *L2 LM*: L2 mono ⇒L2 encoder ⇒L2 decoder ⇒L2 output
- *L1 BT*: L1 mono ⇒L1 encoder ⇒L2 decoder ⇒L2 synthetic ⇒L2 encoder ⇒L1 decoder ⇒L1 output
- *L2 BT*: L2 mono ⇒L2 encoder ⇒L1 decoder ⇒L1 synthetic ⇒L1 encoder ⇒L2 decoder ⇒L2 output

In this architecture, all four losses are computed within the same language so that no supervision is needed.

There are three key structures that underpin the approach to UNMT systems:

Shared BPE Embeddings. Instead of mapping two monolingual embeddings together (Artetxe et al., 2018), the shared BPE embeddings are trained directly on the concatenated source and target monolingual data. This was found efficient and effective for UNMT (Lample et al., 2018).

Encoder–Decoder Language Models. The weights of the deeper layers of the encoders are often shared, to enhance performance. Alternatively, an multi-layer perceptron (MLP) discriminator can be added, to discriminate between the latent representations produced by different encoders.[2]

Back-Translation Models. UNMT borrowed this idea from Sennrich et al. (2016): the back-translation models are trained jointly in both translation directions. Specifically, for one direction, the forward NMT model first generates synthetic target data, and then it is translated back to the source language using the backward model.

3 Chinese–Japanese Sub-character Level UNMT

In addition to validating the effectiveness of UNMT with the Chinese–Japanese language pair, this study has further enhanced the shared information by decomposing characters into ideographs and strokes[3].

[2]It is claimed to be better to have a discriminator that takes the output of the two encoders and to adversarially train it with the translation model (Lample et al., 2018). However, in our experiment, we find this to be effective only for distant language pairs; it makes little difference to the result with similar language pairs, such as Chinese–Japanese, as in our setting. Therefore, we disregard the discriminator here.

[3]In the character-level corpus that we use, the average word length of Chinese and Japanese from dictionary-based tokenizers are 1.7 and 2.2, respectively, which is too short for a BPE algorithm to obtain better shared information. Longer decomposed sequences would be preferable.

3.1 Character Decomposition

Both Chinese and Japanese data are encoded using UNICODE in which similar CJK (Chinese-Japanese-Korean) characters are merged into one type. The CHISE project[4] provides decomposed mapping information from CJK characters to predefined ideograph sequences. There are 394 ideographs and 19 special symbols for "unclear" ideographs. In addition, there are 11 "ideographic description characters" (IDCs) to describe the structural relationship between ideographs, which can help to reduce the ambiguity of the decomposed data.

Based on the CHISE project , we developed a decomposition tool called "textprep" to decompose character-level tokenized data to sub-character-level ideograph and stroke data with no ambiguity[5]. This means that both Chinese and Japanese data can be decomposed to ideograph and stroke sequences and composed back to character sequences. To enable this, a special duplication marker ("") is added in minor ambiguous cases. In addition, all of the ideographs were manually transcribed to stroke sequences. A corpus with no structural information was also created, for comparison reasons, by removing IDCs and adding necessary duplication markers. Table 1 contains examples of various levels of character decomposition in the training corpus.

3.2 Controlling Shared Tokens

Lample et al. (2018) have successfully made 95% of the BPE tokens in the English–German language pair shared across the training set, indicating that the greater the proportion of token sharing, the better a UNMT system will perform. Our study sampled from the same dataset with a controlled rate of token sharing, to gain a better understanding of this notion. Algorithm 1 takes the token sharing rate r, top-k value k, and sample size N as parameters.

Algorithm 1: Sharing Rate Sampling

Data: source/target sentences
Input: r, k, N
Output: source/target sentences with r sharing rate ($sample$)
Init:
$current\_r, vocab, shared\_vocab, sample$;
while $len(sample) < N$ **do**
  $current\_sample \sim$ randomly sample $8 \times k$ sentences;
  calculate sentence-level sharing rate s_r based on $shared\_vocab$;
  sort $sample$ in descending order of s_r;
  if $current\_r < r$ **then**
    select top k sentences;
  else
    select bottom k sentences;
  end
  add selected sentences to $sample$;
  update $current\_r, vocab, shared\_vocab$;
  remove $current\_sample$ from datasets;
end

4 Experiments

To answer the research questions, two lines of experiments were performed. The Japanese–Chinese portion of the Asian Scientific Paper Excerpt Corpus (ASPEC-JC (Nakazawa et al., 2016)) was used. Although this is a parallel corpus, we shuffled it and used it monolingually. The official training/development/testing split contains 670,000 Chinese and Japanese sentences for training and more than 2,000 sentences for evaluating and testing. Word level BLEU scores are used as the evaluation metric.

Sub-character-level UNMT. The baseline is a UNMT system trained on Chinese–Japanese monolingual data, which are first pre-tokenized into words, and then BPE'ed using fastBPE[6]. We call this the character-level baseline because no sub-character-level units are involved. The experiments are to compare it against UNMT systems trained on sub-character-level data, which are directly decomposed from character-level data and then BPE'ed using fastBPE. In sub-character-level data, the presence of structural information was also controlled by adding or removing IDCs.

[4] http://www.chise.org/
[5] https://github.com/vincentzlt/textprep
[6] https://github.com/glample/fastBPE

Granularity		JA–ZH	ZH–JA
Character		24.18 *(29.60)*	29.79 *(40.00)*
Ideograph	w/ IDCs	25.76*	32.61*
	w/o IDCs	25.14* *(32.00)*	32.17* *(42.60)*
Stroke	w/ IDCs	**26.39***	**32.99***
	w/o IDCs	24.75* *(32.10)*	30.59* *(42.20)*

Table 3: BLEU scores ($*$ for statistically significant score against baseline at $p < 0.0001$) of UNMT (larger fonts) and supervised NMT systems (Zhang and Komachi, 2018) (smaller fonts in parentheses) on test sets.

UNMT with different token sharing. We sampled data ($N = 300,000$) from the same monolingual corpus using Algorithm 1 with a controlled token sharing rate (r) of 0.5, 0.7, and 0.9. This is because UNMT systems trained on stroke-level data with IDCs achieved the best performance in preliminary experiments.

For pre-tokenization of the data, Jieba[7] was applied to Chinese using the default dictionary and MeCab[8] was applied to Japanese using the IPA dictionary. For BPE training, the vocabulary size was set to 30,000. We used 4-layer standard transformer (Vaswani et al., 2017) units as our two encoders and decoders. The embedding size was 512; the hidden size of the fully connected network was 2048; the weights of the last three layers of the encoders were shared; the number of multi-attention heads was 8. During training, the dropout rate was set to 0.1 and both vocabularies and embeddings were shared. 10% of input and output sentences were randomly blanked out to add noise to the language model training. We used the Adam optimizer with a learning rate of 0.0001.

5 Results

5.1 Sub-character Level UNMT

Table 3 shows the results for sub-character-level UNMT in both translation directions. Comparing with the character-level baseline, all sub-character-level models have better BLEU scores. In both stroke and ideograph systems, IDCs in the data can further enhance the performance. However, for ideograph systems, removing structural information did not decrease the performance much, whereas a significant drop was observed in stroke systems without structural information. The best UNMT system was trained on stroke data with structural information, in both translation directions. This contrasts with the finding of Zhang and Komachi (2018) on supervised NMT systems: that when both source and target data had the same granularity, ideograph systems outperformed stroke systems in both translation directions.

[7]https://github.com/fxsjy/jieba

[8]http://taku910.github.io/mecab/

r	JA–ZH	ZH–JA
0.5	19.72	25.23
0.7	23.60	28.32
0.9	23.04	28.84

Table 4: BLEU scores with different token sharing rates on test set.

5.2 UNMT with Different Share Token Rates

Table 4 shows the results for UNMT systems using data with different share token rates. When $r = 0.5$, the system recorded the lowest performance; however, when r increased to 0.7 and 0.9, the performance differences became negligible . In contrast with Lample et al. (2018), in our previous sub-character experiments, only 66% to 68% of the tokens were shared but we could still achieve relatively good BLEU scores .

6 Discussion

This study has confirmed the effectiveness of UNMT systems on small Chinese–Japanese datasets, with a much lower token sharing rate than Lampel et al. (2018). Although the BLEU score is not as high as most RNN-based and transformer-based supervised NMT systems, it is still promising, not only because of its translation quality, but also because it greatly broadens the scope of machine translation applications.

6.1 Translation Quality

In both translation directions, there were many synonymous expressions produced that lowered the BLEU score. However, according to native speakers' judgement, they tended to be good translations

Type	Sentence
Reference–JA	図 3 に 「 会 」 が 固有 表現 で ある か 否 か を 判定 する 2 つ の 例文 を 示し た .
Reference–ZH	图 3 所示 的 是 2 个 关于 判断 “ 会 ” 是否是 固有 表达 的 例句 。
Character–JA	図 3 に 示す よう な 2 つ の 判断 について 「 会 」 が 固有 表現 で ある か どう か を 判断 する 例文 を 示す .
Character–ZH	图 3 中 显示 了 判定 “ 会 ” 是 固有 名词 还是 有 2 个 例句 。
Ideograph–JA	図 3 に 示す よう に 2 つ の 判断 「 会 」 が 固有 表現 で ある か どう か について の 例文 を 示す .
Ideograph–ZH	图 3 中 显示 了 判定 “ 会 ” 是否是 固有 名词 的 2 个 例句 。
Stroke–JA	図 3 に 示す の は , 2 つ の 判断 について 「 会 」 が 固有 表現 の 例文 で ある か どう か で ある
Stroke–ZH	图 3 中 显示 了 判定 “ 会 ” 是否是 固有 表达 的 2 个 例句 。
English	Figure 3 showed 2 example sentences of judging whether “会” is an inherent expression.

Table 5: Translation examples from three UNMT models in six translation directions.

in respect of grammaticality, fluency, and naturalness. For example, in Table 5, the character-level system's Chinese translation “中 显示” (“in which shows”) was very close to the reference “所示” (“as shown in”) semantically, and it was consistent in the ideograph-level and stroke-level models. A similar example is “判断” (“judge”) in reference and “判定” (“determine”) in hypothesis. This might be because of the encoder–decoder language models, which successfully grasp the language features and express them in the translation. Consequently, if semantic metrics could be introduced, the performance of UNMT might be better reflected in the results.

6.2 Shared Information and Proportion of Shared Tokens

Zhang et al. (2018) showed that shared information in the form of sub-character-level information can help supervised NMT systems; this study found a similar phenomenon, although with a different granularity preference. This is largely a result of better shared information. For example, in Table 5, despite the fact that translations produced by ideograph and stroke models were better than those of the character model, the stroke model was slightly better than the ideograph model because it translated the Japanese “表現” (“expression”) into Chinese “表达” (“expression”), which was more precise than the ideograph model's “名词” (“none”). However, current unsupervised models still perform poorly on distant language pairs. If the shared information between distant language pairs can be improved, UNMT may work for more general purposes. Additionally, although a low proportion of shared tokens can harm the performance, a high proportion does not linearly improve the performance.

7 Conclusion

The effectiveness of UNMT models on the logographic language pair, Chinese–Japanese, is quite promising, even when using a small training dataset. However, to evaluate its performance more accurately, better semantic metrics are required. Finally, a relatively high proportion of shared tokens is required for good UNMT (around 70%), but a higher shared token rate seems unnecessary.

Acknowledgments

This work was partially supported by JSPS Grant-in-Aid for Young Scientists (B) Grant Number JP16K16117.

References

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2018. A robust self-learning method for fully unsupervised cross-lingual mappings of word embeddings. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1:*

Long Papers, pages 789–798. Association for Computational Linguistics.

Pierre Isabelle, Colin Cherry, and George F. Foster. 2017. A challenge set approach to evaluating machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2486–2496. Association for Computational Linguistics.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation, NMT@ACL 2017, Vancouver, Canada, August 4, 2017*, pages 28–39. Association for Computational Linguistics.

Guillaume Lample, Myle Ott, Alexis Conneau, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2018. Phrase-based & neural unsupervised machine translation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 5039–5049. Association for Computational Linguistics.

Toshiaki Nakazawa, Manabu Yaguchi, Kiyotaka Uchimoto, Masao Utiyama, Eiichiro Sumita, Sadao Kurohashi, and Hitoshi Isahara. 2016. ASPEC: asian scientific paper excerpt corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portoroz, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA.*, pages 311–318. ACL.

Haiyun Peng, Erik Cambria, and Xiaomei Zou. 2017. Radical-based hierarchical embeddings for chinese sentiment analysis at sentence level. In *Proceedings of the Thirtieth International Florida Artificial Intelligence Research Society Conference, FLAIRS 2017, Marco Island, Florida, USA, May 22-24, 2017.*, pages 347–352. AAAI Press.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*. The Association for Computer Linguistics.

Xinlei Shi, Junjie Zhai, Xudong Yang, Zehua Xie, and Chao Liu. 2015. Radical embedding: Delving deeper to Chinese radicals. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 2: Short Papers*, pages 594–598. The Association for Computer Linguistics.

Yota Toyama, Makoto Miwa, and Yutaka Sasaki. 2017. Utilizing visual forms of Japanese characters for neural review classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017, Volume 2: Short Papers*, pages 378–382. Asian Federation of Natural Language Processing.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 5998–6008.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *CoRR*, abs/1609.08144.

Longtu Zhang and Mamoru Komachi. 2018. Neural machine translation of logographic language using sub-character level information. In *Proceedings of the Third Conference on Machine Translation: Research Papers, WMT 2018, Belgium, Brussels, October 31 - November 1, 2018*, pages 17–25. Association for Computational Linguistics.

FTA: a novel feature training approach for classification

Wanwan Zheng
Doshisha University
1-3 Tatara Miyakodani, Kyotanabe-shi, Kyoto-fu, Japan
teiwanwan@gmail.com

Mingzhe Jin
Doshisha University
1-3 Tatara Miyakodani, Kyotanabe-shi, Kyoto-fu, Japan
mjin@mail.doshisha.ac.jp

Abstract

Several studies have been conducted to find the best classification algorithm. Random Forest (RF) and Support Vector Machine (SVM) have been successfully introduced in various prediction models and served as the major data analysis tools that outperform many standard methods. However, RF has difficulties in achieving high accuracy when handling datasets with few instances or variables, and SVM is hard to produce good models if datasets have numerous variables. In this study, the Feature Training Approach (FTA) was proposed, which overcomes the weaknesses of RF and SVM by two trials, namely feature selection and training, SVM ensemble. According to the results of experiments, FTA is quite robust to the two types of data that cannot be well classified, i.e. data with few instances and variables, data with few instances and numerous variables. In most cases, even with different data from different domains, FTA could achieve better performance than RF and SVM.

1 Introduction

Machine learning has become a hot topic in various fields, and classification is a prominent task in machine learning. Data used for classification consists of instance and variable, which can fall into four cases: (1) data with few instances and variables; (2) data with numerous instances and few variables; (3) data with few instances and numerous variables; (4) data with numerous instances and variables. Because enough information is required to complete a statistical description of each class, it is well known that the training of classifiers requires considerable amount of training data (Zhu et al., 2016; Halevy et al., 2009; Mathur and Foody, 2008). However, even if there is considerable amount of data, the classification accuracy of classifiers is not necessarily high. Support Vector Machine (SVM, Cortes and Vapnik, 1995) is an example.

SVM as an effective data analysis tool has been successfully applied to various prediction models. Thanh and Kappas (2018) using Sentinel-2 image data examined and compared the performances of the RF, *k*-Nearest Neighbor (kNN), and SVM for land use/cover classification. According to their findings, SVM produces the highest accuracy with the least sensitivity to the training sample sizes. Kremic and Subasi (2016) applied RF and SVM in facial recognition. As a result, SVM achieves accuracy of 97.94% to the greatest, and RF is 97.17%. Chevalier et al. (2011) compared the performance of SVM with that of Neural Network (NN) in determining air temperature values, and they confirmed the superiority of SVM. Besides, some hybrid methods have been proposed based on SVM. Yong et al. (2015) developed a method based on the combination of Wavelet Transforms (WT) and SVM, which is optimized for those special cases where the real signals contain numerous events in the analyzed temporal window. Tests and trainings were performed using real complex signals, and the results showed the proposed methodology highly efficient. Zheng et al. (2014) proposed to combine *k*-means and SVM to increase the classification accuracy on the Wisconsin Diagnostic Breast Cancer (WDBC) dataset to 97.38%.

However, SVM's classification accuracy is affected by the noise involved in datasets for it uses all variables in tuning models. Thus, the accuracy is relatively low when dealing with high dimensional datasets.

33rd Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 316-323, Hakodate, Japan, September 13-15, 2019

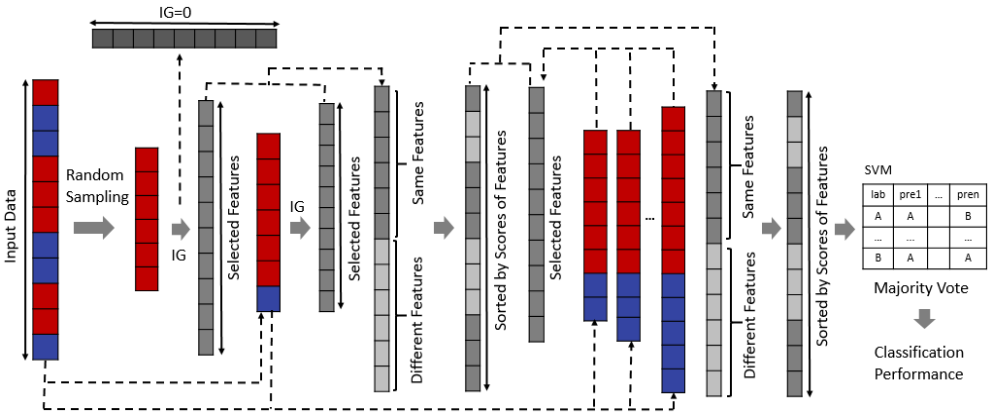


Figure 1: The overall procedure of FTA.

RF has also been extensively used since it's introduced in 2001 (Breiman, 2001). It also has become a standard classification approach in many fields. Couronne et al. (2017) presented a large-scale benchmarking experiment based on 260 real datasets to compare the performance of RF and logistic regression (LR) in prediction. As a result, all measures suggest a significantly better performance of RF. Chelgani et al., (2016) employed RF as a sensible tool for variable importance measurements using various coal properties to predict coke quality. According to the result, RF can further be a reliable and accurate technique to determine complex relationship by fuel and energy investigations. Liu et al. (2013) introduced and investigated RF, Back Propagation Neural Network (BPNN) and SVM to deal with electronic tongue data, and RF is proven to outperform BPNN and SVM.

RF has several advantages over other statistical modeling techniques: (1) capable of dealing with missing values and high-dimensional data; (2) capable of identifying complex interactions between variables and the most important variables; (3) high prediction accuracy; (4) robust against over-fitting. However, from the perspective of random sampling of instances and variables, RF is usually not very accurate when the numbers of samples or variables are small.

In this study, Feature Training Approach (FTA) is proposed as a new classification model which trains features and improves the weaknesses of RF and SVM.

The rest of this article is organized as follows. Sections 2 proposes the approach followed by experiments reported in Section 3. Section 4 gives an explanation why FTA works, and Section 5 draws the conclusions and states limitations and further work.

2 Feature Training Approach

The FTA refers to a two-phase hybrid approach. In the first phase, it performs feature selection and feature training alternately to make a list of selected features. In the second phase, SVM is used to make predictions. The same process is performed K times, and labels are finally determined for test data by majority vote. Figure 1 summarizes the overall procedure of FTA.

Information gain (IG) serves as base feature selection method, which has been validated as a representative feature selection method (Geurts et al., 2018; Chinnaswamy et al., 2017; Wosaiak and Dziomdziora, 2015; Adel et al., 2014). IG measures the reduction in entropy (impurity in an arbitrary collection of examples). With the entropy of Y defined as:

$$H(Y) = -\sum_{y \in Y} p(y) \log_2(p(y)) \quad (1)$$

Where $p(y)$ is the marginal probability density function for the random variable Y.

IG is defined as:

$$\Delta H = H - \frac{m_L}{m} H_L - \frac{m_R}{m} H_R \quad (2)$$

Where m is the total number of instances, with m_k instances belonging to class k (k=1,2,…,k).

IG, a supervised feature selection method, is more independent on the number of training

Split D_{all} into D_{tra} and D_{pre}

Input: (training) data $D_{tra} = \left\{\left(I_i, V_i\right)\right\}_{i=1}^{N}$ I: instance V: variable

1: **For** t=1 to K **do**:

Feature selection and feature training phase

2: Split D_{tra} into N-fold

3: **Input:** Define subspace s by extracting n fold ($n < N$) randomly

$$s = \bigcup_{j \in n_1} s_j \,,\ s' = \bigcup_{m \in n_2} s_m \,,\ n_1 + n_2 = N \,,\ s \cap s' = \varnothing$$

4: Perform feature selection using IG for s

5: Delete $V_i \left(IG_{V_i} = 0\right)$; Record $V_j \left(IG_{V_j} \neq 0\right)$

6: **For** m=1 to n_2 **do**:

7: $s = s + s'_m$

8: Perform feature selection using IG for s

9: Delete $V_{im} \left(IG_{V_i m} = 0\right)$; Record $V_{jm} \left(IG_{V_{jm}} \neq 0\right)$

10: $V_j = V_j \bigcup V_{jm}$, $V_{same} = V_j \bigcap V_{jm}$, $V_{diff} = V'_{same}$

11: $IG_{V_{same}} = mean(V_j, V_{jr})$, $IG_{V_{diff}} = IG_{V_{jm}}$

12: Sort V_j by the scores of IG_{V_j}

13: **End for;**

Output: Extract the top t features as the final list of selected features

Prediction phase

14: **Input:** (testing) data $D_{pre} = \left\{\left(I_i, V_i\right)\right\}_{i=1}^{M}$

15: Apply SVM using the final list of selected features

16: Build training model

17: **Output:** predictions for every instance in D_{pre}

18: **End For;**

Output: majority vote

Figure 2: Pseudo-code FTA.

Data type	#Samples	#Variables	#Datasets
Data with few instances and variables	5×n	13–40	30×10
Data with numerous instances and few variables	40×n–100×n	13–40	30×10
Data with few variables and numerous instances	5×n	294–3,645	30×10
Data with numerous instances and variables	100×n	294–3,645	30×10

n is the number of classes, in the experiment, n=2, 3

Table 1: Information of data.

samples than the unsupervised feature selection method (e.g. principal component analysis, PCA) and distance-based feature selection method (e.g. chi-squared) (Zheng and Jin, 2018). Inspired by this recognition, feature training as the core mechanism in the FTA gradually increases the amount of training samples and updates the list of selected features. In such a way, the same effect as repeated learning with different training data can be obtained. The pseudo code of FTA is shown in Figure 2.

	Reduction of dimension (%)		Mean			Win			p-value		
	min	max	SVM	RF	FTA	SVM	RF	FTA	SVM-RF	SVM-FTA	RF-FTA
Leukemia	75	88	0.8073	0.7954	0.9817	0	0	10		***	***
Bioresponse	93	99	0.5428	0.5695	0.9266	0	0	10		***	***
Gina agnostic	92	96	0.6132	0.5968	0.9229	0	0	10		***	***
Scene	63	91	0.7363	0.6744	0.9450	2	1	10		**	***
Isolet	61	77	0.9588	0.9497	0.9817	7	5	9			
Speech	97	99	0.7700	0.7346	0.8182	0	0	10		*	***
Robert	89	94	0.5949	0.5153	0.9489	0	0	10		***	***
Christine	86	93	0.5332	0.5013	0.9541	0	0	10		***	***
Madelon	89	98	0.4066	0.4176	0.9908	0	0	10		***	***
Arcene	78	98	0.4308	0.4576	0.8052	1	1	10		***	***
Character Font_ARIAL	84	96	0.5613	0.5206	0.8477	0	0	10		***	***
Character Font_CALIBRI	91	98	0.4827	0.3871	0.8484	0	0	10		***	***
Character Font_COURIER	81	99	0.5146	0.5300	0.8861	0	0	10		***	**
Character Font_LUCIDA	93	98	0.4028	0.3755	0.7773	0	0	10		***	***
Character Font_NIRMALA	84	96	0.5793	0.5862	0.9450	0	0	10		***	***
cifar-10-small(0,1,2)	88	97	0.4519	0.4936	0.7871	0	0	10		***	***
cifar-10-small(3,4,5)	80	98	0.3446	0.3341	0.7559	0	0	10		***	***
cifar-10-small(6,7)	79	97	0.6066	0.6017	0.9033	1	0	10		**	**
cifar-10-small(8,9)	76	92	0.6545	0.7112	0.9337	0	0	10		**	*
Eating(1,2,3)	91	95	0.3875	0.4239	0.5829	1	1	8		**	*
Eating(4,5)	84	92	0.5830	0.6878	0.8138	2	1	9		*	
Eating(6,7)	82	92	0.6021	0.7751	0.8545	2	6	8	†	**	
Fashion_Mnist(0,1,2)	65	91	0.8873	0.8762	0.9276	5	4	7			
Fashion_Mnist(3,4,5)	59	82	0.8255	0.8528	0.8785	4	4	8			
har(1,2,3)	65	91	0.7512	0.7912	0.9469	0	0	10		**	**
har(4,5,6)	63	94	0.7167	0.6079	0.8865	0	1	10	†	**	***
svhn(1,2,3)	85	97	0.4645	0.3294	0.7591	0	0	10		**	***
svhn(4,5,6)	93	96	0.3468	0.3183	0.7801	0	0	10		***	***
svhn(7,8)	81	95	0.5131	0.5110	0.8861	0	0	10		***	***
svhn(9,10)	82	96	0.5385	0.5013	0.8888	0	0	10		***	***
mean	**81.0349**	**94.0384**	**0.5869**	**0.5809**	**0.8721**	**0.8333**	**0.8000**	**9.6333**			

*** p < 0.001, ** p < 0.01, * p < 0.05, † p < 0.1

Table 2: Results of the benchmarking experiment using the macro-averaged F-measure. (Data with few instances and numerous variables)

3 Experiments

3.1 Analysis data

Experiments were run on a total of 60 benchmark datasets (30 datasets with numerous variables, 30 datasets with few variables), covering biological data, image data, voice recognition data, physical data and artificial data. Furthermore, to generate the four types of data mentioned in Section 1, random sampling was performed 10 times respectively. The information of data is listed in Table 1.

For data with few instances and variables, the number of instances is set as $5\times n$ (n is the number of classes, each of which has 5 instances), the number of variables is between 13 and 40, and the number of datasets is 30×10 (30 datasets, random sampling was performed 10 times for each dataset). For data with numerous instances, the number of instances is set as 100. The classifications include binary classification and 3-class classification. Macro-averaged F-measure serves as the evaluation metric.

3.2 Experimental results

In this study, for datasets with 5×n instances and 40×n–100×n instances, leave-one-out cross validation (LOOCV) and 10-fold cross validation was conducted, respectively. Furthermore, all the features selected after training were applied as the final list of selected features.

For classifiers, because the probability of overfitting increases with the increase in the number of variables, one of the most challenging tasks is to make correct prediction of data with few instances and numerous variables. The classifier requires the ability to create a learning model that describes the characteristics of data with few instances. Table 2 shows the result of data with few instances and numerous variables.

FTA reduced the dimension of data to the minimum 81.0349% and the maximum of 94.0384% by average. The average of macro-averaged F-measure of FTA, RF, and SVM are 0.8721, 0.5809, 0.5869, respectively, and the average numbers of wins of FTA, RF, and SVM

	Reduction of dimension (%)		Mean			Win			p-value		
	min	max	SVM	RF	FTA	SVM	RF	FTA	SVM-RF	SVM-FTA	RF-FTA
Cardiotocography	78	92	0.4040	0.5424	0.6874	0	3	7		**	
WDBC	27	67	0.8061	0.8644	0.8308	6	7	7			
Vehicle	21	92	0.4485	0.5420	0.6505	0	1	9		**	
Waveform	69	86	0.7152	0.6754	0.8155	3	2	8			†
Software	13	88	0.6742	0.7652	0.8324	3	5	8		*	
Climate	75	86	0.7065	0.7104	0.8733	2	2	8			
HallofFame	14	71	0.5902	0.6106	0.6596	2	5	5			
Fri	80	90	0.5149	0.6064	0.8361	0	2	10		**	*
analcatdata_authorship	66	85	0.9424	0.9526	0.9630	7	7	8			
zernike(1,2,3)	29	61	0.9878	0.8828	0.9878	10	0	10	***		***
zernike(4,5,6)	63	83	0.7702	0.6846	0.8737	0	1	10	†	*	***
zernike(7,8)	36	77	0.8629	0.8821	0.9433	4	5	8			
zernike(9,10)	51	81	0.9056	0.8561	0.9908	4	2	10	†		**
first-order-theorem(1,2,3)	80	100	0.3878	0.4756	0.6544	0	0	10		***	**
first-order-theorem(4,5,6)	78	94	0.3450	0.4416	0.6196	0	0	10		**	†
gesturehaseSegmentation (DHP)	80	97	0.4449	0.4389	0.5033	2	1	7			
gesturehaseSegmentation(RS)	79	94	0.5342	0.5723	0.7395	1	2	9		**	*
hillVally	3	50	0.4666	0.4381	0.5879	2	2	10			
kc1	9	85	0.5328	0.5626	0.6551	2	4	8			
musk	74	94	0.5840	0.5631	0.7876	2	2	9		*	**
ozone-level-8hr	53	94	0.6117	0.7144	0.8569	0	3	9		**	
qsar-biodeg	59	90	0.6517	0.7721	0.8541	3	3	9		*	
semeion(1,2,3)	86	96	0.9175	0.9167	1.0000	2	2	10		**	**
semeion(4,5,6)	87	99	0.9052	0.9049	0.9744	2	3	9		*	*
semeion(7,8)	89	98	0.8861	0.9233	0.9908	2	4	10		**	†
semeion(9,10)	84	96	0.9317	0.8669	0.9800	7	2	10			*
spambase	73	93	0.5727	0.7789	0.7940	1	5	5	*	**	
steel-plateds-fault	44	90	0.5662	0.5756	0.7503	3	0	9		*	*
wall-robot-navigation (1,2)	58	92	0.5008	0.6284	0.7248	1	3	7		*	
wall-robot-navigation (3,4)	33	58	0.9417	0.9817	0.9541	6	9	7			
mean	**56.3667**	**85.9667**	**0.6703**	**0.7043**	**0.8124**	**2.5667**	**2.9000**	**8.5333**			

*** p < 0.001, ** p < 0.01, * p < 0.05,† p < 0.1

Table 3: Results of the benchmarking experiment using the macro-averaged F-measure. (Data with few instances and variables)

are 9.6333, 0.8000, 0.8333, respectively. These average numbers of wins were calculated based on the number of wins of every classifier in terms of macro-averaged F-measure per dataset after 10 times of random sampling. Furthermore, the Tukey's honest significant difference method was employed to verify whether there exists significant difference between any two classifiers. According to the result, significant difference was found between FTA and RF, SVM in most cases.

For classifiers, another challenging task is to correctly predict data with few instances and variables. Halevy et al. (2009) reported that even very complex problems in artificial intelligence may be solved by simple statistical models trained on massive datasets. And numerous research have shown that classification accuracy tends to be positively related to training dataset size (Zhu et al., 2015, Mathur and Foody., 2008, Foody and Mathur, 2004, Pal and Mather, 2003). Because classifiers require enough training data to complete the statistical description of each class, few instances and variables mean that the information used to tune model may probably be insufficient. Table 3 shows the result of data with few instances and variables.

FTA reduced the dimension of data to the minimum of 56.3667% and to the maximum of 85.9667% by average. The average of macro-averaged F-measure of FTA, RF, and SVM are 0.8124, 0.7043, 0.6703, respectively, and the average numbers of wins of FTA, RF, and SVM are 8.5333, 2.9000, 2.5667, respectively. Furthermore, according to the results of Tukey's honest significant difference method, there exists significant difference between FTA and RF, SVM in most cases.

The results of the other two types of datasets are summarized in Table 4 and Table 5, respectively.

For data with numerous instances and variables, it is considered that RF should be good at dealing with such datasets. The average of macro-averaged F-measure of FTA, RF, and SVM are 0.7572, 0.7593, 0.7140, respectively. FTA performs as well as RF does. The average numbers of wins of FTA, RF, and SVM are 5.9333, 3.7333, 1.5667, respectively. Moreover, there exists significant difference between FTA and RF, SVM in almost half cases according to

	Reduction of dimension (%)		Mean			Win			p-value		
	min	max	SVM	RF	FTA	SVM	RF	FTA	SVM-RF	SVM-FTA	RF-FTA
Leukemia	73	84	0.9731	0.9679	0.9854	6	4	10		†	*
Bioresponse	98	100	0.6399	0.6911	0.7418	0	1	9	*	***	*
Gina agnostic	92	95	0.8268	0.8484	0.8700	0	0	10	*	***	*
Scene	52	68	0.8103	0.8177	0.8209	3	6	5			
Isolet	39	44	0.9960	0.9850	0.9957	10	2	10	***		***
Speech	88	92	0.5588	0.5407	1.0000	0	0	10		***	***
Robert	58	75	0.7026	0.7386	0.7131	1	7	2	*		
Christine	84	98	0.6762	0.6668	0.7037	1	0	10			
Madelon	98	100	0.5753	0.5535	0.6458	0	0	10		**	***
Arcene	92	100	0.7242	0.7605	0.8319	0	0	10	*	***	***
Character Font_ARIAL	15	30	0.7650	0.7797	0.7668	1	8	1			
Character Font_CALIBRI	93	99	0.6554	0.6671	0.6772	0	3	7			
Character Font_COURIER	75	96	0.7088	0.7724	0.7045	0	10	0	**		**
Character Font_LUCIDA	85	98	0.5502	0.5804	0.5430	2	7	1			
Character Font_NIRMALA	95	99	0.5781	0.5806	0.6405	0	0	10		***	**
cifar-10-small(0,1,2)	43	63	0.6933	0.6696	0.6900	5	0	5			
cifar-10-small(3,4,5)	84	99	0.5061	0.5163	0.5281	3	1	6			
cifar-10-small(6,7)	61	85	0.7727	0.7857	0.7958	1	2	8			
cifar-10-small(8,9)	66	88	0.7647	0.7678	0.7790	2	1	8			
Eating(1,2,3)	87	92	0.3558	0.6871	0.3826	0	10	0	***	*	***
Eating(4,5)	75	79	0.6679	0.9521	0.6511	0	10	0	***		***
Eating(6,7)	73	82	0.6594	0.9226	0.6620	0	10	0	***		***
Fashion_Mnist(0,1,2)	20	35	0.9513	0.9482	0.9522	3	3	6			
Fashion_Mnist(3,4,5)	9	14	0.9405	0.9399	0.9408	5	5	5			
har(1,2,3)	23	29	0.9640	0.9357	0.9643	4	0	7	***		***
har(4,5,6)	67	77	0.9041	0.9328	0.9119	0	9	1	***		**
svhn(123)	96	99	0.5610	0.6483	0.6564	0	3	7	***	***	
svhn(456)	95	99	0.5544	0.6429	0.6259	0	6	4	***	***	
svhn(78)	93	99	0.7361	0.7829	0.7899	0	3	7	†	*	
svhn(910)	96	100	0.6480	0.6962	0.7461	0	1	9		***	**
mean	**70.8333**	**80.6000**	**0.7140**	**0.7593**	**0.7572**	**1.5667**	**3.7333**	**5.9333**			

*** p < 0.001, ** p < 0.01, * p < 0.05,† p < 0.1

Table 4: Results of the benchmarking experiment using the macro-averaged F-measure.
(Data with numerous instances and variables)

	Reduction of dimension (%)		Mean			Win			p-value		
	Min	max	SVM	RF	FTA	SVM	RF	FTA	SVM-RF	SVM-FTA	RF-FTA
Cardiotocography	66	81	0.4462	0.8958	0.6585	0	10	0	***	***	***
WDBC	10	23	0.9131	0.9471	0.9166	1	10	1	***		***
Vehicle	4	36	0.5709	0.6891	0.5801	0	10	0	***		***
Waveform	55	57	0.8500	0.8436	0.8560	2	3	7			
Software	37	82	0.6109	0.7492	0.6672	0	10	0	***	*	**
Climate	81	88	0.7670	0.7791	0.8135	0	2	8			
HallofFame	0	7	0.7353	0.7507	0.7374	3	8	2			
Fri	84	90	0.6390	0.8368	0.8852	0	0	10	***	***	**
analcatdata_authorship	17	23	0.9897	0.9880	0.9887	5	3	5			
zernike(1,2,3)	2	12	0.9950	0.9910	0.9957	9	0	10	*		**
zernike(4,5,6)	18	27	0.9459	0.9177	0.9442	6	0	5	***		***
zernike(7,8)	10	23	0.9831	0.9696	0.9831	6	0	8	**		**
zernike(9,10)	29	36	0.9821	0.9841	0.9856	3	6	5			
first-order-theorem(1,2,3)	70	94	0.5057	0.5311	0.5127	2	4	5			
first-order-theorem(4,5,6)	56	69	0.5692	0.5914	0.5708	2	6	2			
gesturehaseSegmentation(DHP)	21	49	0.5513	0.6141	0.5532	1	8	1	***		***
gesturehaseSegmentation(RS)	83	93	0.6663	0.6942	0.6448	1	8	1			*
hillVally	23	45	0.5056	0.5291	0.6173	0	0	10		***	***
kc1	5	18	0.7031	0.6963	0.7029	2	3	6			
musk	37	71	0.8371	0.8547	0.8323	2	6	2			
ozone-level-8hr	16	34	0.7439	0.8203	0.7670	0	10	0	***		**
qsar-biodeg	33	57	0.7807	0.8317	0.8140	0	8	2	**	*	
semeion(1,2,3)	26	31	0.9781	0.9707	0.9771	6	0	4	**		*
semeion(4,5,6)	30	36	0.9864	0.9747	0.9864	5	0	6	***		***
semeion(7,8)	67	73	0.9861	0.9867	0.9891	3	3	9			
semeion(9,10)	56	58	0.9920	0.9910	0.9935	6	4	7			
spambase	38	56	0.7010	0.8979	0.7040	0	10	0	***		***
steel-plateds-fault	33	70	0.6311	0.9189	0.6280	0	10	0	***		***
wall-robot-navigation 12	33	58	0.8124	0.9773	0.8184	0	10	0	***		***
wall-robot-navigation 34	4	4	0.9801	0.9925	0.9796	1	10	1	***		***
mean	**34.8000**	**50.0333**	**0.7786**	**0.8405**	**0.8034**	**2.2000**	**5.4000**	**3.9000**			

*** p < 0.001, ** p < 0.01, * p < 0.05,† p < 0.1

Table 5: Results of the benchmarking experiment using the macro-averaged F-measure.
(Data with numerous instances and few variables)

the results of Tukey's honest significant difference method.

For data with numerous instances and few variables, rising the number of instances will bring advantages to RF and SVM. According to the result, the average of macro-averaged F-measure of FTA, RF, and SVM are 0.8034, 0.8405, 0.7786, respectively. The average numbers of wins of FTA, RF, and SVM are 3.9000, 5.4000, 2.2000, respectively. Besides, there exists significant difference between RF and FTA, SVM in almost half cases according to the results of Tukey's honest significant difference method. Therefore, RF is considered the best, followed by FTA and SVM.

4 The reasons why FTA works

The reasons why FTA works are concluded as follows:

1. Because FTA has feature selection process, FTA is expected to work better than SVM in dealing with data with numerous variables.
2. By gradually increasing the amount of training samples and updating the list of selected features, the same effect as repeated learning with different training data is obtained. FTA is expected to be superior to RF in handling data with few instances.
3. Introduction of the majority vote can ensure the high accuracy to a certain extend.

5 Conclusion

This study proposed FTA as a variable choice which is based on feature training. As proven in this benchmark study, FTA (1) provides more accurate models than RF and SVM in handling two types of challenging data which is difficult to make correct prediction for classifiers (i.e. data with few instances and variables, data with few instances and numerous variables), and data with numerous instances and variables; (2) For data with numerous instances and few variables, FTA ranks in the middle of RF and SVM; (3) This time only the well-balanced data was used, whereas, FTA may also work with data with high skew if IG is converted to BNS (Forman, 2003), which was previously shown to substantially improve classification accuracy, especially when dealing with tasks with high skew.

For the limitations of FTA, we do note that FTA is time-consuming especially when dealing with data with numerous instances or variables. This is considered primarily coming from SVM, the feature training and the number of runs in order to make the majority vote. The number of runs was set to 101 in this study, however it might be possible to further improve the model by automatically stopping FTA when a certain great model is made. Furthermore, all the features selected after training were used as the final list of selected features this time, the model may be further improved with a well set of the top *t* features as the final list of selected features.

Caigny et al. (2018) proposed the logit leaf model (LLM), which is constructed based on logistic regression and decision trees. In their experiment, LLM provides more accurate models than logistic regression and decision trees, and performs at least as well as RF and logistic model trees (LMT). As a feature work, the comparison will be made between LLM and FTA. Furthermore, the combination of SVM and IG will also be added as a comparison task in the future work.

References

Aisha Adel, Nazlia Omar, and Adel Al-Shabi. 2014. A comparative study of combined feature selection methods for arabic text classification. Journal of Computer Science, 10(11):2232–2239. https://thescipub.com/abstract/10.3844/jcssp.2014.2232.2239.

Leo Breiman. 2001. Random forests. Machine Learning, 45(1):5–32. https://link.springer.com/article/10.1023/A:1010933404324.

Arno De Caigny, Kristof Coussement, and Koen W. De Bock. 2018. A new hybrid classification algorithm for customer churn prediction based on logistic regression and decision trees. European Journal of Operational Research, 269(2):760-772. https://www.sciencedirect.com/science/article/pii/S0377221718301243.

Arunkumar Chinnaswamy and Ramakrishnan Srinivasan. 2017. Hybrid information gain based fuzzy roughest feature selection in cancer microarray data. In Proceedings of International Conference on Innovations in Power and Advanced Computing Technologies, pages 1–6, Canberra, Australia. https://ieeexplore.ieee.org/document/8244875.

Corinna Cortes and Vladimir Vapnik. 1995. Support vector networks. Machine Learning, 20(3):273– 297.

https://link.springer.com/article/10.1007/BF00994018.

Robert F. Chevalier, Gerrit Hoogenboom, Ronald W. McClendon, and Joel A. Paz. 2011. Support vector regression with reduced training sets for air temperature prediction: a comparison with artificial neural networks. Neural Computing and Applications, 20(1):151–159. https://link.springer.com/article/10.1007/s00521-010-0363-y.

Raphael Couronné, Philipp Probst, and Anne-Laure Boulesteix. 2018. Random forest versus logistic regression: a large-scale benchmark experiment. BMC Bioinformatics, 19(270). https://doi.org/10.1186/s12859-018-2264-5.

S. Chehreh Chelgania and S. S. Matinb and James C.Howerc. 2016. Explaining relationships between coke quality index and coal properties by random forest method. Fuel, 182(15):754–760. https://www.sciencedirect.com/science/article/pii/S0016236116304860.

Giles M. Foody and Ajay Mathur. 2004. A relative evaluation of multiclass image classification by support vector machines. IEEE Transactions on Geoscience and Remote Sensing, pages 1335–1343, https://ieeexplore.ieee.org/abstract/document/1304900/authors#authors.

George Forman. 2003. An extensive empirical study of feature selection metrics for text classification. Journal of Machine Learning Research. 3(2003):1289-1305. http://www.jmlr.org/papers/volume3/forman03a/forman03a.pdf.

Renate Geurts, Karl Ask, Pär Anders Granhag, and Aldert Vrij. 2018. Interviewing to manage threats: Exploring the effects of interview style on information gain and threateners' counter-interview strategies. Journal of Threat Assessment and Management, 5(4):189–204. https://psycnet.apa.org/record/2018-63621-001.

Alon Halevy, Peter Borvig, and Fernando Pereira. 2009. The unreasonable effectiveness of data. IEEE Intelligent Systems, 24(2):8–12. https://static.googleusercontent.com/media/research.google.com/ja//pubs/archive/35179.pdf.

Emir Kremic and Abdulhamit Subasi. 2016. Performance of random forest and svm in face recognition. The International Arab Journal of Information Technology, 13(2):287–293. http://www.ccis2k.org/iajit/PDF/Vol.13,%20No.2/8468.pdf.

Miao Liu, Mingjun Wang, Jun Wanga, and Duo Li. 2013. Comparison of random forest, support vector machine and back propagation neural network for electronic tongue data classification: Application to the recognition of orange beverage and Chinese vinegar. Sensors and Actuators B: Chemical, 177:970–980. https://www.sciencedirect.com/science/article/pii/S0925400512012671.

Ajay Mathur and Giles M. Foody. 2008. Crop classification by a support vector machine with intelligently selected training data for operational application. Computing Reviews, 24(11):503–512. https://www.tandfonline.com/doi/abs/10.1080/01431160701395203.

Phan Thanh Noi and Martin Kappas. 2018. Comparison of random forest, k-nearest neighbor, and support vector machine classifiers for land cover classification using sentinel-2 imagery. Sensors, 18(1). https://doi.org/10.3390/s18010018.

Mahesh Pal and Paul M. Mather. 2003. An assessment of the effectiveness of decision tree methods for land cover classification. Remote Sensing of Environment. 86(4):554-564. https://www.scirp.org/(S(351jmbntvnsjt1aadkposzje))/reference/ReferencesPapers.aspx?ReferenceID=1526874.

Agnieszka Wosiak and Agata Dziomdziora. 2015. Feature selection and classification pairwise combinations for high-dimensional tumour biomedical datasets. Schedae Informaticae, 24:53–62. http://www.ejournals.eu/sj/index.php/SI/article/view/6334.

D. De Yong, S. Bhowmik, and F. Magnago. 2015. An effective power quality classifier using wavelet transform and support vector machines. Expert Systems with Applications, 42(15):6075–6081. https://www.sciencedirect.com/science/article/pii/S0957417415002328.

Bichen Zheng, Sang Won Yoon, and Sarah S.Lam. 2014. Breast cancer diagnosis based on feature extraction using a hybrid of k-means and support vector machine algorithms. Expert Systems with Applications, 41(4):1476–1482. https://www.sciencedirect.com/science/article/pii/S0957417413006659.

Zhihua Zhou and Ji Feng. 2018. Deep forest. National Science Review, 6(1):74–86. https://arxiv.org/abs/1702.08835.

Bi-Directional Decoder Model with Efficient Fine-Tuning of Embedding for Named Entity Recognition

Panuwat Assawinjaipetch
Japan Advanced Institute of Science and Technology
a.panuwat@jaist.ac.jp

Kiyoaki Shirai
Japan Advanced Institute of Science and Technology
kshirai@jaist.ac.jp

Virach Sornlertlamvanich
Sirindhorn International Institute of Technology
virach@siit.tu.ac.th

Sanparith Marukatat
Thailand's National Electronics and Computer Technology Center
sanparith.marukatat@nectec.or.th

Abstract

Named Entity Recognition (NER) is one of the important tasks in natural language processing. In this paper, we propose a novel method for NER in Japanese. It consists of three deep learning modules: a Character Encoder, Word Encoder and Tag Decoder, which are implemented by Long Short-Term Memory (LSTM) or Bi-Directional LSTM (BiLSTM). Pre-trained character and word embeddings are used as the input of our model. Our new idea is to combine a forward and backward LSTM at Tag Decoder. This enables us to consider named entity (NE) tags of both the previous and succeeding words in the classification, while only a previous NE tag was taken into account in most of the past studies. We also introduce a separate fine-tuning of the embedding that enables us to efficiently fine-tune the parameters of the character and word embeddings. In this method, the parameters of embeddings and other model parameters are trained separately. In an experiment using a large Japanese named entity tagged corpus, the F1-score of our proposed method was 0.944, which was better than the baseline by 0.06 points.

1 Introduction

Named Entities Recognition (NER) is the task of identifying named entities, such as person, location, organization, and so on, in a text. An NER system is often used as a core system in various types of Natural Language Processing (NLP) including question answering, information retrieval, dialogue system, topic modeling, etc. In modern research, NER systems are implemented as classifiers using abstract representation of a sentence as features, where an abstract representation is obtained by deep learning architectures.

NER is usually defined as a sequential labeling problem. Regarding a word sequence of a given sentence as a time sequence, named entity (NE) tags for words are determined one by one from the first to last words. In order to classify an NE tag of a word in a certain time step, it is common to use an NE tag determined in the previous time step as a feature for classification. In this research, we propose a novel method to use the NE tags of not only the previous but also the succeeding words in a deep learning model.

On the other hand, fine-tuning of the word embedding is widely applied in deep learning models for NLP. That is, the word embedding is pre-trained using a huge amount of texts in a general domain, then the parameters of the word embedding are updated using a relatively small amount of data that is specific to the target domain. Another contribution of this paper is to propose a novel method for fine-tuning of the word and character embeddings. It performs parameter estimation with a deep neural network and fine-tuning of embeddings separately.

2 Related work

2.1 Natural language processing with deep learning

Recently, deep learning has been actively studied in the NLP research field. Collobert et al. (2011) intro-

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 324-333, Hakodate, Japan, September 13-15, 2019

duced a neural network model for NLP. Their system used minimal feature engineering but achieved promising results. However, the proposed feed forward network could not capture relations between the words in a sentence, although they can be a useful feature for various NLP tasks. Recurrent Neural Network (RNN) has been proposed, which is well-known for its ability to detect hidden relationship between words. Later, RNN and its variants have been applied for many NLP applications. RNN can be generally classified into three types. The first one is the traditional RNN called the Hopfield network, proposed by Hopfield (1982). The second one is the most popular network, called Long Short-Term Memory (LSTM), proposed by Hochreiter and Schmidhuber (1997). It was also the first network that tried to mitigate the vanishing gradient problem. The third one is Gate Recurrent Unit (GRU) proposed by Cho et al. (2014). Since the second and third networks were less sensitive to the problem of vanishing gradient, most modern research tends to use them rather than the traditional RNN. However, it is still uncertain which is better, LSTM or GRU. Finally, models that combined neural network in forward and backward directions, such as bi-directional LSTM (BiLSTM) (Graves and Schmidhuber, 2005), achieved further improvement due to their ability to capture left and right contexts.

2.2 Deep neural network for named entity recognition

The modern neural architectures for NER can be broadly classified into categories according to their representation of an input sentence. The representation can be based on words, characters, and other features such as affix *n*-gram, as well as combinations of these.

2.2.1 NER based on word embedding

In the usual deep neural networks for NER, a sequence of words in a sentence is given as an input. Usually, each word is represented by a word embedding and given to the neural networks. Collobert et al. (2011) first introduced a convolutional neural network that accepted a word sequence as an input. Then, several methods of RNN that handled a sequence of words were proposed (Mesnil et al., 2013; Nguyen et al., 2016). Huang et al. (2015) proposed LSTM with Conditional Random Field (CRF) (Lafferty et al., 2001) that achieved an 84.26% F1 score on the CoNLL-2003 English data set (Tjong Kim Sang and De Meulder, 2003). By slightly modifying this model, Shao et al. (2016) proposed a window-based bi-directional LSTM for NER. Neural network models for NER on specific domains (e.g. the medical domain) have also been investigated (Chalapathy et al., 2016; Xu et al., 2018).

2.2.2 NER based on character embeddings

Each sentence is taken to be a sequence of characters in several previous methods. Each character is converted into a vector representation by character embedding. The potential of the character NER neural model was first highlighted by Kim et al. (2016). The character based architecture has the ability to tackle an out-of-vocabulary problem and can improve the performance of NER in morphologically rich languages. The architecture was applied to various languages such as Vietnamese (Pham and Phuong, 2017) and Chinese (Dong et al., 2016). Kuru et al. (2016) applied a character based model to 7 different languages.

2.2.3 NER based on word and character embeddings

Several studies have proven that the incorporation of both word and character sequences in a neural network model can contribute to develop a strong NER system. Ma and Hovy (2016) and Chiu and Nichols (2016) proposed such models and achieved 91.21% and 91.62% F1-score on the CoNLL-2003 English data set, respectively. Misawa et al. (2017) proposed a model using word and character emdeddings for Japanese NER. Lample et al. (2016) and Yang et al. (2016) applied BiLSTM and GRU for feature extraction at both the word and character levels.

2.3 Output layer in neural based NER models

Most neural network models for NER consist of two parts. One is constituted by the layers to obtain an abstract representation of an input from word and/or character embeddings. The other is constituted by the layers that determine an NER tag for each word based on the abstract representation. Hereafter, we call the latter the *hidden2tag* layer. CRF has been commonly used in the *hidden2tag* layer. However,

other networks are also used. For example, **?**) and Mesnil et al. (2013) proposed methods to use RNN and feed-forward network in the *hidden2tag* layer.

However, to the best of our knowledge, no bi-directional RNN or LSTM has been used in the *hidden2tag* layer. As discussed in Section 1, NER is usually regarded as a sequential labeling problem, where the NER tags of words are determined one by one. Furthermore, the NE tag of the previous word is commonly used in the classification of the NE tag for the current word. However, the NE tag of the succeeding word cannot be used as a feature, since it is not determined yet in a sequential labeling. If the NE tags are determined in a backward direction (from the last to first word), the succeeding NE tag can be used, but the previous NE tag cannot. Therefore, the bi-directional RNN or LSTM cannot be applicable to the *hidden2tag* layer. Intuitively, both the NER tags of the previous and succeeding words are effective for NER. This paper proposes a way to use both of them.

Mesnil et al. (2013) proposed a model called bi-directional Jordan-type RNN for the slot filling task in spoken language. Their model also considered both the previous and succeeding output tags, but at time t only the output tags in the near words from $t-T$ to $t+T$ were used. Our model combines the ordinary forward and backward LSTM that can take long dependencies into account.

3 Proposed method

3.1 Task

We define classes of named entities following Sekine's extended named entity hierarchy (version 7.1.0)(Sekine and Nobata, 2004)[1], which consists of 200 fine grained named entity classes. Since the number of NE classes is large, coarse grained NE types in the hierarchy are used. Table 1 shows a list of the 26 NE classes in our NER task. An extended named entity annotated corpus in Japanese (Hasimoto et al., 2008) was used to develop our method.[2] In the corpus, named entities are annotated with IOB encoding, where the NE classes are those in Sekine's extended named entity hierarchy. The corpus consists of 8,228 articles, 53,224 distinct words, and 2,226,147 tokens. It is about 7.4 times larger than the CoNLL-2003 English corpus, which consists of 1,393 articles and 301,418 tokens. As preprocessing, we used the tool CaboCha (Taku Kudo, 2002) for word segmentation, POS tagging, and chunking.

Table 1: List of named entity classes

God, Percent, Location, Latitude Longitude, Product, Ordinal Number, Name Other, Numex Other, Multiplication, School Age, Timex, Age, Natural Object, Disease, Person, Organization, Facility, Colour, Money, Point, Rank, Countx, Frequency, Measurement, Event, Period

The graphical notation of the task definition is illustrated in Figure 1. An input of our model is a sentence represented as a sequence of words $\{w_1 \cdots w_n\}$. The number of the words for each sentence is fixed at n: padding is used when the length of a sentence is less than n. A sentence is also represented as a sequence of characters $\{c_1 \cdots c_m\}$. Similarly, the number of the characters in a sentence, denoted by m, is also fixed. Following the IOB encoding of named entities, the model predicts an output tag t_i for each word w_i. t_i is represented by B_x, I_x or O, where x stands for a type of a name entity (e.g. "Organization", "Person").

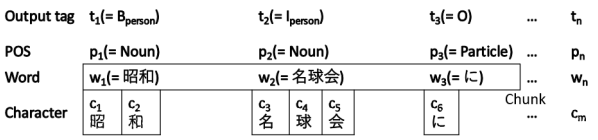


Figure 1: Task definition

3.2 Model

Our proposed model is based on Shen's architecture (Shen et al., 2018) , which achieved a 90.89% F1-score on the CoNLL-2013 dataset. An overview of our model is shown in Figure 2. It consists of three modules: Character Encoder, Word Encoder and Tag Decoder. Tag Decoder corresponds to the *hidden2tag* layer.

[1] https://nlp.cs.nyu.edu/ene/

[2] Although the corpus includes newspaper articles and white papers, only news texts were used in this study.

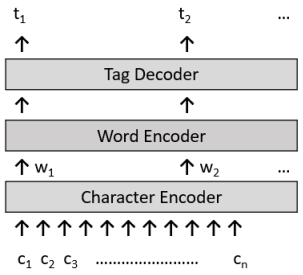


Figure 2: Overview of NER model

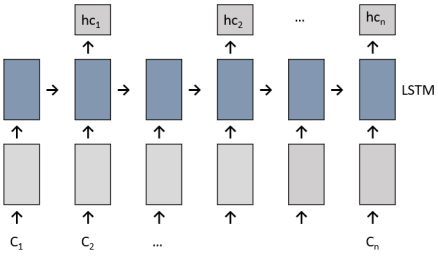


Figure 3: Character Encoder

3.2.1 Character Encoder

Character Encoder produces an abstract representation of a word from a sequence of characters. The motivation of introducing this module is that some Japanese characters indicate types of named entities. For example, the character c_5 "会(*kai*)" in Figure 1 literally means "association". A proper noun including it tends to be classified as an organization. Similarly, a proper noun including the character "岳(*gaku*)", which means "mountain", tends to be a location.

The architecture of Character Encoder is shown in Figure 3. Character embedding is entered as an input of Character Encoder. Character embedding is pre-trained from the training corpus by the skip-gram model implemented by the *word2vec* tool (Mikolov et al., 2013). In our preliminary experiment, it was found that a single direction LSTM was slightly better than BiLSTM. Thus, we apply a single direction LSTM, while Shen et al. (2018) used BiLSTM in Character Encoder. The hidden state of the last character of each word is passed to the next module. The output of Character Encoder applied to the *i*th word is denoted by hc_i.

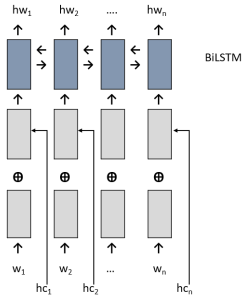


Figure 4: Word Encoder

3.2.2 Word Encoder

Word Encoder produces contextual information of the words in a sentence. The architecture of Word Encoder is shown in Figure 4. The word embedding of w_i and the output of Character Encoder hc_i are concatenated, then this is passed to the BiLSTM model. Finally, the hidden state of each word hw_i is obtained as the output of this module. It is almost the same as Shen's original model (Shen et al., 2018) except that CNN was used in their model. In addition, the word embedding is pre-trained from the training data by the skip-gram model using the *word2vec* tool.

3.2.3 Tag Decoder

For each word w_i, Tag Decoder predicts the output vector (denoted by o_i) that represents the distribution of the scores of the output tags. The architecture of Tag Decoder is shown in Figure 5. The output of Word Encoder hw_i and the previous output vector o_{i-1} are concatenated and passed to LSTM. The output of LSTM (hd_i) is augmented by two additional features. One is the POS embedding pe_i that represents the information of the POS (p_i) of each word. The other is the Japanese particle embedding jp that represents the syntactic information of the chunk.[3] Since we believe that both POS and the Japanese particle are effective for NER, the hw_i are concatenated with pe_i and jp. Then, they are entered to a

[3]"Particle" is one of the POSs and represents a case maker in Japanese. It plays an important grammatical role, especially in determining the type of a chunk. For example, a chunk "noun + *ga*" represents a nominative case of a predicate, while "noun + *ni*" represents a dative case ("*ga*" and "*ni*" are Japanese particles). For each chunk, the particle that appears in the rightmost position (denoted by JP in Figure 5) is identified, which determines the grammatical role of the chunk. Then the embedding of JP, denoted by jp, is added to hd_i of all the words in the chunk.

feed forward network (FFN) to determine the output vector o_i. As for the final result, a single NE tag t_i for each word is determined by the index of the highest value in the output vector o_i.

3.2.4 Two directional Tag Decoder

We propose a new tag decoder that uses the information of the previous NE tag (the output vector o_{i-1}) and the succeeding NE tag (o_{i+1}) for the decoding of t_i, since we can assume that both of them are effective features. Note that BiLSTM cannot be applied as Tag Decoder. o_{i+1} is not predicted by the model at time i when the NE tags are determined in the forward direction, and using the backward direction leaves o_{i-1} undetermined. In our method, two unidirectional LSTM models are trained in the training phase and combined in the test phase as follows.

- M_f, a model using a forward LSTM in Tag Decoder, is trained. In this model, o_{t-1} is added as an input of LSTM at time i, as shown in Figure 5.
- M_b, a model using a backward LSTM in Tag Decoder, is trained. In this model, o_{t+1} is added as an input of LSTM at time i.
- In order to recognize named entities from an unknown sentence, M_f and M_b are applied and the output vectors o_i^f and o_i^b are obtained. Then, the average of $o_i = (o_i^f + o_i^b)/2$ is calculated. The NE tag at time i is determined by the index of the highest score of o_i.

3.3 Separate embedding fine-tuning

In our method, the parameters of the word and character embeddings are fine-tuned, that is, they are updated through training of the NER model. However, since the number of parameters is increased by the fine-tuning, not only does this involve a high computational cost, but also the fine-tuned embedding parameters might not fit the NER task well. Our method, called *separate embedding fine-tuning*, tackles these problems. The basic idea is to train the *embedding parameters* and *model parameters* separately. The model parameter are the parameters except for the word and character embeddings, including LSTM in Character Encoder, BiLSTM in Word Encoder, part-of-speech embedding, Japanese particle embedding, LSTM in Tag Decoder, and FNN in Tag Decoder. Our separate embedding fine-tuning is carried out as follows.

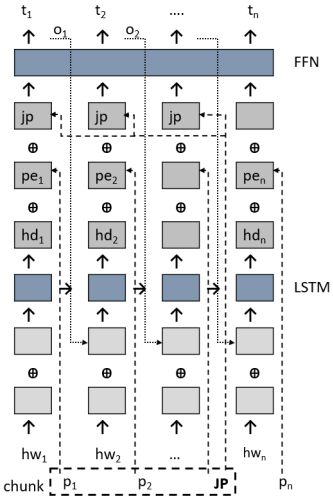


Figure 5: Tag Decoder

1. Word and character embeddings are pre-trained. They are used as initial embedding parameters.
2. The classification model is trained until the loss function is saturated. In this step, only the model parameters are estimated, while the embedding parameters are fixed.
3. The model is trained again, where only the embedding parameters are updated and the model parameters are fixed. the embedding parameters are fine-tuned until the loss function is saturated.

4 Evaluation

4.1 Experimental settings

The news articles in the NE tagged corpus were divided into about 90,000 sentences, then split into training, development and test data-sets. The proportions of the training, development and test data are 80%, 10% and 10% respectively. Each subset contains named entities of almost all 26 classes. Among 52,208 NE types in all the data-sets, 45,097(86%) NEs are ambiguous, i.e. they have two or more NE classes, while 7,111(14%) NEs have one NE class. Table 3 shows the statistics of the data-sets.

Table 2: Definition of models

	Character Encoder	Word Encoder	Tag Decoder	pre-trained	fine-tuning	vector size hidden	pos	jp
Shen's model	BiLSTM	BiLSTM	forward LSTM	no	–	128	–	–
TDf-small	LSTM	BiLSTM	forward LSTM	no	–	128	128	128
TDf	LSTM	BiLSTM	forward LSTM	yes	whole	256	64	64
TDb	LSTM	BiLSTM	backward LSTM	yes	whole	256	64	64
TDfb	LSTM	BiLSTM	forward&backward	yes	whole	256	64	64
TDfb-sep	LSTM	BiLSTM	forward&backward	yes	separate	256	64	64

Table 3: Dataset

Data	Sentence	Token	NE
Training	71,854	1,781,685	187,814
Development	8,980	222,403	23,278
Test	8,980	222,073	23,283

The models were trained on Google Colaboratory, which allows us to use one Tesla K80 GPU. With the development data, we use ADAM(Kingma and Ba, 2014) for the optimization of the hyper parameters in the following configuration:

1. Number of hidden units in LSTM/BiLSTM = 128 or 256
2. Optimizer ADAM learning $\alpha = [10^{-4}, 10^{-7}]$, $\beta_1 = 0.5$, $\beta_2 = 0.999$, $\varepsilon = 10^{-8}$

Table 2 summarizes methods compared in this experiment. First, we implemented almost the same model as that of (Shen et al., 2018) as the baseline. Next, TDf-small, our base model with a relatively small neural network, was trained for quick comparison with the baseline. A major difference between them is that inclusion of POS and Japanese particles is only applied by TDf-small.

The rest of the models use pre-trained character and word embeddings, as indicated in the column of "pre-trained" in Table 2. In these models, we increased the number of hidden units from 128 to 256, since this was able to improve the F1-score in our preliminary experiment. On the other hand, we reduced the dimension of the vector of POS (*pos*) and Japanese particles (*jp*) from 128 to 64, considering the additional computational cost for the fine-tuning of the word and character embeddings. TDf, TDb and TDfb use forward, backward and both LSTM in Tag Decoder respectively, where all parameters including word/character embeddings are trained simultaneously. Finally, TDfb-sep was trained with our separate embedding fine-tuning proposed in subsection 3.3.

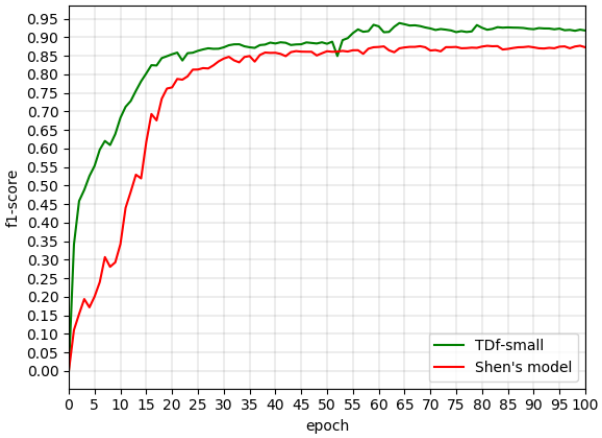


Figure 6: F1-score on the development data of Shen's and our model

The precision, recall and F1-score of each named entity class on the test data were measured in the experiment. However, for the comparison, a micro average of the F1-score for 26 NE classes is used as the major evaluation criterion.

4.2 Results and discussion

Figure 6 shows the F1-score on the development data of the baseline and our model. It shows that TDf-small clearly outperforms Shen's model, which is one of the state-of-the-art models. The number of the epochs for training the models was determined so that the F1-score becomes the highest on the development data, then the NER performance of two models in the test data was measured. The F1-score of TDF-small

Table 4: Performance of NER on the test data

Model	F1-score	statistical test
Shen's model	0.8839	–
TDf	0.9316	–
TDb	0.9302	$p < 0.05$ (vs TDf)
TDfb	0.9337	$p < 0.05$ (vs TDf)
TDfb-sep	0.9440	$p < 0.01$ (vs TDfb)

Table 5: F1-score comparison between TDf and TDb

NE class	TDf	TDb	dif.*
Product	0.942	0.946	−0.004
Person	0.957	0.958	−0.001
Timex	0.985	0.983	+0.002
Countx	0.940	0.929	+0.011
Organization	0.903	0.901	+0.002
Periodx	0.967	0.966	+0.001
Ordinal_Number	0.891	0.890	+0.001
Location	0.951	0.946	+0.005
Facility	0.865	0.858	+0.007
Event	0.884	0.877	+0.007
Age	0.987	0.985	+0.002
Percent	0.990	0.987	+0.003
Natural_Object	0.853	0.817	+0.036
Disease	0.947	0.900	+0.047
Multiplication	1.00	1.00	0
Rank	0.896	0.912	−0.016
Numex_Other	0.731	0.686	+0.045
Frequency	0.667	0.511	+0.156
Point	0.891	0.871	+0.020
Measurement	0.966	0.942	+0.024
Money	0.991	0.994	−0.003
School_Age	0.962	0.905	+0.057
Color	0.895	0.729	+0.166
God	0.571	0.333	+0.238
Name_Other	0.769	0.566	+0.203
Latitude_Longtitude	0.00	0.00	0

* dif. means TDf − TDb

was 0.9254, which was obviously better than Shen's model, namely 0.8839. These results prove that the inclusion of POS and Japanese particles is effective.

Next, we evaluated our proposed methods using forward, backward, and both LSTM in Tag Decoder. Table 4 presents the F1-score of our models on the test data, while Table 5 compares TDf and TDb for each NE class. The model with forward LSTM in Tag Decoder slightly outperforms the backward model in the overall F1-score in Table 4 and most of the NE classes in Table 5. However, TDb is better than TDf for some NE classes, namely, Product, Person, Money and Rank. When forward and backward LSTM are combined, the F1-score is slightly improved from the model using only forward LSTM. Although the difference is small, it is confirmed by McNemar's test that TDfb is significantly better than TDb at the 95% confidence level. From the above results, we can conclude that combining forward and backward LSTM in Tag Decoder is effective.

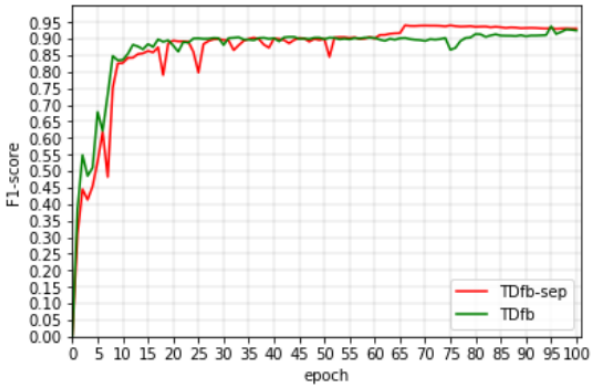


Figure 7: F1-score on the development data of TDfb and TDfb-sep

We evaluated the separate embedding fine-tuning by comparing TDfb and TDfb-sep. Figure 7 shows the change of F1-score on the development data of these two models. In the training of TDfb-sep, the parameters of the character and word embeddings were fixed until the 60^{th} epoch where the loss function saturates in the training. After the 61^{st} epoch, the parameters of the character and word embeddings were fine-tuned, while the model parameters were fixed. It can be seen in Figure 7 that the F1-score is sharply improved at the 61^{st} epoch. Furthermore, in the comparison on the test data in Table 4, the F1-score of TDfb-sep is significantly better than that of TDfb, at the 99% confidence level. These results indicate that the idea of updating the model parameters and embedding parameters separately is effective for training the neural based NER model.

Finally, the precision(P), recall(R) and F1-score(F) of our best model, TDfb-sep, for each NE class are shown in Table 6. The last column "NE" shows the number of named entities in the test data. Among the 26 NE classes, the F1-scores for 17 classes are higher than 90%, which are satisfying results for a practical NLP system. The F1-score is lower than 80% when the number of named entities in the test data (also in the training data) is small, as

Table 6: Performance of NER for each class

NE class	P	R	F	NE
Product	0.961	0.949	0.955	5580
Person	0.970	0.969	0.969	3021
Timex	0.976	0.990	0.983	2215
Countx	0.955	0.932	0.943	1333
Organization	0.940	0.910	0.925	2649
Periodx	0.968	0.978	0.973	495
Ordinal_Number	0.940	0.915	0.927	328
Location	0.971	0.939	0.955	3413
Facility	0.855	0.943	0.894	916
Event	0.908	0.885	0.897	766
Age	0.985	0.992	0.988	384
Percent	0.981	0.997	0.988	303
Natural_Object	0.931	0.773	0.845	436
Disease	0.900	0.960	0.929	150
Multiplication	1.00	1.00	1.00	13
Rank	0.974	0.912	0.942	205
Numex_Other	0.773	0.699	0.734	73
Frequency	0.476	0.769	0.588	13
Point	0.892	0.916	0.904	154
Measurement	0.937	0.966	0.951	292
Money	0.988	0.990	0.989	411
School_Age	0.916	0.974	0.944	78
Color	0.780	0.914	0.842	35
God	0.400	0.500	0.444	4
Name_Other	0.600	0.800	0.686	15
Latitude_Longtitude	0.00	0.00	0.00	0
micro average	0.944	0.944	0.944	23282
macro average	0.845	0.868	0.854	23282

is the case with Numex_Other, Frequency, God, and Name_Other. This might be caused by the insufficiency of the training data. However, the F1-scores of the named entities that are often regarded as important reach over or around 90%, such as Timex (98.3%), Person (96.9%), Product(95.5%), Location (95.5%), Organization (92.5%), and Event (89.7%).

5 Conclusion

This paper proposed the novel method of deep learning for Named Entity Recognition in Japanese. Our model consisted of three neural network modules: Character Encoder, Word Encoder and Tag Decoder. In addition to word embedding and character embedding, two important features were added. One was the part-of-speech that is widely used in various NLP tasks, the other was the Japanese particles, which play a significant grammatical role in Japanese.

The first contribution of this paper was to combine forward and backward LSTM in Tag Decoder. The information of both left and right contexts was thought to be necessary for an accurate NER. However, since the NER tag of one or the other of either the previous or succeeding words is inapplicable in a sequential labeling model, BiLSTM could not be simply applied. To use both the previous and succeeding NE tags for classification, models using forward LSTM and backward LSTM in Tag Decoder were separately trained, then, in the test phase, the NE tag of each word was determined by the sum of the probability distributions of the NE tags of the two models. Our model using both directions of LSTM slightly outperformed the models with forward or backward LSTM in our experiment.

The second contribution was to propose a method of fine-tuning of the word and character embeddings. Although fine-tuning an embedding has been a promising approach to improve the performance of deep learning models for NLP, it increases the number of parameters considerably. In our approach, the model parameters were first trained with pre-trained and fixed word and character embeddings, then the parameters of the word and character embeddings were fine-tuned with the fixed model parameters. This method was able to improve the F1-score by 0.01 point in our experiment. Furthermore, our best model was obviously better than the baseline, and achieved an F1-score of 0.944.

In the future, we will explore other effective features to be added to the neural network in order to improve the model performance. We will also investigate various adjustment methods for the hyper parameter estimation. Another important line of future research is to investigated whether the model with two directional tag decoder is effective for NER of other languages such as English. We plan to evaluate our model on CoNLL-2003 English dataset.

Acknowledgements

This research is financially supported by the National Science and Technology Development Agency (NSTDA), National Research University Project, Thailand Office of the Higher Education Commission, Japan Advanced Institute of Technology (JAIST), and Infrastructure Engineering Research Unit, Sirindhorn International Institute of Technology (SIIT), Thammasat University (TU).

References

Raghavendra Chalapathy, Ehsan Zare Borzeshi, and Massimo Piccardi. 2016. Bidirectional LSTM-CRF for clinical concept extraction. *ClinicalNLP 2016*, page 7.

Jason P. C. Chiu and Eric Nichols. 2016. Named entity recognition with bidirectional LSTM-CNNs. *Transactions of the Association for Computational Linguistics*, 4:357–370.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *Proceedings of the Empiricial Methods in Natural Language Processing*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. 2011. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12(Aug):2493–2537.

Chuanhai Dong, Jiajun Zhang, Chengqing Zong, Masanori Hattori, and Hui Di. 2016. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition. In *NLPCC/ICCPOL*.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks*, 18(5–6):602–610.

Taiichi Hasimoto, Takashi Inui, and Koji Murakami. 2008. Constructing extended named entity annotated corpora. *Technical Report*, 008-NL-188:113–120. Written in Japanese.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

John J. Hopfield. 1982. Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences of the USA*, 79(8):2554–2558.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional LSTM-CRF models for sequence tagging. *CoRR*, abs/1508.01991.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander M Rush. 2016. Character-aware neural language models. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations*, volume abs/1412.6980.

Onur Kuru, Ozan Arkan Can, and Deniz Yuret. 2016. Charner: Character-level named entity recognition. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 911–921.

John D. Lafferty, Andrew McCallum, and Fernando C. N. Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning*, pages 282–289, San Francisco, CA, USA. Morgan Kaufmann Publishers.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. In *The 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 260–270.

Xuezhe Ma and Eduard Hovy. 2016. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 1064—-1074, Berlin, Germany.

Grégoire Mesnil, Xiaodong He, Li Deng, and Yoshua Bengio. 2013. Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *INTERSPEECH*, pages 3771–3775.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119.

Shotaro Misawa, Motoki Taniguchi, Yasuhide Miura, and Tomoko Ohkuma. 2017. Character-based bidirectional LSTM-CRF with words and characters for Japanese named entity recognition. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 97–102.

Thien Huu Nguyen, Avirup Sil, Georgiana Dinu, and Radu Florian. 2016. Toward mention detection robustness with recurrent neural networks. In *Proceedings of IJCAI Workshop on Deep Learning for Artificial Intelligence*, volume abs/1602.07749, New York, USA.

Hoang Pham and Le-Hong Phuong. 2017. End-to-end recurrent neural network models for Vietnamese named entity recognition: Word-level vs. character-level. In *The 15th International Conference of the Pacific Association for Computational Linguistics*.

Satoshi Sekine and Chikashi Nobata. 2004. Definition, dictionaries and tagger for extended named entity hierarchy. In *Proceedings in the 4th International Conference on Language Resources and Evaluation*, pages 1977–1980.

Yan Shao, Christian Hardmeier, and Joakim Nivre. 2016. Multilingual named entity recognition using hybrid neural networks. In *The Sixth Swedish Language Technology Conference*.

Yanyao Shen, Hyokun Yun, Zachary C. Lipton, Yakov Kronrod, and Animashree Anandkumar. 2018. Deep active learning for named entity recognition. In *International Conference on Learning Representations*.

Yuji Matsumoto Taku Kudo. 2002. Japanese dependency analysis using cascaded chunking. In *Proceedings of the 6th Conference on Natural Language Learning*, pages 63–69.

Erik F Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL*, volume 4, pages 142–147. Association for Computational Linguistics.

Kai Xu, Zhanfan Zhou, Tianyong Hao, and Wenyin Liu. 2018. A bidirectional LSTM and conditional random fields approach to medical named entity recognition. In *International Conference on Advanced Intelligent Systems and Informatics*, pages 355–365.

Zhilin Yang, Ruslan Salakhutdinov, and William W. Cohen. 2016. Multi-task cross-lingual sequence tagging from scratch. *CoRR*, abs/1603.06270.

Making Metaphors: A Quantitative Analysis of Metaphor Production and Interpretation in Japanese Using a Multimodal Task

Brian J. Birdsell
Hirosaki University
brian@hirosaki-u.ac.jp

Natsuko Tatsuta
Hirosaki University
tatsuta@hirosaki-u.ac.jp

Hiroaki Nakamura
Hirosaki University
hiroaki@hirosaki-u.ac.jp

Abstract

Two key features of Conceptual Metaphor Theory are that metaphors appear in multiple modes of communication from language to gestures to pictures and that metaphors scaffold our understanding of abstract concepts by grounding them in embodied, physically experienced concepts. In an exploratory study, we investigated metaphor production and interpretation using cross modal stimuli (verbal and pictorial). Native Japanese participants viewed an abstract word in the textual mode, in the form of an incomplete copula metaphor (Friendship is ...), and then saw six images of concrete entities (castle, heater, colored pencils, etc.). They chose one of these image concepts to complete the copula metaphor and then provided an interpretation of it. In this paper, we first analyze these choice selections using descriptive statistics. Results indicate that there is a wide amount of variability among these selected responses. Secondly, we analyze the interpretations, which use (1) external or systemic properties of the pictorial entity; (2) situational functions or actions performed on the entity; or (3) some form of introspection related to the entity.

1 Introduction

Metaphor describes one thing (often referred to as the topic) in terms of another (the source or vehicle), as in "Hope is the thing with feathers" (from a poem by Emily Dickinson). In this case, "hope" is the topic and in this sentence it's correlated to the source, "the thing with feathers", which one might presume to be a bird. This is considered a nominal metaphor (e.g., X <copula> Y), consisting of two parts, the topic, which tends to be more abstract, and the source, which tends to be a physical or concrete entity. This combinatorial ability is paramount for communication and is widespread in everyday language use, not just in poetry. Moreover, metaphor is not only used for verbal communication, but also nonverbal communication through gestures (Cienki and Müller, 2008), pictures (Forceville and Urios-Aparisi, 2009) and music (Zbikowski, 2008). Under this perspective, metaphor is viewed as being only derivatively part of language and in fact, conceptual in nature (commonly referred to as, Conceptual Metaphor Theory; see Lakoff, 1987; Lakoff and Turner, 1980).

In this paper, we first provide some theoretical background on the differing processes of concrete and abstract concepts, focusing initially on Dual Coding Model and then on embodied theories of cognition. Specifically, our review addresses abstract concepts and the role of metaphor, affect, and language. Finally, we argue for an approach to abstract metaphor construction as a dynamic process that highlights the fluidity, flexibility, and variability of concepts. Then we discuss a unique exploratory study that utilized both verbal and pictorial stimuli for a metaphor production and interpretation task. Moreover, we investigate the chosen sources by the participants to complete the

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 334-344, Hakodate, Japan, September 13-15, 2019

metaphors and what semantic features they used to interpret them. This study thus sheds light on the tight combinatorial and entrenched networks between some concepts, as well as the ad hoc process to interpret newly formed and unfamiliar combinations.

2 Constructing Meaning: The Dual Coding Model

The Dual Coding Model (Paivio, 1971, 2007) claims that word meanings are represented in two different systems – one for nonverbal codes called "imagens" and the other for verbal codes called "logogens". The nonverbal system relies on multimodal representations (tactile, visual, olfactory) of the concepts while the verbal system is based on a linguistic system of knowledge for the concept and associative networks (curiosity – wandering). According to this theory, concrete concepts recruit equally from both imagery and verbal processes, but for abstract language, verbal processes predominate (Paivio, 2007). Therefore, concrete concepts have an advantage since they receive dual-content, which has been called a *concreteness effect,* or the effect that we tend to process concrete words faster and remember them better than abstract words. This model highlighted the importance of sensorimotor systems for processing concrete concepts, but on the other hand, claimed that abstract concepts are purely part of the verbal system. This model foreshadowed a growing movement within the cognitive sciences towards a greater awareness of the body for meaning construction.

3 Embodied Theories of Meaning Construction

Over the past couple decades, the field of cognitive science has gone through a major paradigm shift from the traditional computational theory of mental processes, as algorithmic operations on abstract symbols, to one where concepts, objects, or events are grounded in sensorimotor, perceptual and emotional systems, commonly referred to as embodied cognition (Barsalou, 2008; Gibbs, 2005; Pecher and Zwaan, 2005). As opposed to earlier amodal or disembodied perspectives, an embodied perspective views meaning to be grounded in knowledge of action and objects. In an early theoretical approach to embodiment, Barsalou (1999, 2017) proposed that when processing a word like *guitar*, aggregated information from perception, action, and internal states are recruited, which results in a simulation of that object. This may include the shape, texture, sound, how it's played, and past interactions with it. These simulations or reactivated perceptual input provide linguistic meaning to the concept. Barsalou (1999) labeled this theory as perceptual symbol systems.

Studies using fMRI provided evidence for embodiment by showing that when one reads action verbs (e.g., *lick, pick, or kick*) corresponding to certain body parts (e.g., *face, arms, legs* respectively), adjacent or overlapping areas were activated in the motor and premotor cortex for that action (Hauk, Johnsrude, and Pulvermüller. 2004). In another study, Buccino et al. (2005) also found similar results. In their study, they found that when participants listened to verbal language corresponding to a body part (foot/leg; hand/arm), this modulated the activity of the motor system for the effector involved in that listening activity. Such results suggest that action verbs are coded into the same premotor and motor cortices used when one actually performs the action (for a review of the role of the motor system in language comprehension see Fischer and Zwaan, 2008). Evidence goes beyond action words to include the sensory modes. For instance, odor-related words (e.g., cinnamon) have been shown to elicit increased activity in olfactory regions compared to neutral words (Gonzalez et al., 2006). In sum, there is a considerable amount of accumulating evidence that supports an embodied view for language comprehension, especially in regards to concrete concepts.

However, a common criticism of embodied theories is that they have difficulty in explaining abstract concepts (since they are commonly viewed as not deriving from sensorimotor or perceptual content) and the creation of novel concepts that also lack an experiential basis (Borghi et al., 2017; Dove, 2011). In the next section, we review two possible explanations for abstract concepts: (1) they are grounded in metaphors, which has been called a strong version of embodiment (Meteyard et al., 2012), and (2) abstract concepts are both linguistic and embodied (Dove, 2009, 2011; Vigliocco et al., 2009) and are grounded in emotional and interoceptive states (Kousta et al.,

2011), as well as events and situations (Barsalou and Wiemer-Hastings, 2005).

3.1 A Strong Version: Metaphor as a Bridge

A strong version of embodied cognition argues that not only concrete concepts rely on sensorimotor simulation, but abstracts concepts, too (Gallese and Lakoff, 2005, Gibbs, 2006) and metaphors act as a "bridge" between embodiment and abstraction (Jamrozik et al., 2016).

Grady (1997) theorized that some associations between topic and source are so deeply entrenched, common, and found across multiple languages that he called them primary metaphors (e.g., DIFFICULTY IS HEAVY; POWER IS UP; INTIMACY IS CLOSENESS). So, the physical concepts, which are experienced through the sensorimotor systems (i.e., heaviness), map in a unitary direction onto non-sensory abstract concepts (i.e., difficulty). Lakoff and Johnson (1999) suggest that these types of metaphor are obligatorily learned during cognitive development. For instance, for AFFECTION IS WARMTH, the child conflates the affection of the caregiver with the sensation of bodily warmth. Thus, affection, as an abstract psychological feeling, becomes fused with the sensory bodily experience of warmth. These experientially motivated primary metaphors are widespread and have been shown to take place at the conceptual level. For instance, a situation of difficulty may include some type of burden like harboring a secret. Slepian et al. (2012) found that when people recalled or suppressed an important secret like infidelity, they estimated hills to be steeper and distances to be farther. The researchers interpreted this as evidence showing that harboring a secret physically weighs people down and thus influences their perception.

This co-activation of the topic-source concepts is automatic and unconscious. Accordingly, when one thinks of a difficult situation or an intimate relationship, one also conceptualizes a physical weight or spatial closeness respectively. Yet these sources may map onto multiple and varying abstract concepts. For instance, weight is also mapped onto importance due to repeated experiences with heavy objects, which require more effort, in terms of physical strength or cognitive planning, as compared with lighter objects (Jostmann, Lakens, and Shubert, 2009).

There is also empirical support in the field of neuroscience for the metaphor as a "bridge" view. For instance, in an fMRI study that specifically examined the grounding of metaphor in sensorimotor systems found that textural metaphors (e.g., *she had a rough day*) activated texture-selective somatosensory cortex, compared to literal matched sentences (e.g., *she had a bad day*) (Lacey, Stilla, and Sathian, 2012). Despite this widespread support for the metaphor view for explaining abstract word meaning comprehension (e.g., Jamrozik et al., 2016), its explanatory power is limited to abstract concepts that have clear sources (e.g., bad – rough; burden – weight; intimacy – closeness), which are numerous, but not exhaustive of all abstract concepts.

3.2 Abstract Knowledge Grounded in Affect, Events, and Situations

In a weaker version of embodied cognition, Vigliocco et al. (2009) proposed that experiential (sensorimotor and affective), as well as linguistic (verbal, associative networks) information contribute to the representation of all concepts. In this view, concrete word meanings have preponderance for experiential information, but abstract ones for affective and linguistic information. This emphasizes the importance of emotion for abstract meaning construction and how abstract words also have an experiential basis, but also how emotion allows for "learning, or bootstrapping, of abstract knowledge" (Meteyard et al., 2012 p. 800).

Another approach that falls under a weaker version of embodied cognition proposes that abstract concepts are fundamentally different from concrete ones for they simulate concrete situations and introspective experiences (Barsalou and Wiemer-Hastings, 2005). More specifically concrete concepts have a focus object (e.g., guitar), but abstract ones are more diffuse, are used in a wider variety of contexts, and are more complex. Developing this model, Barsalou et al. (2008) proposed what they called LASS or Language And Situated Simulation framework, which claims that linguistic and situated simulations are continuously interacting with each other. The linguistic is involved with more superficial processing of word meanings whereas deeper processing involves sensorimotor simulations.

In sum, abstract concepts are highly heterogeneous. Some abstract concepts may recruit entrenched metaphorical mappings onto concrete concepts while others have high affective associations and still others activate a scene that simulates one to mentally run a situation.

4 Dynamic Concepts: Flexibility, Fluidity, and Variability

Another important issue for metaphor production research is the flexibility of concepts. There is growing evidence suggesting that the architecture of the semantic system is experientially based (or at least partially for abstract concepts, as presented in the previous section) and moreover variable across timescales and contexts, as well as individual processing preferences and abilities (Yee and Thompson-Schill, 2016). That is to say, concepts don't necessarily have conceptual cores nor are they static. Variability is far more common than often assumed and even entrenched features of concepts are not always automatically activated (Lebois, Wilson-Mendenhall, and Barsalou 2015). In addition, Yee et al. (2016) argue that there is never a "no context" or "neutral context" situation for even the goal of the task can influence conceptual activation. This approach aims to move away from seeing *concepts* in the head, as static objects, to an an approach that emphasizes the dynamic process of making meaning by way of *conceptualizing*, as an active process (Casasanto and Lupyan, 2015).

Reconstructing the meaning of words is highly modulated by the individual's personal experiences with such words. This is especially the case with abstract nouns since they span a wide range of contexts and thus subjective experiences are crucial for their representations (Wiemer-Hastings and Xu, 2005). Some researchers even contend based on the dynamic influences of numerous variables such as the body, the environment, past experiences, and relevant goals that you can never represent the same concept twice (Connell and Lynott, 2014).

5 Overview of the Study

In this current study, we aim to investigate whether or not some abstract concepts (LOVE, ANGER), are more commonly combined with some physical and concrete source concepts and if this shows up across a large group of individuals. The other possibility is that topic/source combinations vary widely when individuals are confronted with a decision task like a metaphor completion task. For instance, when asked to think about the concept of LOVE, one likely activates long-term memory and recent experiences with this abstract concept. This may include people, such as family (house) or the performance of lovers (masks). For others, this abstract concept might simply evoke an explosive emotion. Still for others, they might consider LOVE as requiring time to grow (forest) or in contrast leading to entrapment (spider web). The point here is that when primed to complete a metaphor for this abstract concept from a set of six corresponding possible sources, which are presented in the pictorial mode, these concurrent contextual cues from the images activate certain features of LOVE. If the two concepts were indeed fused by an entrenched metaphorical mapping, one would assume based on the participants desire to reduce the cognitive load of the task, this entrenched source would be frequently selected to complete the metaphor.

5.1 Research Questions

1. Do subjects show any preference for certain source stimuli to complete the metaphors?
2. What selected features or properties of the selected source concepts do the subjects use to interpret the metaphors?

6 Method

6.1 Participants

97 native Japanese speakers took part in this study (36 female, M_{age} = 20, SD_{age} = 1.07). Participants were recruited from 1st year Liberal Arts courses at the university.

6.2 Material

The material developed for this study consisted of two parts: a list of abstract concepts (metaphor topics) and six corresponding images (potential sources to complete the metaphor). In total, there were 20 abstract concepts used in this study, collated from a list of common and familiar abstract nouns, as well as previous research that have similarly used abstract topics as a prompt for a copula metaphor production task (see Shibata et

al., 2007; Terai et al., 2015) (e.g., 好奇心, kōkishin 'CURIOSITY'; see Appendix A for the full list),. Each abstract noun (e.g., 友情 yūjō 'FRIENDSHIP') had a unique set of six corresponding images that represented concrete physical entities (e.g., space heater, castle, colored pencils, etc.). These images were selected as having potential semantic features that could be mapped onto the topic (e.g., warmth, protection, variety). In addition, we had conducted a previous study with these images where we interviewed participants afterwards about the images and subsequently removed ones, they had deemed confusing or difficult to understand. One strategy for compiling these sets of images for each abstract topic was to do a search for the abstract word on Google and then look at images that appeared in this list. We also aimed at using pictures that were simplistic and had only one primary entity visible in the image. The material was inputted into a Google Form.

6.3 Procedure

Students individually sat at computers in a language lab. First, they signed a consent form that explained the purpose of the study and then opened up the Google Form and saw an example that explained the steps to complete the metaphor task. Specifically, the participant first saw an abstract topic concept in the format of an incomplete metaphor in the textual mode (愛情・・・だ *aijō* ... da 'LOVE IS ...') and then six corresponding images (see Figure 1). Beneath the images, there was an input box where they completed the metaphor by choosing one of the images as the source and then there was a second input box where they provided an interpretation of it. They did this for each of the 20 abstract concept/image sets.

Figure 1: The 6 images to complete the metaphor for 愛情 *aijō* 'LOVE'

7 Results and Discussion

The first step in analyzing the data involved a descriptive analysis for the source selections for each of the topics (see Table 1). The distribution of selected sources across all six images highlights the flexibility of conceptual combination and the possibility of fusing an abstract concept to numerous physical entities. For instance, one would presume under a conceptual metaphor model that FRIENDSHIP (as an abstract concept closely related to intimacy) would activate the physical sensation of "warmth", which is a salient feature of a heater, but this was one of the lower selected sources by the respondents.

TOPIC	% of Selected SOURCES (Images)		
FRIENDSHIP 友情 *yūjō*	*Colored Pencils* 26%	*Medicine* 22%	*Battery* 16%
	Shoes 15%	*Heater* 8%	*Castle* 7%
LOVE 愛情 *aijō*	*Masks* 22%	*Forest* 10%	*House* 20%
	Spider Web 16%	*Explosion* 24%	*Suitcase* 3%
ANXIETY 不安 *fuan*	*Storm* 33%	*Tangled knot* 32%	*Crutches* 12%
	Cliff 11%	*Hotpot* 8%	*Blender* 0%
CURIOSITY 好奇心 *kōkishin*	*Dandelion* 29%	*Lighter* 24 %	*Map* 21%
	Paintbrush 11%	*Grass* 6 %	*Forest Path* 2 %
EDUCATION 教育 *kyōiku*	*Opened Door* 26%	*Key* 29%	*Globe* 8%
	Construction 14%	*Hammer* 2%	*Dumbbell* 18%

Table 1: Five TOPICS and corresponding selected SOURCES (see Appendix A for the full list)

This variability was widespread and occurred in 16 of the 20 topics where no one source accounted for more than 40% of the responses. So, in answer to our first research question, only 20% of the topics appear to elicit a preferential source. These topics likely have strong associative connections to these concrete concepts (MIND – *sponge*), as they appeared across a large population of participants.

TOPIC	% of Selected SOURCES (Images)		
HOPE 希望 *kibō*	*Rainbow* 55%	*Birds* 15%	*Stoplights* 12%
	Stethoscope 5%	*Merry-go-round* 3%	*Leaf* 3%

HONESTY 正直 *shōjiki*	*Sword* 50%	*Straight Road* 20%	*Magnifying Glass* 10%
	Handshake 8%	*Mountaintop* 4%	*Watering Can* 0%
MIND 心 *kokoro*	*Sponge* 46%	*Space* 22%	*Flowering Vase* 14%
	Window 10%	*Computer* 2%	*Rug* 2%
CULTURE 文化 *bunka*	*Tree (roots)* 43%	*Salad* 14 %	*Handcuffs* 14%
	Fence 11%	*Computer Network* 6%	*Pillars* 5%

Table 2: TOPICS with high percentages of single SOURCE selections

Despite one source being heavily weighed in the above four topics, there was still rich diversity among the responses, which shows the idiosyncrasies and emotional valence of these abstract concepts. For instance, CULTURE was widely associated with a *tree*, which has positive meaning, but in contrast, 14% of the respondents chose a source with negative meaning (*handcuffs*).

The second goal of this research was to investigate the semantic features used by the participants to interpret their newly constructed metaphors. In order to do this, we used the coding scheme developed by Wu and Barsalou (2009). Table 2 shows the analysis for the topic, FRIENDSHIP. This topic constrained and provided context to these six images, which forced the participants to look for features that could be used to provide some semantic structure to their newly created metaphors. Through this analysis, we can observe the wide range of conceptual content that is projected from the sources onto this abstract concept. For instance, those who selected *colored pencils* tended to describe FRIENDSHIP as being diverse. This property comes from the systemic property of colored pencils, as in the fact, that these pencils come in many different colors. In contrast, others provided an introspective property of this entity by describing how *colored pencils* enable one to live a more enriching life. That is to say, these *colored pencils* provide one the tools to live a more meaningful life, which likely refers to their intrinsic property for making art. Another commonly chosen source was *medicine*, whereby the participants focused on the healing power of FRIENDSHIP. Moreover, those who chose *battery* focused on how relationships between friends occasionally need to be recharged or can even become dangerous if they are overused. Another salient systemic property of batteries is that they alter between being fully charged and weakly charged and again participants used this feature to map onto the topic and how friendships similarly oscillate in strength and weakness. Unexpectedly, 15% of the participants selected *shoes* to complete the metaphor and the interpretations highlighted many different properties of *shoes* ranging from the physical attribute of *shoes*, such as they come in pairs (the image of two) to relational meaning, as in, *shoes* are used to travel through life with, similar to a FRIENDSHIP. Surprisingly only a small percent chose a *heater* and as expected they interpreted this metaphor by describing how FRIENDSHIP makes one feel warm. Finally, for *castle*, which was least selected source image, participants tended to focus on the systemic property of *castles*, as in being strong and unbreakable. In contrast, other respondents focused on the situation of people working together to build a *castle*, which relates to how FRIENDSHIP is a building process.

TOPIC: FRIENDSHIP IS …	
Image SOURCES	Coding Scheme of Interpretations
1. *Colored Pencils* (26%)	**E_{sys}** variation /diversity **S_F** draw colorful pictures **I_c** colorful pictures *enables* one to live a more enriching life
2. *Medicine* (22%)	**S_F** helps one overcome pain; gives one energy **S_F** possibility of addiction; makes one crazy; possibility of overdoes
3. *Battery* (16%)	**E_{sys}** sometimes weak sometimes strong **I_c** relationship *requires* occasional recharging **I_c** dangerous *if* overused
4. *Shoes* (15%)	**E_{CE}** they come in pairs; not alone **S_F** walk together in life with
5. *Heater* (8%)	**E_{sys}** warmth
6. *Castle* (7%)	**E_{sys}** strength, unbreakable **S_A** takes time to build up, but after it's built one knows all the details **I_R** fun, a remarkable symbol

Table 3: Image SOURCES and interpretations (see Appendix B for the coding scheme)

A second example, CURIOSITY (see Table 4), also illustrates the great diversity of representation of this abstract concept based on the selected sources. For instance, those who selected a *dandelion* considered the situation and how the wind or an agent (such as a friend) performs an action upon this object by blowing it, causing the seeds to spread and eventually fall to the ground and germinate, which results in the budding of a new flower. So, using this source, the participants viewed curiosity as something that takes a long temporal timeframe, as compared to those who chose a *lighter*. Using this source, the participants saw the temporariness of CURIOSITY, as something short-lived like a flickering flame. Moreover, a number of the respondents who selected *lighter* as the source also included the potential danger of this entity (i.e., "curiosity killed the cat"). This did not show up in any of the other selected sources. Also, in contrast to the before mentioned examples, those who selected a *map* focused on the situational function of this object, in that it acts as a guide for one to move through unknown territories. The fourth most commonly selected image was the *paintbrush*. Again, some participants focused on the perceptual property of this entity and viewed CURIOSITY as being colorful. Others recalled some situation of a *paintbrush* and the freedom it provides one to paint as one wants to paint or a more specific situation of one throwing the paints onto the ground in search of a color, which highlights the exploratory nature of CURIOSITY. A few selected the *grass* picture, as the image to complete the metaphor and primarily interpreted this with systemic properties of *grass*. Interestingly of all the entity properties of *grass*, they tended to focus on its strength to grow anywhere, even in difficult places, and its persistence and expansiveness. These interpretations focus on the grittiness of CURIOSITY and how it often requires one to overcome difficulties, which might include resistance from others or society, and its need for commitment and endurance. Again, this contrasts considerably with a *lighter*, which associated CURIOSITY with a certain amount of fleetingness. Finally, a couple participants chose the image source, a *forest path*, and one interpretation pointed out introspectively how CURIOSITY moves you along a path into the future. This closely relates to the conceptual metaphor, LIFE IS A JOURNEY, and CURIOSITY is the trigger that helps you to move forward.

TOPIC: CURIOSITY IS …	
Image SOURCES	Coding Scheme of Interpretations
1. *Dandelion* (29%)	**E_{CE}** seeds > **S_A** a friend (or wind) blows the seeds > **E_B** seeds fall and take root (and flower)
2. *Lighter* (24%)	**E_{sys}** fleetingness **S_F** potentially dangerous
3. *Map* (21%)	**S_F** takes you to places; guides you in the unknown **I_O** maps expand your mind
4. *Paintbrush* (11%)	**E_{CE}** colorful **S_A** freedom to paint; dumping the colors onto the floor **I_R** exciting
5. *Grass* (6%)	**E_{sys}** grows anywhere; expansive; persistent; strong; casually grows; grows in difficult places **S_A** searching for a bug in a grassy place (an unforgettable experience)
6. *Forest Path* (2%)	**E_{sys}** continuous **I_O** moves you along the path towards the future

Table 4: Image SOURCES and interpretations

In summary, to answer our second research question, participants used different techniques to interpret the metaphors. For instance, sometimes they focused on the properties of the entity. This may include external components such as the fact that dandelions have seeds or a systemic property such as the flame of a lighter is a momentary flicker. Another strategy commonly used is to focus on some action property and this often involves simulating some situation related to the entity (as in, a map guides you through the unknown). Finally, some participants used an introspective property of the entity to interpret the metaphor. In this case, they have learned that maps are used for exploring new areas and thus provide a tool for expanding one's knowledge structures and it is then this feature that is mapped onto the topic, CURIOSITY. At the same time, certain salient properties of the entity were not used by the participants like a *map* is made of paper or that a *lighter* is commonly used for cooking or lighting a cigarette or that a *dandelion* is a plant.

8 Conclusion

In this paper, we describe an exploratory study that investigated metaphor production across two different modes (textual, pictorial). Providing the participants with a visual concrete image for the source activated a nonverbal, multimodal code, independent of linguistic content. The participants had to associate an abstract topic with one of these images. The data presented here shows a wide range of selected sources for a majority of the topics, which highlights the flexibility of thought, as well as the looseness of abstract conceptual content. One can view CULTURE as a *tree*, which grows and provides a group of people the roots or structure to live. At the same, another views CULTURE as confining and restrictive and thus associates it with *handcuffs* or a *fence*. This study highlights that when one is presented with an abstract concept and then six images, one does not retrieve static conceptual features, but conceptualizes, which is an active process of meaning making. This is likely dependent on a number of variables, ranging form past knowledge structures to emotions to salient thoughts that exact moment when the subjects were completing the task.

Another important point that this paper raises has to do with automatic activation of conceptual metaphors. To return to one of our examples, FRIENDSHIP, what was most revealing to us is the fact that few participants (8%) selected the heater. Those who did select this source mapped, as expected, the most salient property of a heater, warmth, onto the abstract concept, FRIENDSHIP. For us, we had predicted this source selection would be the strongest since it touches on a very conventional conceptual metaphor, yet this was not the case. This questions whether or not the conceptual structure of such an entrenched metaphor is indeed obligatorily accessed, especially when it crosses over into a different mode (pictorial).

As abstract concepts have been notably difficult to explain from an embodied perspective, future research needs to continue to look at this dynamic process. One area that could be investigated in future studies is to look at how stable these generated metaphors are at the individual level. Would participants' metaphors change over time? This would involve a longitudinal study that would ask participants to complete the metaphor tasks in this study twice, separated by a specific length of time, and then compare their response selections. Since abstract concepts are a key element to human societies and the cognitive architecture of humans, it is highly relevant to explore research methods that provide greater insight into our understanding of them.

Acknowledgments

This study was supported by JSPS KAKENHI Grant Number JP19K00566.
Images used in this study came from vectorstock.com.

References

Barsalou, L. W. (1999). Perceptual symbol systems. *Behavioral and Brain Sciences, 22*, 577–660.

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645.

Barsalou, L. W. (2017). Cognitively plausible theories of concept composition. In *Compositionality and concepts in linguistics and psychology* (pp. 9–30). Springer, Cham.

Barsalou, L. W., and Wiemer-Hastings, K. (2005). *Situating abstract concepts*. In D. Pecher & R. A. Zwaan, (pp. 129–163). New York: Cambridge University Press.

Barsalou, L. W., Santos, A., Simmons, W. K., and Wilson, C. D. (2008). Language and simulation in conceptual processing. In M. De Vega, A. M. Glenberg, and A. C. Graesser (Eds.), *Symbols, embodiment, and meaning* (pp. 245–283). Oxford: Oxford University Press.

Borghi, A. M., Binkofski, F., Castelfranchi, C., Cimatti, F., Scorolli, C., & Tummolini, L. (2017). The challenge of abstract concepts. *Psychological Bulletin, 143*(3), 263–292.

Buccino, G., Riggio, L., Melli, G., Binkofski, F., Gallese, V., and Rizzolatti, G. (2005). Listening to action-related sentences modulates the activity of the motor system: a combined TMS and behavioral study. *Cognitive Brain Research, 24*(3), 355–363.

Casasanto, D., and Lupyan, G. (2015). All concepts are ad hoc concepts. In E. Margolis and S. Laurence (Eds.), *The conceptual mind: New directions in the study of concepts* (pp. 543–566). Cambridge: MIT Press.

Cienki, A., and Müller, C. (Eds.). (2008). *Metaphor and gesture* (Vol. 3). John Benjamins Publishing.

Connell, L., and Lynott, D. (2014). Principles of representation: Why you can't represent the same concept twice. *Topics in Cognitive Science, 6*(3), 390–406.

Dove, G. (2009). Beyond perceptual symbols: A call for representational pluralism. *Cognition, 110(*3), 412–431.

Dove, G. (2011). On the need for embodied and dis-embodied cognition. *Frontiers in Psychology, 1*, 242, 1–13.

Fischer, M. H., and Zwaan, R. A. (2008). Embodied language: A review of the role of the motor system in language comprehension. *The Quarterly Journal of Experimental Psychology, 61*(6), 825–850.

Forceville, C., and Urios-Aparisi, E. (Eds.). (2009). *Multimodal metaphor* (Vol. 11). Walter de Gruyter.

Gallese, V., and Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in conceptual knowledge. *Cognitive Neuropsychology, 22*(3-4), 455–479.

Gibbs Jr, R. W. (2005). *Embodiment and cognitive science*. Cambridge University Press.

Gibbs Jr, R. W. (2006). Metaphor interpretation as embodied simulation. *Mind and Language, 21*(3), 434–458.

González, J., Barros-Loscertales, A., Pulvermüller, F., Meseguer, V., Sanjuán, A., Belloch, V., and Ávila, C. (2006). Reading cinnamon activates olfactory brain regions. *Neuroimage, 32*(2), 906–912.

Grady, J. (1997). *Foundations of meaning: Primary Metaphors and Primary Scenes*. Unpublished doctoral dissertation, University of California, Berkeley.

Hauk, O., Johnsrude, I., and Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron, 41*(2), 301–30

Jamrozik, A., McQuire, M., Cardillo, E. R., and Chatterjee, A. (2016). Metaphor: Bridging embodiment to abstraction. *Psychonomic Bulletin and Review, 23*(4), 1080–1089.

Jostmann, N. B., Lakens, D., and Schubert, T. W. (2009). Weight as an embodiment of importance. *Psychological Sscience, 20*(9), 1169–1174.

Kousta, S. T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: why emotion matters. *Journal of Experimental Psychology: General, 140*(1), 14–34.

Lacey, S., Stilla, R., and Sathian, K. (2012). Metaphorically feeling: comprehending textural metaphors activates somatosensory cortex. *Brain and Language, 120*(3), 416–421.

Lakoff, G. (1987). *Women, fire and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.

Lakoff, G., and Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press, Chicago.

Lakoff, G., and Johnson, M. (1999). *Philosophy in the flesh* (Vol. 4). New York: Basic books.

Lebois, L. A., Wilson-Mendenhall, C. D., and Barsalou, L. W. (2015). Are automatic conceptual cores the gold standard of semantic processing? The context-dependence of spatial meaning in grounded congruency effects. *Cognitive Science, 39*(8), 1764–1801.

Meteyard, L., Cuadrado, S. R., Bahrami, B., and Vigliocco, G. (2012). Coming of age: A review of embodiment and the neuroscience of semantics. *Cortex, 48*(7), 788–804.

Paivio, A. (1971). *Imagery and verbal processes*. New York: Holt, Reinhart and Winston.

Paivio, A. (2007). *Mind and its evolution: A dual-coding theoretical approach*. Mahwah, NJ: Lawrence Erlbaum.

Pecher, D., and Zwaan, R. A. (Eds.). (2005). *Grounding cognition: The role of perception and action in memory, language, and thinking.* Cambridge University Press.

Shibata, M., Abe, J., Terao, A., & Miyamoto, T. (2007). Neural bases associated with metaphor comprehension -An fMRI study-. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society 14*, 339–353.

Slepian, M. L., Masicampo, E. J., Toosi, N. R., and Ambady, N. (2012). The physical burdens of secrecy. *Journal of Experimental Psychology: General, 141*(4), 619–624.

Terai, A., Nakagawa, M., Kusumi, T., Koike, Y., & Himura, K. (2015). Enhancement of visual attention proceeds the emergence of novel metaphor interpretations. *Frontiers in Psychology 6*,1–8.

Vigliocco, G., Meteyard, L., Andrews, M., & Kousta, S. (2009). Toward a theory of semantic representation. *Language and Cognition, 1*(2), 219–247.

Wiemer-Hastings, K., & Xu, X. (2005). Content differences for abstract and concrete concepts. *Cognitive Science, 29*(5), 719–736.

Wu, L. L., & Barsalou, L. W. (2009). Perceptual simulation in conceptual combination: Evidence from property generation. *Acta Psychologica, 132*(2), 173–189.

Yee, E., and Thompson-Schill, S. L. (2016). Putting concepts into context. *Psychonomic Bulletin and Review, 23*(4), 1015–1027.

Zbikowski, L. M. (2008). Metaphor and music. In R. W. Jr. Gibbs (Ed.), *The Cambridge handbook of metaphor and thought* (pp. 502–524). Cambridge: Cambridge University Press.

Appendix A. Complete list of metaphor topics and selected sources

Metaphor Topic	Metaphor Source (6 images - nonverbal)
LOVE 愛情 *aijō*	Explosion (24%), Masks (22%), House (20%), Spider web (16%), Forest (10%), Suitcase (3%)
FRIENDSHIP 友情 *yūjō*	Colored Pencils (26%), Medicine (22%), Battery (16%), Shoes (15%), Heater (8%), Castle (7%)
CURIOSITY 好奇心 *kōkishin*	Dandelion (seeds) (29%), Lighter (24%), Map (21%), Paintbrush (11%), Grass (6%), Forest Path (2%)
LONELINESS 孤独 *kōdoku*	Desert (27%), Hole (20%), Rain cloud (19%), Barbed Wire (11%), Wall (10%), Desert (sunset focus) (5%), Pliers (3%)
AN IDEA アイディア *aidia*	Tree (37%), Butterfly (14%), Light Bulb (14%), Sailboat (12%), Bathtub (11%), Passport (4%)
MOTIVATION 動機付け *dōkizuke*	Seedling (35%), Fire (23%), Alarm Clock (12%), Racecar (12%), Cupcake (8%) Coffee Cup (6%)
DISAPPOINTMENT 失望 *shitsubō*	Crevice (31%), Downward Escalator (23%), Sinking Ship (18%), Birdcage (15%), Fly (insect) (3%), Jump Rope (chord) (3%)
THE ECONOMY 経済 *keizai*	Rollercoaster (38%), Scale (21%), Tightrope Walking (19%), The Tides (18%), Plant (3%), Slot Machine (2%),
JEALOUSY 嫉妬 *shitto*	Shark (27%), Sickness (26%), Stove top (24%), Wrench (9%), Megaphone (7%), Coffee Press (4%),
OLD AGE 老年 *rōnen*	Autumn Leaves (28%), Wine (25%), Blank Notebook (18%), Rocking Chair (18%), Fossilized shell (7%) Bridge (4%),
ANGER 怒り *ikari*	Volcano (39%), Bomb (25%), Lightning (16%), Octopus (7%), Fried Eggs and Bacon (5%), Lit Matches (3%)
FREEDOM 自由 *jiyū*	Wings (22%), Bicycle (21%), Paper Airplane (21%), Notebook (11%), Rocket (space) (7%), Swing (3%)
ANXIETY 不安 *fuan*	Storm (33%), Tangled Knot (32%), Crutches (12%), Cliff (11%), Hotpot (8%), Blender (0%),
SYMPATHY 同情 *dōjō*	Umbrella (rain) (18%), Heart (15%), Pillow (12%), Glasses (12%), Compass (8%), Sunshine (7%)
SUCCESS 成功 *seikō*	Ladder (25%), Money (24%), Diamond (16%), Star (12%), Fireworks (7%), Darts (6%)
HONESTY 正直 *shōjiki*	Sword (50%), Straight Road (20%), Magnifying Glass (10%), Handshake (8%), Mountaintop (4%), Watering Can (0%)
CULTURE 文化 *bunka*	Tree (roots) (43%), Handcuffs (14%), Salad (14%), Fence (11%), Computer Network (6%), Pillars (5%),
THE MIND 心 *kokoro*	Sponge (46%), Space (22%), Flower Vase (14%), Window (10%), Computer (2%), Rug (2%),
HOPE 希望 *kibō*	Rainbow (55%), Birds (15%), Stoplights (12%), Stethoscope (5%) Merry-go-round (3%), Leaf (3%)
EDUCATION 教育 *kyōiku*	Key (29%), Opened Door (26%), Dumbbell (18%), Construction (14%), Globe (8%), Hammer (2%),

Note: The totals do not always add up to 100% due to some responses not being able to be categorized for not following instructions (i.e., Love is love).

Appendix B. Coding of interpretations

The coding used in this study was adapted from Wu and Barsalou (2009).

Entity Properties: $\mathbf{E_{sys}}$ a global systemic property of the entity *(states, conditions, abilities, traits)*; $\mathbf{E_{CE}}$ an external component of the entity; $\mathbf{E_B}$ an action that is characteristic of an entity's behavior

Situation Properties: $\mathbf{S_F}$ the function or role the entity serves for the individual; $\mathbf{S_A}$ an action that a participant performs in a situation;

Introspective Properties: $\mathbf{I_c}$ contingency between two or more aspects of a situation *(if, enable, cause, because, depends, requires)* $\mathbf{I_R}$ representational state in the mind of a situational participant *(beliefs, ideas)* $\mathbf{I_O}$ an operation on a cognitive state *(retrieval, learning)*

Multiple Pivots in Statistical Machine Translation for Low Resource Languages

Sari Dewi Budiwati[1,2], **Masayoshi Aritsugi**[3]
[1]Computer Science and Electrical Engineering
Graduate School of Science and Technology, Kumamoto University, Japan
[2]School of Applied Science, Telkom University, Indonesia
[3]Big Data Science and Technology
Faculty of Advanced Science and Technology, Kumamoto University, Japan
saridewi@st.cs.kumamoto-u.ac.jp, aritsugi@cs.kumamoto-u.ac.jp

Abstract

We investigate many combinations of multiple pivots of four phrase tables on a low resource language pair, i.e., Japanese to Indonesia, in phrase-based Statistical Machine Translation. English, Myanmar, Malay, and Filipino from Asian Language Treebank (ALT) were used as pivot languages. A combination of four phrase tables was examined with and without a source to target phrase table. Linear and Fillup Interpolation approaches were employed with two measurement parameters, namely, data types and phrase table orders. The dataset was divided into two data types, i.e., sequential and random. Furthermore, phrase table orders comprise of two, viz., descending and ascending. Experimental results show that the combination of multiple pivots outperformed the baseline. Moreover, the random type significantly improved BLEU scores, with the highest improvement by +0.23 and +1.02 for Japanese to Indonesia (Ja-Id) and Indonesia to Japanese (Id-Ja), respectively. Phrase tables order experiments show a different result for Ja-Id and Id-Ja. The descending order has a higher percentage as much as 87.5% compared to the ascending order in Ja-Id. Meanwhile, the ascending order obtained more than 90% in Id-Ja. Finally, the combination of multiple pivots attempt shows a significant effect to reduce perplexity score in Ja-Id and Id-Ja.

1 Introduction

Statistical Machine Translation (SMT) needs parallel corpora in order to learn translation rules. Parallel corpora are bilingual texts where one of the corpora is an exact translation of the other. Some European languages achieve high-quality translation results with BLEU score more than 40 (Koehn, 2005; Ziemski et al., 2016) by using millions of line parallel corpora and the availability of linguistic tools, e.g., morphological analyzer, POS (part of speech) taggers, and stemmer. Unfortunately, except for Chinese and Japanese, Asian languages have limited parallel corpora with few thousands of line sentences (Riza et al., 2016; Nomoto et al., 2018; Tiedemann, 2012). Moreover, most of the Asian languages still lack linguistic tools and it is thus difficult to achieve the same translation results as European.

With the limited parallel corpora, there are two strategies to achieve high-quality translations, namely building parallel corpora and utilizing existing corpora (Trieu, 2017). Building parallel corpora is difficult since it can be time-consuming and expensive, and needs experts. Therefore, many researchers have focused on utilizing existing corpora, i.e., using pivot approaches (Utiyama and Isahara, 2007; Paul et al., 2009; Habash and Hu, 2009; El Kholy et al., 2013; Dabre et al., 2015; Trieu and Nguyen, 2017; Ahmadnia et al., 2017; Budiwati et al., 2019). Instead of direct translation between a language pair, pivot approaches use the third language as a bridge to overcome the parallel corpora limitation. Pivot approaches arise as preliminary assumption that there are enough parallel corpora between source-pivot and pivot-target languages.

In previous research, English has been the main choice of pivot languages. However Wu and Wang

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 345-355, Hakodate, Japan, September 13-15, 2019

(2007) and Paul et al., (2013) showed that non-English as a pivot language can improve translation quality for certain language pairs. Wu and Wang (2007) showed that using Greek as a pivot language has improved the translation quality compared to English in French to Spanish language pair. Greek as pivot language obtained +5.00 points, meanwhile English obtained +2.00 points. Paul et al., (2013) showed that from 420 experiments language pair in Indo-European and Asian languages, 54.8% is preferable using non-English as the pivot language. Moreover, Wu and Wang (2007) and Dabre et al., (2015) showed promising results by using more than one non-English language. Wu and Wang (2007) showed that using 4 languages, namely Greek, Portuguese, English, and Finnish outperformed the baseline BLEU score with more than +5.00 points. Dabre et al., (2015) also showed that using 7 non-English, namely Chinese, Korean, Marathi, Kannada, Telugu, Paite and Esperanto pivot languages outperformed the baseline BLEU score with more than 3.00 points in Japanese to Hindi language pair.

In this paper, we investigate many combinations of multiple pivots of four phrase tables on low resource language pairs. To make the discussion of this paper concrete, we use Japanese to Indonesia (Ja-Id) and Indonesia to Japanese (Id-Ja) language pairs as an example of them. First, we generate single pivot phrase table by each pivot language, i.e., English, Myanmar, Malay, and Filipino from Asian Language Treebank (ALT). We generate phrase tables by using different approaches, namely Cascade, Triangulation, Linear Interpolation (LI), and Fillup Interpolation (FI). Second, we chose which single pivot approaches have the best result. Last, the combinations of multiple pivots phrase tables were examined with and without a source to target (src-trg) phrase table.

We measured the effect of many combinations of multiple pivots by two parameters, namely data types and phrase table orders. The dataset was divided into two data types, i.e., sequential and random. Sequential type means that the dataset remains unchanged. Meanwhile, random type means the dataset was shuffled before being processed into SMT framework. Furthermore, phrase tables order comprises of two, viz., descending and ascending. Descending order arranges the four phrase tables from highest to lowest according to their BLEU scores. Ascending order is the opposite.

Our contributions are as follows:

- The use of *with and without* src-trg phrase table initiated by the fact that some language pairs have a small parallel corpus, while the others have none. We showed that for the language pair which does not have an src-trg parallel corpus, the translation could be accomplished with multiple pivots and produce high BLEU scores. Furthermore, employing the small src-trg parallel corpus could improve BLEU score more.

- The use of random data type became factors to make better translation results. We showed that the random data type has a significant improvement in translation results. The random data type could be applied in another language pair which has the same characteristics dataset as ALT, i.e., texts originating in English and translated into other languages.

- Phrase table orders can have some effect on perplexity scores. We showed that different phrase tables orders produced different perplexity scores in the experiments of this paper. We thus can say that the phrase tables order should be considered in the multiple pivots.

This paper is organized as follows. Section 2 discusses the availability of parallel corpora and efforts to improve the translation result in Ja-Id language pair. Sections 3 and 4 explain the SMT methodology and pivot approaches. Section 5 describes the experimental setup of many combinations of multiple pivots phrase tables. Section 6 discusses results. Section 7 concludes the paper.

2 Related Work

Current freely available Ja-Id parallel corpora are Asian Language Treebank (ALT) (Riza et al., 2016), TUFS Asian Language Parallel Corpus (TALPCo) (Nomoto et al., 2018), and OPUS (Tiedemann, 2012). ALT is a parallel treebank from English Wikinews to ten languages, i.e., English, Japanese, Indonesia, Khmer, Malay, Myanmar (Burmese), Filipino, Laotian, Thai and Vietnamese. ALT com-

prises of 20,106 sentences annotated with word segmentation, POS tags, and syntax information. The annotation information is limited to Japanese, English, Myanmar and Khmer languages. TALPCo is a parallel corpus of basic vocabulary words and example sentences in five languages, i.e., Japanese, English, Burmese (Myanmar), Indonesian and Malay. TALPCo comprises of 1,372 sentences and only the Burmese (Myanmar) data have been annotated for tokens and parts of speech (POS). OPUS is a collection of translated texts from movies subtitles, localization files (GNOME, Ubuntu, KDE4), Quran translations and a collection of translated sentences from Tatoeba. The parallel corpora of OPUS Ja-Id comprises of 2.9 M sentences from a different domain.

Several approaches have been done in Ja-Id machine translation as shown in Table 2, i.e., pivot languages (Paul et al., 2009), stemmer and removing particles (Simon and Purwarianti, 2013), lemmatization and reordering model (Sulaeman and Purwarianti, 2015), and neural machine translation (Adiputra and Arase, 2017). If we compare these approaches with their BLEU scores in Table 1, Paul et al., (2009) obtained the highest BLEU scores, i.e., 53.13 for Ja-Id and 55.52 for Id-Ja. This result denotes that high-quality translation results can be achieved with enough parallel corpora and certain strategy, e.g., pivot languages.

3 Statistical Machine Translation

Statistical Machine Translation (SMT) is based on a log-linear formulation (Och and Ney, 2002). Let *s* be a source sentence (e.g., Japanese) and *t* be a target sentence (e.g., Indonesia), SMT system outputs the best target translation t_{best} as follows

$$\begin{aligned} t_{\text{best}} &= \arg\max_{t} p(t|s) \\ &= \arg\max_{t} \sum_{m=1}^{M} \lambda_{\text{m}} h_{\text{m}}(t|s) \end{aligned} \tag{1}$$

where $h_m(t|s)$ represents feature function, and λ_{m} is the weight assigned to the corresponding feature function (Wu and Wang, 2007). The feature function $h_m(t|s)$ is a language model probability of target language, phrase translation probabilities (both directions), lexical translation probabilities (both directions), a word penalty, a phrase penalty, and a linear reordering penalty. The weight (λ_{m}) can be set by minimum error rate training (Och, 2003).

4 Pivot Methods

Pivot translation is a translation from a source language (SRC) to a target language (TRG) through an intermediate pivot (or bridging) language (PVT) (Paul et al., 2009). Several pivot approaches are sentence translation, triangulation and synthetic corpus.

4.1 Sentence translation

The sentence translation strategy or cascade uses two independently trained SMT systems (Utiyama and Isahara, 2007). These two independently systems are SRC-PVT and PVT-TRG systems. First, given a source sentence *s*, then translate it into *n* pivot sentences p_1, p_2, ..., p_{n} using an SRC-PVT system. Each p_{i} has eight scores namely language model probability of the target language, two phrase translation probabilities, two lexical translation probabilities, a word penalty, a phrase penalty, and a linear reordering penalty. The scores are denoted as h^{e}_{i1}, h^{e}_{i2}, ..., h^{e}_{i8}. Second, each p_{i} is translated into *n* target sentences t_{i1}, t_{i2}, ..., t_{in} using a PVT-TRG system. Each t_{ij} (*j*= 1, ..., n) also has the eight scores, which are denoted as $h^{\text{t}}_{\text{ij1}}$, $h^{\text{t}}_{\text{ij2}}$, ..., $h^{\text{t}}_{\text{ij8}}$. The situation is as follows:

$$\begin{aligned} SRC\text{-}PVT &= p_{\text{i}}(h^{\text{e}}_{\text{i1}}, h^{\text{e}}_{\text{i2}}, ..., h^{\text{e}}_{\text{i8}}) \\ PVT\text{-}TRG &= t_{\text{ij}}(h^{\text{t}}_{\text{ij1}}, h^{\text{t}}_{\text{ij2}}, ..., h^{\text{t}}_{\text{ij8}}). \end{aligned} \tag{2}$$

We define the score of t_{ij}, $S(t_{ij})$, as

$$S(t_{\text{ij}}) = \sum_{m=1}^{8} (\lambda^{\text{e}}{}_{\text{m}} h^{\text{e}}{}_{\text{im}} + \lambda^{\text{t}}{}_{\text{m}} h^{\text{t}}{}_{\text{ijm}}) \tag{3}$$

where $\lambda^{\text{e}}{}_{\text{m}}$ and $\lambda^{\text{t}}{}_{\text{m}}$ are weights set by performing minimum error rate training (Och, 2003). Finally, t_{best} will be

$$t_{\text{best}} = \arg\max_{t_{\text{ij}}} S(t_{\text{ij}}). \tag{4}$$

4.2 Triangulation

Triangulation, or known as phrase table translation is an approach for constructing an SRC-TRG translation model from SRC-PVT and PVT-TRG translation models (Hoang and Bojar, 2016). First, we

Experiments	Paul et al., (2009)		Simon et al., (2013)		Sulaeman et al., (2015)		Adiputra et al., (2017)
	Ja-Id	Id-Ja	Ja-Id	Id-Ja	Ja-Id	Id-Ja	Ja-Id
Baseline	52.90	55.52	0.06364	0.10424	0.0065	0.1369	9.34
Proposed	53.13	54.12	0.08806	0.08342	0.172	0.1652	6.45

Table 1: BLEU score comparison of related work.

Experiments	Paul et al., (2009)	Simon et al., (2013)	Sulaeman et al., (2015)	Adiputra et al., (2017)
Baseline	SMT	SMT	SMT	SMT
Proposed approaches	SMT with single pivot Cascade	SMT with stemmer	SMT with reordering model	NMT with biRNN
Dataset	160K of BTEC	500	1,132 of JLPT	725,495 of OPUS and ALT

Table 2: Proposed approaches and dataset of the related works.

train two translation models for SRC-PVT and PVT-TRG, respectively. Second, we build an SRC-TRG translation model with **p** as a pivot language.

Given a sentence **p** in the pivot language, the pivot translation model can be formulated as follows (Wu and Wang, 2007):

$$\begin{aligned} p(\mathbf{s}|\mathbf{t}) &= \sum_{p}(p(\mathbf{s}|\mathbf{t},\mathbf{p}))p(\mathbf{p}|\mathbf{t}) \\ &\approx \sum_{p}(p(\mathbf{s}|\mathbf{p}))p(\mathbf{p}|\mathbf{t}) \end{aligned} \tag{5}$$

where **s** and **t** are source and target translation model, respectively.

The triangulation translation model is often combined with SRC-TRG translation model, called phrase table combination. There are 3 ways to combine triangulation with SRC-TRG translation model, namely Linear Interpolation (LI), Fillup Interpolation (FI), and Multiple Decoding Paths (MDP). Linear Interpolation is performed by merging the tables and computing a weighted sum of phrase pair probabilities from each phrase table giving a final single table. Fillup Interpolation does not modify phrase probabilities but selects phrase pair entries from the next table if they are not present in the current table. Multiple Decoding Paths (MDP) method of Moses which uses all the tables simultaneously while decoding ensures that each pivot table is kept separate and translation options are collected from all the tables (Dabre et al., 2015).

More than one pivot language can be used to improve the quality of the translation performance, this is called multiple pivots. If we use n pivot languages and combine with SRC-TRG translation model, then the estimation of phrase translation probability and the lexical weight are as follows (Ahmadnia et al., 2017):

$$P(s|t) = \sum_{i=1}^{n} \alpha_{\mathrm{i}} P_{\mathrm{i}}(s|t) \tag{6}$$

$$P(s|t,\alpha) = \sum_{i=1}^{n} \beta_{\mathrm{i}} P_{\mathrm{i}}(s|t,\alpha) \tag{7}$$

where $P(s|t)$ and $P(s|t,\ \alpha)$ are the phrase translation probability and the lexical weight trained with SRC-TRG corpus estimated by using pivot language, while α_{i} and β_{i} are interpolation coefficients. Last, $\sum_{i=1}^{n} \alpha_{\mathrm{i}} = 1$, and $\sum_{i=1}^{n} \beta_{\mathrm{i}} = 1$.

5 Description of Languages, Dataset Scenarios and Experiments

In this section, we first describe the characteristics of pivot languages. Further, we explain how dataset is divided and used.

5.1 Languages involved

We use six datasets from ALT, i.e., Japanese, Indonesia, English, Myanmar, Malay and Filipino. Japanese and Indonesia datasets were used to build the direct translation as Baseline model. The Japanese language is an SOV language, while Indonesia is SVO language. Therefore, we chose pivot languages based on the similarity of a word order with Japanese or Indonesia. English and Malay have the same word order as Indonesia. Meanwhile, Myanmar has the same word order as Japanese. Filipino was chosen to evaluate the effect of VOS language. The word order and languages family can be seen in Table 3.

5.2 Dataset scenario

We divide the dataset into two data types, namely sequential (seq) and random (rnd). Sequential type

Languages	Word of order	Language Family
Japanese	SOV	Japonic
Indonesia	SVO	Austronesian
English	SVO	Indo-European
Myanmar	SOV	Sino-Tibetan
Malay	SVO	Austronesian
Filipino	VOS	Austronesian

Table 3: Language characteristics.

means that the dataset remains unchanged. Meanwhile, random type means the dataset was shuffled before used in SMT framework. We used `random.shuffle()` method from python library.

We divide datasets into 8.5K for training (*train*), 2K for tuning (*dev*) and 1K for the evaluation *(eval*). Overall, we conduct 132 experiments, i.e., 4 Baselines, 32 SRC-PVT and PVT-TRG, 64 single pivots, and 32 multiple pivots.

5.3 Experimental setup

We used Moses decoder (Koehn et al., 2007) and Giza++ for word alignment process, phrase table extraction and decoding. We used 3-gram KenLM (Heafield, 2011) for language model, MERT (Och, 2003) for tuning and BLEU (Papineni et al., 2002) for evaluation from Moses package.

5.3.1 Single pivot

In the single pivot, we implement four approaches, i.e., Cascade, Triangulation, Linear Interpolation (LI) and Fillup Interpolation (FI). In the Cascade approach, we construct SRC-PVT and PVT-TRG systems, where the first system translates the source language input into the pivot language and the second system takes the translation result as input and translates into the target language. As a result, we construct 16 SRC-PVT and 16 PVT-TRG systems.

In the Triangulation approach, we construct phrase tables as follows:

- Pruning the SRC-PVT and PVT-TRG phrase table from the Cascade experiments using *filter-pt* (Johnson et al., 2007). The pruning activity intended to minimize the noise of SRC-PVT and PVT-TRG phrase table.
- Merging two pruning phrase tables using *Tm-Triangulate* (Hoang and Bojar, 2015). The result is denoted as `TmTriangulate` phrase table.

In the Linear Interpolation approach, we combine `TmTriangulate` and SRC-TRG phrase table with *dev* phrase table as a reference. The result is called `TmCombine` phrase table. In Fillup interpolation, we use *backoff* mode thus the result is called `TmCombine-Backoff` phrase table. We use *tmcombine* and *combine-ptables* tools to construct `TmCombine` and `TmCombine-Backoff` phrase tables.

5.3.2 Multiple pivots

In multiple pivots, first, we observe BLEU scores result from each approach in a single pivot. Then, we employ phrase tables from the best pivot approaches into the next step, i.e., the combination of multiple pivots. As shown in Figure 1 and Figure 2, the Linear and Fillup Interpolation approaches have higher BLEU scores compared to Baseline. Therefore, we use the four phrase tables from Linear and Fillup Interpolation approaches, i.e., English phrase table (EnPT), Myanmar phrase table (MyPT), Malay phrase table (MsPT) and Filipino phrase table (FiPT).

Next, we combine the four phrase tables based on the single pivot BLEU score, viz., descending and ascending orders. Descending order sorts the four phrase tables from highest to lowest according to their BLEU scores. Ascending order is the opposite. For example, the BLEU scores of Linear Interpolation approach are 11.34 for EnPT, 12.21 for MyPT, 12.11 for MsPT, and 12.15 for FiPT. For descending order, we put the four phrase tables, i.e., MyPT, FiPT, MsPT, and EnPT, respectively. Meanwhile, for ascending order, we put the four phrase tables, i.e., EnPT, MsPT, FiPT, MyPT, respectively.

The combinations of multiple pivots phrase tables were examined with and without an SRC-TRG phrase table, as follows:

- Merging of four phrase tables without SRC-TRG phrase table using Linear Interpolation approach. The result is denoted as All-LinearInterpolate `All-LI`.

- Merging of four phrase tables without SRC-TRG phrase table using Fillup Interpolation approach. The result is denoted as All-FillupInterpolation `All-FI`.

- Combining `All-LI` with SRC-TRG phrase table using Linear Interpolation approach. The result is denoted as `Base-LI`.

- Combining `All-FI` with SRC-TRG phrase table using Fillup Interpolation approach. The result is denoted as `Base-FI`.

6 Result and Discussion

In this section, we will discuss results based on BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) and perplexity scores. BLEU score is a metric for evaluating the generated sentence compared to the reference sentence. High BLEU scores indicate a better system. Perplexity score is frequently used as a quality measure for language models (Sennrich, 2012). Lower perplexity scores indicate that the language model is better compared to higher perplexity score. We used the query from KenLM (Heafield, 2011) to get the perplexity including OOV (Out of Vocabulary). OOV is unknown words that do not appear in the training corpus. We show the perplexity scores of the target language test dataset according to the 3-gram language model trained on the respective training dataset.

6.1 Baseline translation results

The Baseline is a direct translation between languages pair, namely Ja-Id and Id-Ja. We construct two Baseline systems in each language pair, based on data types, i.e., sequential and random.

Baseline BLEU scores are given in column 2 of Table 4 and Table 5 for Ja-Id and Id-Ja, respectively. As shown in the tables, Baseline Random obtained higher BLEU score compared to Baseline Sequential. The BLEU score of Baseline Random Ja-Id is 12.17, +0.21 points higher compared to Baseline Sequential. Meanwhile, the BLEU score of Baseline Random Id-Ja is 12.00, +1.00 points higher compared to Baseline Sequential.

Baseline perplexity scores are given in Figure 3 and Figure 4 for Ja-Id and Id-Ja, respectively. As shown in the figures, the Ja-Id and Id-Ja perplexity scores of Random data type obtained higher point compared to the Sequential data type. Perplexity score of Ja-Id in Random data type has 384.59, while Sequential data type has 291.51. Furthermore, perplexity score of Id-Ja in Random data type has 81.58, while Sequential data type has 71.94.

The results denote that Random data type obtained higher BLEU score but it has OOV issue, compared to Sequential data type. In the next section, we showed our efforts to reduce perplexity scores by using multiple pivots.

6.2 Multiple pivots translation results

6.2.1 Single pivot results

The Triangulation approach was the worst approach in Ja-Id and Id-Ja. All the results of Triangulation have smaller BLEU score compared to Baseline. The Cascade approach also has lower scores compared to Baseline, except three experiments in Sequential data type by using Malay and English as a pivot language. The three experiments outperformed the Baseline by range from +0.05 to 1.18 points. However, we didn't use the Cascade results because of its different technique compared to other approaches. The Cascade approach did not combine phrase tables such as Linear and Fillup Interpolation. The Cascade approach used two independently systems, i.e., SRC-PVT and PVT-TRG. The SRC-PVT system translates the Japanese text into the pivot language. The PVT-TRG system takes the translation result as input and translates into Indonesian text.

The Linear Interpolation (LI) and Fillup Interpolation (FI) approaches show significant result in Ja-Id and Id-Ja. Both approaches have higher BLEU scores compared to Baseline, by more than 75% experiments. This was shown in Figure 1 and Figure 2 for Ja-Id and Id-Ja, respectively.

In terms of language, Myanmar became a main option as pivot language in Ja-Id Sequential data type. Meanwhile, Ja-Id Random data type has two options of pivot language, i.e., Malay, and Myanmar. Surprisingly, Myanmar also became a main option as pivot language in Id-Ja Sequential and Random data types. As we look to the language characteristics in Table 3, Myanmar has the same word order as Japanese while Malay has the same word

order as Indonesia. The results denote that word order closely related to the source or target language should be considered when choosing pivot language.

In terms of data type, Sequential or Random data types could be chosen in Ja-Id. Both data types have increased the BLEU scores by 75% of experiments. Random data type was preferable in Id-Ja because the highest improvement points were achieved by +1.84 compared to Baseline. The results denote that data type is an important parameter to consider to improve the BLEU score.

In terms of perplexity score, the LI and FI approaches in different data types are unable to reduce the scores. The single pivot language even increased the perplexity scores as shown in Figure 3 and Figure 4. We showed how to reduce the perplexity scores by using multiple pivots in the next section.

6.2.2 Multiple pivots results

From the single pivot, LI and FI become the best approach to improve the BLEU scores compared to the Baseline. Therefore, we use the phrase tables from both approaches and we did combinations of multiple pivots phrase tables, i.e., All-LI, All-FI, Base-LI, and Base-FI, as described in Section 5.3.

For example in Ja-Id of All-LI, we combine the four phrase tables from the single pivot LI approach by descending and ascending orders. First, we observe the BLEU scores of LI Sequential data type are 11.34 for EnPT, 12.21 for MyPT, 12.11 for MsPT, and 12.15 for FiPT. Next, we combine the four phrase tables according to their BLEU scores in descending order, i.e., MyPT, FiPT, MsPT, and EnPT, respectively. Last, we combine the four phrase tables according to their BLEU scores in ascending order, i.e., EnPT, MsPT, FiPT, MyPT, respectively. As a result, the BLEU scores have different scores for descending and ascending orders, i.e., 12.01 and 12.20, respectively. The results are shown in Figure 5.

We did not use SRC-TRG phrase table in All-LI and All-FI approaches, and their BLEU scores outperformed Baseline. The results denote that the translation could be accomplished with multiple pivots and still produce high BLEU scores without using SRC-TRG phrase table. Moreover, the translation results could have higher BLEU scores if there is a small SRC-TRG phrase table, as in Base-LI and Base-FI results.

The combinations of multiple pivots phrase tables have different effects on the BLEU scores, when we used different order. In Ja-Id, the descending order was preferable because more than 87.5% experiments result outperformed the Baseline. In Id-Ja, the ascending order was preferable because all the experiments outperformed the Baseline. The results are shown in Figure 5 and Figure 6 for Ja-Id and Id-Ja, respectively.

In terms of data type, most of the results of Ja-Id outperformed the Baseline, excluding the Base-FI Random data type. Meanwhile, all the results of Id-Ja outperformed the Baseline. The highest improvement score was obtained by Base-LI Random data type in Ja-Id descending, by +0.23 points. Meanwhile, the highest improvement was obtained by ALL-FI Sequence data type in Id-Ja ascending, as much as +1.84 points. The results indicate that data types have a significant effect to improve the BLEU scores.

In terms of perplexity scores for Ja-Id, All-LI and All-FI show poor results. However, the perplexity scores could be reduced in Random data type of Base-LI and Base-FI. Both approaches use SRC-TRG phrase table in the combination process. The results show that the SRC-TRG phrase table has a significant impact on reducing the perplexity score. Meanwhile, the perplexity scores in Id-Ja could be reduced without using the SRC-TRG phrase table. Moreover, the Base-LI and Base-FI results have lower perplexity scores compared to All-LI and All-FI. We show the perplexity scores in Figure 7 and Figure 8 for Ja-Id and Id-Ja, respectively.

We summarize the results of single and multiple pivots in Table 4 and Table 5. We show BLEU scores of best approaches in Figure 9 and Figure 10, and the perplexity scores of best approaches in Figure 11 and Figure 12.

7 Conclusion and Future Work

In this paper, we showed experiment results of single and multiple pivots in Ja-Id and Id-Ja. We used English, Myanmar, Malay, and Filipino as pivot languages in single pivot. We implemented four approaches, i.e., Cascade, Triangulation, Linear Inter-

Data Type	Baseline	Single Pivot				Multiple Pivots	
		Cascade	Triangulation	LI	FI	Desc	Asc
Sequential	11.96	12.01 (MS)	9.71 (EN)	12.21 (MY)	12.27 (MY)	12.23 (Base-LI)	12.37 (Base-FI)
Random	12.17	11.81 (MS)	9.62 (FI)	12.22 (MS)	12.29 (MY)	12.40 (Base-LI)	12.27 (All-FI)

Table 4: Best BLEU score in baseline, single and multiple pivots for Japanese to Indonesia

Data Type	Baseline	Single Pivot				Multiple Pivots	
		Cascade	Triangulation	LI	FI	Desc	Asc
Sequential	11.00	12.18 (MS)	8.26 (EN)	12.03 (MY)	12.40 (MY)	12.15 (Base-LI)	12.84 (ALL-FI)
Random	12.00	11.13 (MS)	9.17 (MS)	12.84 (MY)	12.88 (MY)	12.74 (All-FI)	13.02 (ALL-FI)

Table 5: Best BLEU score in baseline, single and multiple pivots for Indonesia to Japanese

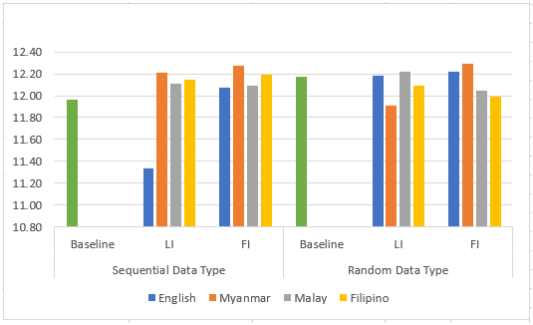


Figure 1: Single pivot BLEU scores of Ja-Id for LI and FI approaches.

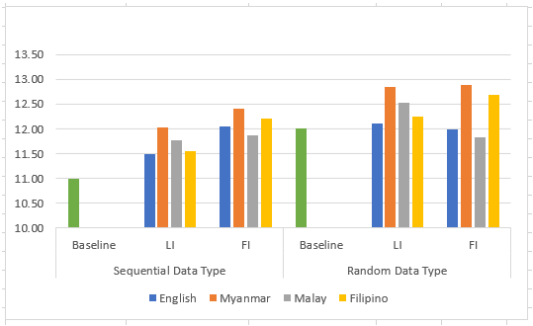


Figure 2: Single pivot BLEU scores of Id-Ja for LI and FI approaches.

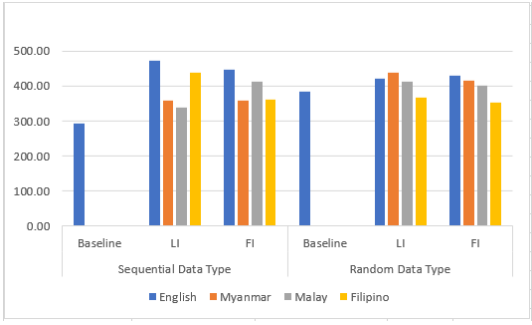


Figure 3: Perplexity Score of Ja-Id single pivot for LI and FI approaches.

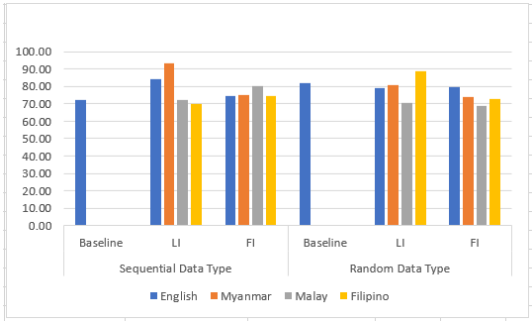


Figure 4: Perplexity Score of Id-Ja single pivot for LI and FI approaches.

polation (LI) and Fillup Interpolation (FI) in single pivot. We found that LI and FI approaches outperformed the Baseline. In multiple pivots, we implemented four approaches, i.e., All-LI, All-FI, Base-LI, and Base-FI. We found that most of all approaches in multiple pivots outperformed the Baseline.

We divided the dataset into two data types in single and multiple pivots, namely sequential and random. The data types showed different effects on the language pairs. In Ja-Id of single pivot, sequential or random could be chosen to improve the BLEU score. Both data types have increased the BLEU scores by 75% of experiments. However, random data type was preferable in Id-Ja because the highest improvement points were achieved by +1.84. Random data type was preferable for Ja-Id and Id-Ja in multiple pivots. The highest improvement points were achieved by +0.23 and 1.84 for Ja-Id and Id-Ja, respectively.

In multiple pivots, we combined the four phrase tables from the best single pivot approaches, i.e.,

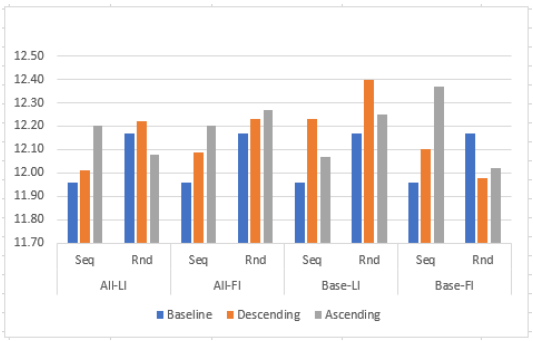


Figure 5: BLEU score for Ja-Id in multiple pivots.

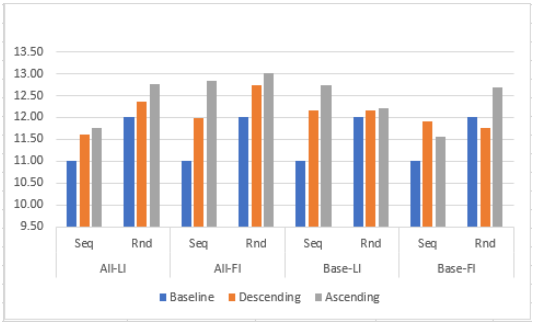


Figure 6: BLEU score for Id-Ja in multiple pivots.

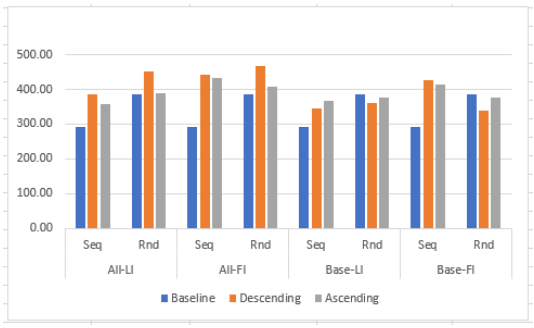


Figure 7: Perplexity score for Ja-Id in multiple pivots.

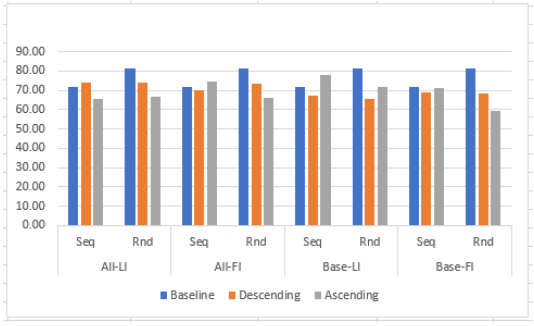


Figure 8: Perplexity score for Id-Ja in multiple pivots.

Linear Interpolation (LI) and Fillup Interpolation (FI). The combinations of multiple pivots phrase tables were examined with and without src-trg phrase table. We measured the effect by phrase tables orders, i.e., descending and ascending. From the experiment results, the descending order was preferable in Ja-Id. Meanwhile, the ascending order was preferable in Id-Ja.

In the experiments, we did not show the combinations of two or three phrase tables as in (Wu and Wang, 2007). This will be included in our future work to give a better explanation on whether the combinations of two or three phrase tables will give better improvement compared to four phrase tables. Furthermore, the combination of the best phrase tables from each data type should be taken into account for next future research.

References

Cosmas Krisna Adiputra and Yuki Arase. 2017. Performance of Japanese-to-Indonesian Machine Translation on Different Models. In *The 23rd Annual Meeting of the Society of Language Processing*. The Association for Natural Language Processing.

Benyamin Ahmadnia, Javier Serrano, and Gholamreza Haffari. 2017. Persian-Spanish Low-Resource Statistical Machine Translation Through English as Pivot Language. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017, Varna, Bulgaria, September 2 - 8, 2017*, pages 24–30.

Sari Dewi Budiwati, Al Hafiz Akbar Maulana Siagian, Tirana Noor Fatyanosa, and Masayoshi Aritsugi. 2019. DBMS-KU Interpolation for WMT19 News Translation Task. In *Proceedings of the Fourth Conference on Machine Translation*, pages 340–345, Florence, Italy. Association for Computational Linguistics.

Raj Dabre, Fabien Cromieres, Sadao Kurohashi, and Pushpak Bhattacharyya. 2015. Leveraging Small Multilingual Corpora for SMT Using Many Pivot Languages. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1192–1202. Association for Computational Linguistics.

Ahmed El Kholy, Nizar Habash, Gregor Leusch, Evgeny Matusov, and Hassan Sawaf. 2013. Language Independent Connectivity Strength Features for Phrase

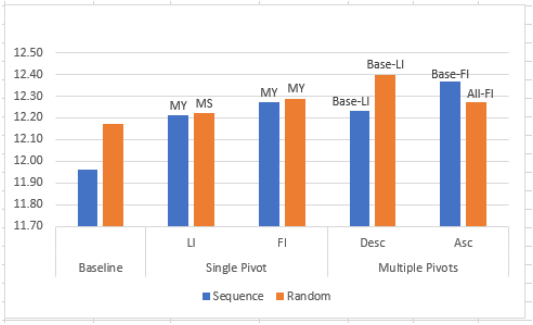


Figure 9: BLEU scores of Ja-Id best approach.

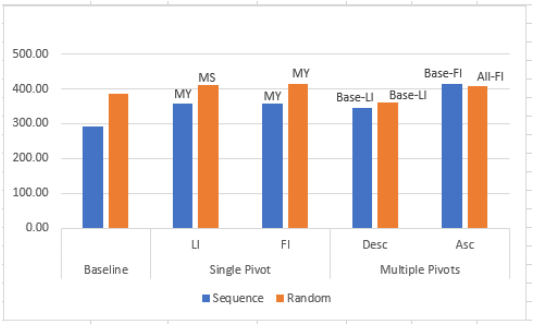


Figure 11: Perplexity scores of Ja-Id best approach.

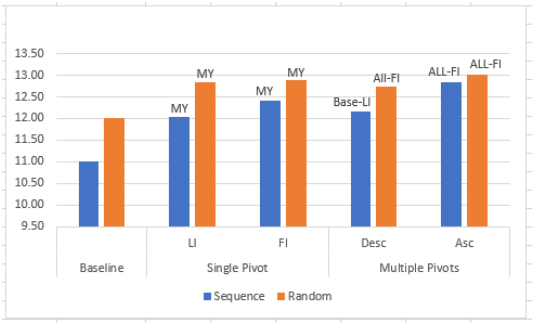


Figure 10: BLEU scores of Id-Ja best approach.

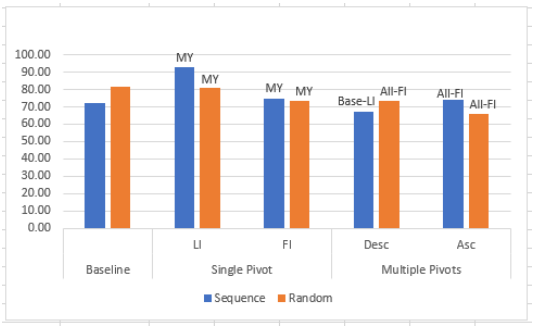


Figure 12: Perplexity scores of Id-Ja best approach.

Pivot Statistical Machine Translation. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 412–418. Association for Computational Linguistics.

Nizar Habash and Jun Hu. 2009. Improving Arabic-Chinese Statistical Machine Translation Using English As Pivot Language. In *Proceedings of the Fourth Workshop on Statistical Machine Translation*, StatMT '09, pages 173–181, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kenneth Heafield. 2011. KenLM: Faster and Smaller Language Model Queries. In *Proceedings of the EMNLP 2011 Sixth Workshop on Statistical Machine Translation*.

Duc Tam Hoang and Ondřej Bojar. 2015. TmTriangulate: A Tool for Phrase Table Triangulation. *The Prague Bulletin of Mathematical Linguistics*, 104:75–86.

Duc Tam Hoang and Ondrej Bojar. 2016. Pivoting Methods and Data for Czech-Vietnamese Translation via English. In *Proceedings of the 19th Annual Conference of the European Association for Machine Translation*, pages 190–202.

Howard Johnson, Joel Martin, George Foster, and Roland Kuhn. 2007. Improving Translation Quality by Discarding Most of the Phrasetable. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open Source Toolkit for Statistical Machine Translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180. Association for Computational Linguistics.

Philipp Koehn. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Conference Proceedings: the tenth Machine Translation Summit*, pages 79–86, Phuket, Thailand. AAMT, AAMT.

Hiroki Nomoto, Kenji Okano, David Moeljadi, and Hideo Sawada. 2018. TUFS Asian Language Parallel Corpus (TALPCo). In *Proceedings of the Twenty-Fourth Annual Meeting of the Association for Natural Language Processing*.

Franz Josef Och and Hermann Ney. 2002. Discriminative Training and Maximum Entropy Models for Statistical Machine Translation. In *Proceedings of the 40th Annual Meeting of Association for Computational Linguistics*, ACL '02, pages 295–302, Strouds-

burg, PA, USA. Association for Computational Linguistics.

Franz Josef Och. 2003. Minimum Error Rate Training in Statistical Machine Translation. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 160–167, Stroudsburg, PA, USA. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*.

Michael Paul, Hirofumi Yamamoto, Eiichiro Sumita, and Satoshi Nakamura. 2009. On the Importance of Pivot Language Selection for Statistical Machine Translation. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Short Papers*, NAACL-Short '09, pages 221–224, Stroudsburg, PA, USA. Association for Computational Linguistics.

Michael Paul, Andrew Finch, and Eiichiro Sumita. 2013. How to Choose the Best Pivot Language for Automatic Translation of Low-Resource Languages. *ACM Trans. Asian Lang. Inf. Process.*, 12(4):14:1–14:17, October.

H. Riza, M. Purwoadi, Gunarso, T. Uliniansyah, A. A. Ti, S. M. Aljunied, L. C. Mai, V. T. Thang, N. P. Thai, V. Chea, R. Sun, S. Sam, S. Seng, K. M. Soe, K. T. Nwet, M. Utiyama, and C. Ding. 2016. Introduction of the Asian Language Treebank. In *2016 Conference of The Oriental Chapter of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA)*, pages 1–6, Oct.

Rico Sennrich. 2012. Perplexity Minimization for Translation Model Domain Adaptation in Statistical Machine Translation. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, EACL '12, pages 539–549, Stroudsburg, PA, USA. Association for Computational Linguistics.

H. S. Simon and A. Purwarianti. 2013. Experiments on Indonesian-Japanese statistical machine translation. In *2013 IEEE International Conference on Computational Intelligence and Cybernetics (CYBERNETICSCOM)*, pages 80–84, Dec.

M. A. Sulaeman and A. Purwarianti. 2015. Development of Indonesian-Japanese Statistical Machine Translation Using Lemma Translation and Additional Postprocess. In *2015 International Conference on Electrical Engineering and Informatics (ICEEI)*, pages 54–58, Aug.

Jorg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Hai-Long Trieu and Le-Minh Nguyen. 2017. A Multilingual Parallel Corpus for Improving Machine Translation on Southeast Asian Languages. In *Proceedings of MT Summit XVI, vol.1: Research Track*.

Hai-Long Trieu. 2017. *A Study on Machine Translation for Low-Resource Languages*. Ph.D. thesis, Japan Advanced Institute of Science and Technology, September.

Masao Utiyama and Hitoshi Isahara. 2007. A Comparison of Pivot Methods for Phrase-Based Statistical Machine Translation. In *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pages 484–491. Association for Computational Linguistics.

Hua Wu and Haifeng Wang. 2007. Pivot Language Approach for Phrase-based Statistical Machine Translation. *Machine Translation*, 21(3):165–181, September.

Michal Ziemski, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations Parallel Corpus v1. 0. European Language Resources Association (ELRA), May.

Semi-supervised learning for all-words WSD using self-learning and fine-tuning

Rui Cao **Jing Bai** **Wen Ma** **Hiroyuki Shinnou**
Ibaraki University, Department of Computer and Information Sciences
4-12-1 Nakanarusawa, Hitachi, Ibaraki JAPAN 316-8511
{18nd305g, 19nd301r, 19nd302, hiroyuki.shinnou.0828}
@vc.ibaraki.ac.jp

Abstract

In this paper, we propose a semi-supervised learning method using self-learning and fine-tuning for all-words word sense disambiguation (WSD). The all-words WSD can be regarded as a sequence labeling problem, so we use a bidirectional Long Short-term Memory (LSTM) to solve it. Furthermore, we propose the semi-supervised learning method to improve that LSTM model, where self-learning is essentially used. In general, self-learning is the method for a classification problem, not for a sequence labeling problem. To apply self-learning to an all-words WSD, the LSTM model is trained by not accumulating the loss from the low probability label. We also construct the model with additional labeled data and then fine-tune by using the original labeled data. As result, the precision has been improved from the precision of the model learned from only initial labeled data.

1 Introduction

In this paper, we propose a semi-supervised learning method using self-learning for all-words word sense disambiguation (WSD). WSD is a task to identify the sense of a polysemy word in a sentence, and hence is essential in semantic analysis. However, its use in an actual system is difficult because the general WSD is developed for limited target words only. Thus, an all-words WSD that provides senses to all polysemy words in a given sentence should be developed.

Normally, WSD can be solved through supervised learning. Thus, labeled training data, that are example sentences with sense tags, are required for each word of WSD. In an all-words WSD, a large number of words with sense tags are necessary because the target word is unlimited.

Thus, unsupervised learning should also be considered (Tanigaki et al., 2013; Komiya et al., 2015; Suzuki et al., 2018). However, a problem regarding accuracy exists in this case. Under such situation, the corpus with sense tags has been gradually prepared. Recently, the all-words WSD in a supervised learning framework has been attempted to address (Shinnou et al., 2017b; Shinnou et al., 2018). However, the currently available corpus with sense tags is limited and we cannot obtain a sufficient accuracy. Therefore, we attempt to develop an all-words WSD with high accuracy through semi-supervised learning.

Semi-supervised learning is a method used in training classifiers from a small amount of labeled data and large amount of unlabeled data. In the case of all-words WSD, the unlabeled data means a plain corpus. Because obtaining a large amount of plain corpus is easy, semi-supervised learning is promising approach for all-words WSD. Therefore, we propose the semi-supervised learning method to improve that LSTM model, where self-learning is essentially used. In general, self-learning is the method for a classification problem, not for a sequence labeling problem. To apply self-learning to the all-words WSD, the LSTM model is trained by not accumulating the loss from the low probability label. We construct the model with additional labeled data and then fine-tune by using the original labeled data. As result, the precision has been im-

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 356-361, Hakodate, Japan, September 13-15, 2019

proved from the precision of the model learned from only initial labeled data.

2 Related Work

Many studies on semi-supervised learning for classifiers are already available. Co-training (Blum and Mitchell, 1998) and expectation-maximization (EM) (Nigam et al., 2000) algorithm are the popular and conventional methods. Co-training is a method utilized to improve classifier reciprocal by using two independent views. In the EM algorithm, a generation model $p(x;\theta)$ has been set and considered the label as potential variable to construct $p(z|x)$.

Based on this idea, the semi-supervised learning can be divided into two categories. The first one is employing a classifier trained by the original labeled data and then fine-tuning the classifier by data with a probability label. Self-learning (Abney, 2007) and label propagation (Zhu and Ghahramani, 2002) also belong to this category.

The second one is mapping data to space.[1] Initially, mapping unlabeled data into space which can divide them well, then mapping labeled data to that space. Finally, the process identifies and constructs classifiers in that space. Generally, if the data can be mapped into a low-dimensional space,a small amount of labeled data is sufficient to estimate the boundaries between classes. hence, the semi-supervised learning can be approved. The multi-body theory (Rifai et al., 2011) and method using generation model (Cozman et al., 2003) belong to this category. Additionally, the semi-supervised learning method using deep generation model has a similar framework with the semi-supervised learning using the generation model. Thus, we consider the method of mapping the unlabeled data into space that can accurately divide them to be used by the network. (Kingma et al., 2014; Rasmus et al., 2015; Salimans et al., 2016)

The pre-trained method is a representative of the semi-supervised learning for a sequence labeling model (Peters et al., 2017; Qi et al., 2009). To training the identify vector as input, which can be recognition by a recognizer from the unlabeled data, and added it to the training and test data. The recent pre-training method used for a network-based language model, referred to as ELMo (Peters et al., 2018), also belongs to this type. BERT (Devlin et al., 2018) also belongs to the same framework which was developed from ELMo.

For the all-words WSD, some unsupervised learning using the topic model has been proposed (Boyd-Graber et al., 2007; Komiya et al., 2015). This should be easily extended to semi-supervised learning because a generative model has been established.

3 All-words WSD Based on Bidirectional LSTM

The all-words WSD can be regarded as a sequence labeling problem that provides labels (sense) to each word in the input word sequence. An LSTM is used when the sequence labeling problem handles a neural network and corresponds to the time series by learning from the hidden layer of time t and the state of input from $t-1$. It is also a model that addresses the time series data, Natural language processing can treat word sequence from words and sentences that are regarded as the time series data. Therefore, the word after time t which be paying attention is available, and then the data can be also analyzed from the reverse direction. The model in (Figure 1) is using forward direction and reverse direction LSTM while obtaining the output for time t. Hence, the model is referred to as bidirectional LSTM.

4 Bidirectional LSTM with Self-Learning

Self-leaning utilizes the current classifier to provide a label with the probability for the unlabeled data and considers the labels with high probabilities as correct labels. By adding the data to the labeled data (training data), the accuracy of the classifier is gradually increased. In self-learning of the sequence labeling model, the sequence labeling model receives unlabeled word strings as inputs and provides a label with probability for each word. Thus, the labels with high and low probabilities are mixed and the word sequence cannot be simply added to the training data. Therefore, self-learning for a sequence labeling model has two problems: (1) enhancing the training data and (2) using increased training data.

[1] generally contains a lower dimensional space than the original data.

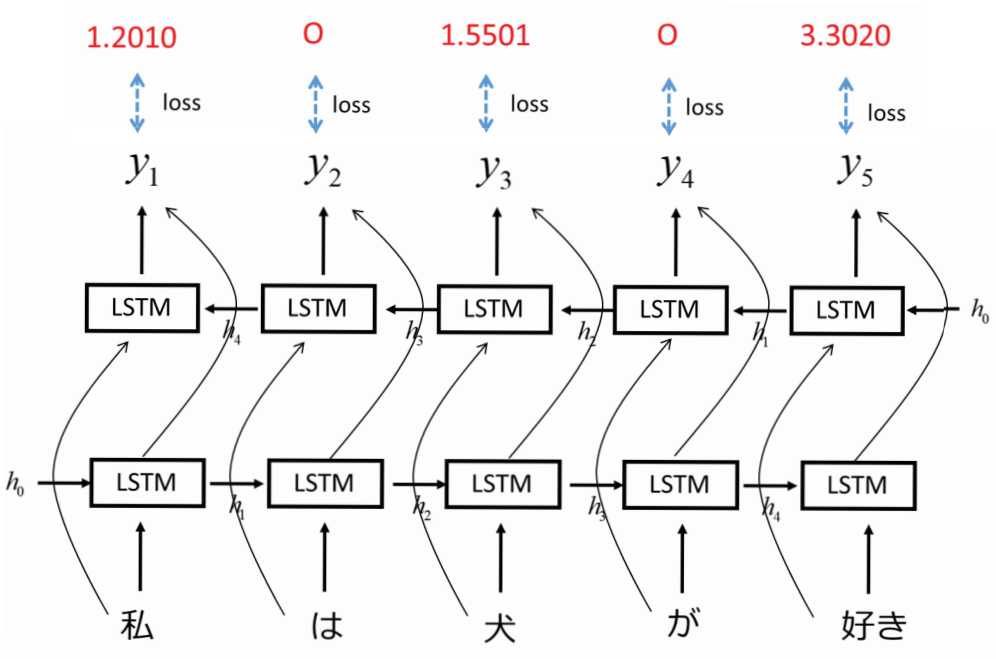


Figure 1: Learning of a bidirectional LSTM

4.1 Avoiding learning from low probability labels

For the first problem mentioned in the previous section, we do not learn from a label with low probability (confidence degree). Thus, the sequence labeling model provides labels with probabilities for each unlabeled word and then adds the word list to the training data regardless of the probability. Performing this process using LSTM is easy. For each word in LSTM, $loss_i$ is obtained from the difference between the output value and label of the word w_i, thereby accumulating the loss. When the processing is completed up to the end of the sentence, the network parameters are updated based on the accumulated $\sum_i loss_i$. If a label with low confidence degree exists. then $loss_i = 0$ is acceptable.

4.2 Using supplemental labeled data

For the second problem described in the previous section, the following three approaches are considered. In this case, the training data with label are assumed to be D. and the labeled data with probability obtained through self-learning are assumed to be A.

In this study, we attempt the following three approaches and then determine the most effective approach.

(a) Using $D \cup A$ in training the bidirectional LSTM model

(b) Using D in training the bidirectional LSTM model and A to fine-tune the model

(c) Using A in training the bidirectional LSTM model and D to fine-tune the model

5 Experiment

In this study, the sense ID in the Word List by Semantic Principles (WLSP) provided by National Institute for Japanese Language and Linguistics is regarded as sense. the Japanese sense dataset, Balanced Corpus of Contemporary Written Japanese (BCCWJ) tagged with WLSP, has been released from National Institute for Japanese Language and Linguistics (NINJAL) (Kato et al., 2017). We utilize it as a sense-tagged corpus for Japanese all-words WSD. Approximately 10% of this data is used as test data T, whereas the rest are labeled training data D. Regarding the number of sentences, D has 12,482 sentences and T has 1,498 sentences. Moreover, unlabeled data U are used in self-learning with regard

to the label. We used 100,000 sentences that are randomly extracted from the Mainichi Shimbun from 1993 to 1999.

Two layers were used as a bidirectional LSTM model. To convert the words into distributed representations we used the nwjec2vec (Shinnou et al., 2017a), which is exiting Japanese distributed expression data without learning.

Then, we utilized D in training the bidirectional LSTM model and evaluated it using T, where T was divided into 36,263 words by using the system. Considering that division of 2212 words was different from the correct answer data, the remaining 34,051 words (sense) were used as the evaluation subject. Meanwhile, 18,522 words are polysemy. The correct answer rate of these 18,522 words was determined as the correct answer rate of the all-words WSD. Figure 2 show the results. Moreover, the abscissa represents the number of epochs during the learning of the bidirectional LSTM, whereas the ordinate represents the correct answer data as described previously. The correct answer rate of the model was obtained after 18 epoch with the best value of 0.799. Because the system in (Shinnou et al., 2018) was used, the correct answer rate of the model constructed after 20 epochs where the value of 0.796 the base correct answer rate is 0.796.

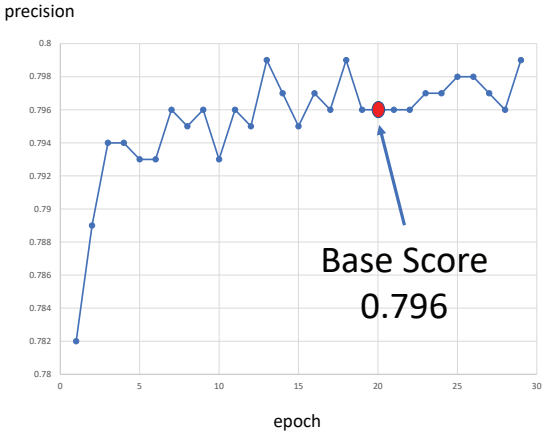


Figure 2: Using only D in training the model

Then, the model constructed after 20 epochs to the given U label with probability was used the label whose probability is less than 0.8 was replaced with the label of -1 to construct a supplemental version of the labeled data A.

(a) Using $D \cup A$ in training the model

We used $D \cup A$ as the new data to train the bidirectional LSTM model and then employed T to evaluate it. Figure 3 show the training results. The correct answer rate in this method was increased to 0.798.

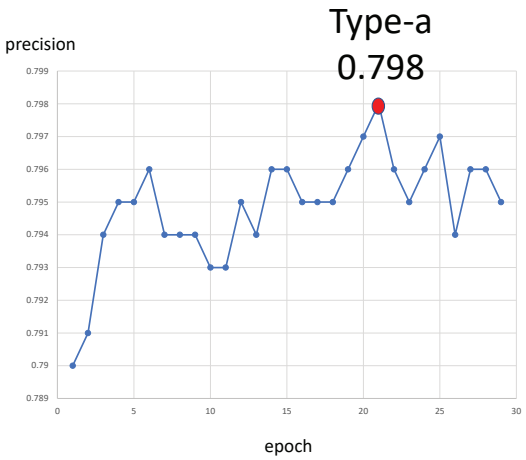


Figure 3: Using $D \cup A$ in training the model

(b) Fine-tuning ($D \rightarrow A$)

We first used D to train the bidirectional LSTM model, then A to fine-tune it, and finally T to evaluate it.Figure 4 shows the training results. In this case, the correct rate was reduced to 0.794.

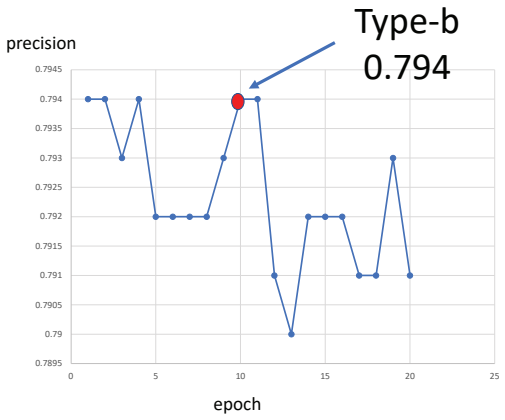


Figure 4: Fine-tuning ($D \rightarrow A$)

(c) Fine-tuning ($A \rightarrow D$)

We employed A to train the bidirectional LSTM model. D to fine-tune it. and T to evaluate it. Figure 5 shows the training result. In this case, the correct rate was increased to 0.799.

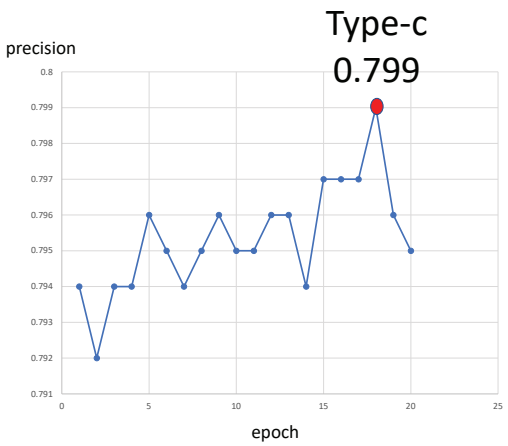


Figure 5: Fine-tuning ($A \rightarrow D$)

6 Discussion

About how to use the enhanced data, In (c) approach which creates the model based on enhanced and fine-tuning it with original labeled data. As shown in Figure 5, the correct answer rate is increased gradually, which is higher than of the base sequence labeling model. Therefore, semi-supervised learning method through self-learning can be considered to be a promising method.

However, the correct answer rate has a minimal improvement. Thus, self-learning was not effective in this experiment. Particularly in the self-learning of the discriminator, because information that can acquire new knowledge in the enhanced training data does not exist, using the semi-supervised learning is assumed to be ineffective. In the case of sequence labeling problem, we anticipated that the outcome would be good for the diversity label combination. However, this experiment did not work well.

The effect may be caused by modifying the amount of data (100,000 sentences in this experiment) or the parameter of the threshold (0.8 in this experiment) with the pseudo-label, which is regarded as the appropriate label. Therefore, we will examine these appropriate values in the future.

In addition, adjusting the amount of loss for every word in the learning process for the LSTM model may be effective. In this experiment, we set the weights to 0 when the probability based on the confidence degree is less than 0.8, and the others were set to 1. It is considered if the set weights as probability based on the confidence degree will get more appropriate for processing self-learning processing. The question of this point also will be investigated as the future problem.

7 Conclusion

In this paper, we proposed a semi-supervised learning method using self-learning for all-words word WSD. The all-words WSD is regarded as a sequence labeling problem, so we used a bidirectional LSTM to solve it. To improve that LSTM model, we attempts semi-supervised learning for it, where self-learning is essentially used. In general self-learning is for a classification problem, not for a sequence labeling problem. To apply self-learning to our problem, the LSTM model is trained by not accumulating the loss from the low probability label. We also proposed a method to train the model with additional labeled data and then to fine-tune by using the original labeled data. As result, the precision has been improved from the precision of the model learned from only initial labeled data. This improvement is just small. Hence, our proposed method is a little effective. In the future, we will try the loss from the probability based on the confidence degree.

References

Steven Abney. 2007. *Semisupervised learning for computational linguistics*. CRC Press.

Avrim Blum and Tom Mitchell. 1998. Combining labeled and unlabeled data with co-training. In *Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100. ACM.

Jordan Boyd-Graber, David Blei, and Xiaojin Zhu. 2007. A Topic Model for Word Sense Disambiguation. In *EMNLP-CoNLL-2017*.

Fabio G Cozman, Ira Cohen, and Marcelo C Cirelo. 2003. Semi-supervised learning of mixture models. In *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*, pages 99–106.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep

bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Sachi Kato, Masayuki Asahara, and Makoto Yamazaki. 2017. Annotation of 'word list by semantic principles' information on 'balanced corpus of contemporary written japanese'. In *Processing of NLP 2017*, pages 306–309 (In Japanese).

Diederik P Kingma, Shakir Mohamed, Danilo Jimenez Rezende, and Max Welling. 2014. Semi-supervised learning with deep generative models. In *Advances in Neural Information Processing Systems*, pages 3581–3589.

Kanako Komiya, Yuto Sasaki, Hajime Morita, Minoru Sasaki, Hiroyuki Shinnou, and Yoshiyuki Kotani. 2015. Surrounding Word Sense Model for Japanese All-words Word Sense Disambiguation. In *PACLIC-29*, pages 35–43.

Kamal Nigam, Andrew McCallum, Sebastian Thrun, and Tom Mitchell. 2000. Text Classification from Labeled and Unlabeled Documents using EM. *Machine Learning*, 39(2/3):103–134.

Matthew E Peters, Waleed Ammar, Chandra Bhagavatula, and Russell Power. 2017. Semi-supervised sequence tagging with bidirectional language models. *arXiv preprint arXiv:1705.00108*.

Matthew E. Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of NAACL*.

Yanjun Qi, Pavel Kuksa, Ronan Collobert, Kunihiko Sadamasa, Koray Kavukcuoglu, and Jason Weston. 2009. Semi-Supervised Sequence Labeling with Self-Learned Features. In *Data Mining, 2009. ICDM'09. Ninth IEEE International Conference on*, pages 428–437. IEEE.

Antti Rasmus, Mathias Berglund, Mikko Honkala, Harri Valpola, and Tapani Raiko. 2015. Semi-supervised learning with ladder networks. In *Advances in Neural Information Processing Systems*, pages 3546–3554.

Salah Rifai, Yann N Dauphin, Pascal Vincent, Yoshua Bengio, and Xavier Muller. 2011. The manifold tangent classifier. In *Advances in Neural Information Processing Systems*, pages 2294–2302.

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. 2016. Improved techniques for training gans. In *Advances in Neural Information Processing Systems*, pages 2234–2242.

Hiroyuki Shinnou, Masayuki Asahara, Kanako Komiya, and Minoru Sasaki. 2017a. nwjc2vec: Word Embedding Data Constructed from NINJAL Web Japanese Corpus (in Japanese). *Natural Language Processing*, 24(5):705–720.

Hiroyuki Shinnou, Kanako Komiya, Minoru Sasaki, and Shinsuke Mori. 2017b. Japanese all-words WSD system using the Kyoto Text Analysis ToolKit. In *PACLIC-31*, pages 392–399.

Hiroyuki Shinnou, Rui Suzuki, and Kanako Komiya. 2018. All-words WSD assigning Bunruigoi ID by using Bidirectional LSTM (in Japanese). In *NINJAL Language Resources Workshop, P2-04-E*.

Rui Suzuki, Kanako Komiya, Masayuki Asahara, Minoru Sasaki, and Hiroyuki Shinnou. 2018. All-words Word Sense Disambiguation Using Concept Embeddings. In *LREC-2018*.

Koichi Tanigaki, Mitsuteru Shiba, Tatsuji Munaka, and Yoshinori Sagisaka. 2013. Density Maximization in Context-Sense Metric Space for All-words WSD. In *ACL-51*, volume 1, pages 884–893.

Xiaojin Zhu and Zoubin Ghahramani. 2002. *Learning from labeled and unlabeled data with label propagation*. Technical Report CMU-CALD-02-107, Carnegie Mellon University.

A Reinforced Improved Attention Model for Abstractive Text Summarization

Yu Chang Hang Lei Xiaoyu Li Yiming Huang
School of Information and Software Engineering
University of Electronic Science and Technology of China
Chengdu, China
realchangyu@gmail.com {hlei, xiaoyuuestc}@uestc.edu.cn yiminghwang@gmail.com

Abstract

In recent times, RNN-based sequence-to-sequence attentional models have achieved good performance on abstractive summarization. However, numerous problems regarding repetition, incoherence, and exposure bias are encountered when applying these models. In this work, we propose a novel architecture that augments the standard sequence-to-sequence attentional model and a new training method combining reinforcement learning. We evaluate our proposed method on the CNN/Daily Mail dataset. The empirical results demonstrate the superiority of our proposed method in the abstractive summarization.

1 Introduction

Abstractive text summarization is an important aspect of natural language processing (NLP), which requires the machine to automatically generate a paragraph of general content (Wang et al. 2018), such as news title summarization (Kraaij, Spitters, and Hulth 2002) and abstract summarization (Barzilay and McKeown 2005), after reading an article. Nevertheless, compared with other NLP tasks, automatic summarization exists numerous problems. For example, unlike machine translation tasks where input and output sequences often share similar lengths, summarization tasks are more likely to have input and output sequences greatly imbalanced. There are two methods to summarization: extractive and abstractive. Whereas the extraction method collects abstracts only from paragraphs (usually entire sentences) that are extracted directly from the source text (Neto, Freitas, and Kaestner 2002; Dorr, Zajic, and Schwartz 2003; Martins and Smith 2009; Berg-Kirkpatrick, Gillick, and Klein 2011; Nallapati, Zhai, and Zhou 2017), the abstract method may generate new words and phrases that are not present in the source text (Ranzato et al. 2015; Nallapati et al. 2016; See, Liu, and Manning 2017; Zhou et al. 2018; Gehrmann, Deng, and Rush 2018).

With the success of the sequence-to-sequence (seq2seq) mode (Bahdanau, Cho, and Bengio 2014; Sutskever, Vinyals, and Le 2014), it is possible to use recurrent neural networks (RNN) to read articles and generate topics. However, there are some problems with the conventional seq2seq model. First, before the start of the summary generation task, a fixed size vocabulary needs to be established, and each word of the text is replaced by its index in the vocabulary when the text is processed. However, most source articles will have out-of-vocabulary (OOV) words that are not in the vocabulary list, such as names of people, place names, scores, etc. When these words are encountered in the conventional seq2seq model, they can only be regarded as unrecognized words (UNK) (Gulcehre et al. 2016), so the output will often appear as well. Second, when generating a summary of multiple sentences, it is common to generate repeated words or sentences (See, Liu, and Manning 2017). In addition, exposure bias is problem in sequence gen-

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 362-369, Hakodate, Japan, September 13-15, 2019

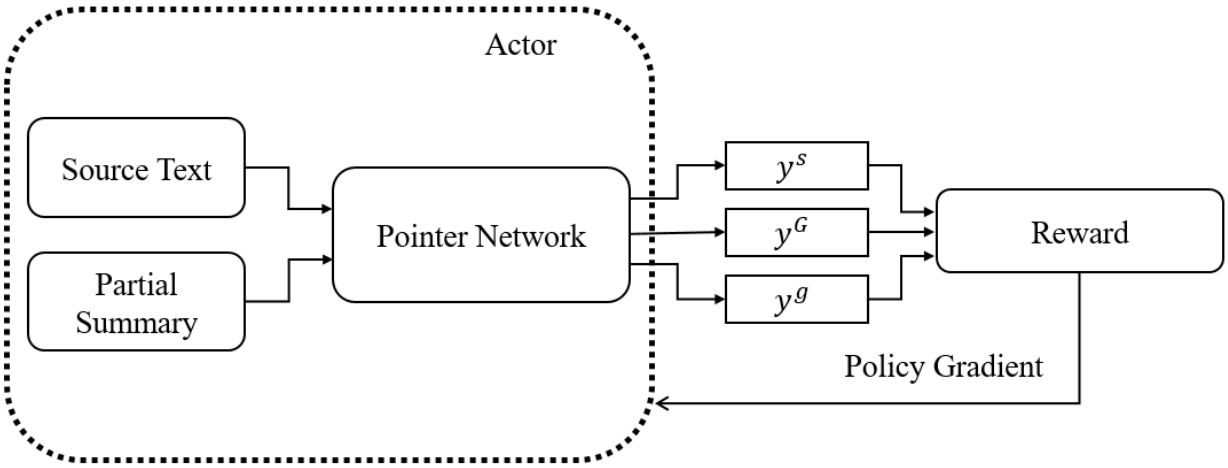


Figure 1：Model Overview. For each pointer network output distribution, a specific action y^s is sampled and the greedy action y^g is extracted and action y^G is the ground truth.

eration task (Ranzato et al. 2015).

In the model training process, each input word of the decoder uses the last word correctly output in the training sample, while in the testing stage, each input word of the decoder is its last output word, which results in the deviation between the test and training results.

The main contributions of this paper include: In this paper a new approach is proposed based on the pointer network (Vinyals, Fortunato, and Jaitly 2015) jointly with an improved attention mechanism, to solve the OOV word problem by using high attention words as candidate outputs based on different attention to the input text, that is, words in the input text can be copied into the output. Besides, the proposed model has been optimized by employing the reinforcement training. Next, We improve the temporal attention (Sankaran et al. 2016) and decoder self-attention (Paulus, Xiong, and Socher 2017). Benefiting from our approaches, repetition is reduced by storing the attention of the history input word and decoding the attention between words in different time steps. We abstract the text abstract model into a reinforcement learning model (Mnih et al. 2015). In the training process, the decoding input of each time step is the output of the previous time step, and the ROUGE score (Lin 2004) of the generated abstract and reference abstract is taken as the reward, which is solved by the policy gradient (Thomas and Brunskill 2017) to solve the exposure deviation problem.

2 Models

The symbols we will use are defined as follows: n_e represents the length of the encoder, n_d represents the length of the decoder, $x=\{x_1, x_2, \cdots x_{n_e}\}$ represents the encoder input word vector sequence, $h=\{h_1, h_2, \cdots, h_{n_e}\}$ represents the output sequence of the encoder, $s=\{s_1, s_2, \cdots, s_{n_d}\}$ represents the output sequence of the decoder, $y=\{y_1, y_2, \cdots, y_{n_d}\}$ represents the final output of the word vector sequence, $y^*=\{y_1^*, y_2^*, \cdots, y_{n_d}^*\}$ represents the ground-truth of training samples, $[a, b]$ represents the combination of a and b into one vector.

The overall structure of the model is shown in Figure 1.

The overall process of training is as follows:

(1) The input word sequence passes through the embedded layer to obtain the same length vector, which is then fed into the encoder.

(2) After encoding all the input text, the encoded information is fed to the decoder.

(3) Feed the real sample or its own output from the previous moment into the decoder (detailed in 2.4) to get the output of the current moment.

(4) Calculate temporal attention and decode self-attention to get the context vector of the encoder and decoder.

(5) Feed the context vector and decoder output into the generated and pointer network to get the output of the word.

(6) Calculate rewards based on output words and real samples, and train the entire network using the policy gradient method.

Only the first 5 steps are required for the test, and the decoder input of step 3 is the output of the previous moment.

The following is a detailed introduction to the basic structure, temporal attention, decoding self-attention, generation and pointer network, loss function and reinforcement learning.

2.1 Basic Structure and Temporal Attention

Our basic structure references (Nallapati and Xiang 2016), the encoder uses a single-layer bidirectional LSTM, consisting of a forward LSTM($LSTM^f$) and a backward LSTM ($LSTM^b$), the encoder's the i time step output $h_i = [h_i^f, h_i^b]$.

In order to prevent the generation of repetitive words, the temporal attention is introduced, that is, in each decoding time step to save attention, in the new time step to get attention divided by the sum of historical attention, weaken the previously high focus of the part, enhance the previously less attention to the part. The output of the t time step of the decoder for the attention a_t^e of each time step of the encoder is calculated as follows:

$$e_{ti} = \left(v^e\right)^T \tanh\left(W_h^e h_i + W_s^e s_t + b_{ti}^e\right)$$

$$\alpha_t^e = \begin{cases} \exp(e_t) & t = 1 \\ \dfrac{\exp(e_t)}{\exp\left(\sum_{j=1}^{t-1} e_j\right)} & \text{other} \end{cases}$$

$$a_t^e = \text{softmax}\left(\alpha_t^e\right)$$

where v^e , W_h^e , W_s^e and b_{ti}^e are learnable parameters.

In the traditional attention mechanism, historical attention is not preserved, so the calculation formula of the traditional attention mechanism a_t^e is $a_t^e = \text{softmax}\left(e_t^e\right)$.

We can get the context vector c_t^e of the encoder based on the attention of each output of the encoder at the t time step:

$$c_t^e = \sum_{i=1}^{n_e} a_{ti}^e h_i$$

2.2 Decoding Self-attention

In addition to the temporary temporal attention mechanism, we also introduce decoding self-attention in order to be able to focus on previously generated words and prevent duplication when generating new words. At the $t > 1$ time step, the decoder outputs attention to the output of the $0 < j < t$ time Step a_t^d :

$$e_{tj}^d = \left(v^d\right)^T \tanh\left(W_h^d s_j + W_s^d s_t + b_{tj}^d\right)$$

$$a_t^d = \text{softmax}\left(e_t^d\right)$$

where v^d , W_h^d , W_s^d and b_{tj}^d are learnable parameters.

At the $t = 1$ time step, the decoder context vector c_t^d is a 0 vector. When $t > 1, c_t^d$:

$$c_t^d = \sum_{k=1}^{j} a_{tk}^d s_i$$

2.3 Generate and Pointer Network

The final output word in the t time step is distributed as P_v , indicating the probability of each word being output in the word list, and is related to the context vector c_t^e of the encoder, the context vector c_t^d of the decoder, and the current output s_t of the decoder, using linear function and softmax to calculate:

$$P_v^t = \text{softmax}(W_{out}[c_t^e, c_t^d, s_t] + b_{out})$$

where W_{out} and b_{out} are learnable parameters.

However, P_v^t only decides that a word in the word list should be output. If a word in the original text is needed but not in the word list, it cannot be solved. Therefore, we use pointer network to determine whether a word should be copied based on the attention to the input word.

We define the variable P_{gen}^t to determine the probability of outputting a word based on P_v^t , then $1 - P_{gen}$ represents the probability of copying a word:

$$P_{gen}^t = \sigma\left(w_{ce}^{t\ T} c_t^e + w_{cd}^{t\ T} c_t^d + w_s^{t^T} s_t + b_{gen}^t\right)$$

where w_{ce}^t , w_{cd}^t , w_s^t and b_{gen}^t are learnable parameters, σ is the sigmoid activation function.

Combining P_v^t and pointer network, we get the probability of the final output word y :

$$P^t(y) = P_{gen}^t P_v^t(y) + (1 - P_{gen}^t) \sum_{i=1}^{n_e} a_{ti}^e (x_i = y)$$

Of course, if the word y does not exist in the word list, then $P_v^t(y) = 0$.

2.4 Loss Function and Reinforcement Learning

When training RNN to do sequence generation tasks, the most common method is teacher forcing(Williams and Zipser 1989), which trains the network at each time step of decoding with maximum likelihood estimation as the target. Maximizing likelihood estimation is equivalent to minimizing the loss function below:

$$L_{ML} = -\sum_{t=1}^{n_d} \log P(y_t^* \mid y_1^*, \cdots, y_{t-1}^*, x)$$

Firstly, using such loss function, the decoder input is real output when training, and the decoder output is its own output when testing, it will cause exposure bias. Secondly, there is a certain deviation between the target of likelihood estimation and the evaluation index (such as ROUGE), the value of loss function will decrease, but the ROUGE will increase, or vice versa.

We use reinforcement learning to solve the above two problems. For exposure bias, use the output of the decoder itself as input to the next decoder during training. For the deviation between the optimization target and the evaluation index, using the principle of reinforcement learning, the evaluation index is directly taken as the target, and the network is trained by the strategy gradient.

We use the entire network as the actor, the ROUGE-L score of the actor's output y as a reward, denoted as $R(y)$, the maximum value is 1 and the minimum value is 0. So the task target is to maximize the reward, that is, the loss function $L_{RL}(\theta)$ is the negative expectation reward:

$$L_{RL}(\theta) = -E_{y \sim P_\theta(y)}[R(y)]$$

where θ represents all trained parameters, $P_\theta(y) = P(x) \prod_{t=1}^{n_d} P_\theta(y_t \mid y_1, \cdots, y_{t-1}, x)$ represents the probability of actor output sentence y .

According to the policy gradient algorithm, we get the gradient of the loss function about θ:

$$\nabla_\theta L_{RL}(\theta) = -E_{y \sim P_\theta(y)}[R(y) \nabla_\theta \log P_\theta(y)]$$

In order to reduce the variance of the gradient, we use a policy gradient algorithm with baseline, and its loss function is as follows:

$$L_{RL} = -(R(y^s) - R(y^g)) \sum_{t=1}^{n_d} \log P(y_t^s \mid y_1^s, \cdots, y_{t-1}^s, x)$$

where y^s represents output sampled according to distributed $P(y_t^s \mid y_1^s, \cdots, y_{t-1}^s, x)$, y^g represents the output obtained according to distributed $P(y_t^g \mid y_1^g, \cdots, y_{t-1}^g, x)$ greed.

In the above formula, $R(y^g)$ is the baseline and $R(y^s)$ is the target. When both L_{ML} and L_{RL} are considered in training, the network needs to be updated separately based on two loss functions. Therefore, the storage space occupied during training (the memory used when using the GPU) is twice times the use of a single loss function.

Considering the diversity of the training samples, the output y^g is inherently quite random, so we use $R(y^g)$ as the optimization target and $R(y^s)$ as the baseline. The new modified loss function is:

$$L_{RL} = -(R(y^g) - R(y^s)) \sum_{t=1}^{n_d} \log P(y_t^g \mid y_1^g, \cdots, y_{t-1}^g, x)$$

This way, when using both loss functions, you do not need to save the intermediate parameters of output, just save the intermediate parameters that generate, and when you update, the two loss functions can be updated at the same time. Therefore, the storage space occupied during the training process is half of the previous formula and can achieve the same effect.

3 Related Work

Automatic text summarization models are usually divided into abstract models and extraction models. Early work focused on methods based on extraction and compression. From Rush (et al. 2015) for the first time to apply modern neural network to abstra-

Model	ROUGE-1	ROUGE-2	ROUGE-L
Lead-3 [See et al., 2017]	39.2	15.7	35.5
SummaRuNNer [Nallapati et al., 2017]	39.6	16.2	35.3
PointerGenerator+Coverage [See et al., 2017]	39.53	17.28	36.38
Inconsistency Loss [Hsu et al., 2018]	40.68	17.97	37.13
ML+RL [Paulus et al., 2017]	39.87	15.82	36.90
Ours			
Storing attention	37.14	15.35	34.59
Improved attention	39.57	17.15	36.83
Improved attention + RL	**40.75**	**18.03**	**38.11**

Table 1: ROUGE F1 results for various models and ablations on the CNN/Daily Mail test set.

ctive text summarization, abstract models show excellent performance. These models include the use of recurrent neural networks (RNN), where encoder and decoder are constructed using either Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber 1997) or Gated Recurrent Unit (GRU) (Cho et al. 2014), attention (Nallapati and Xiang 2016), coverage (Chen et al. 2016; See, Liu, and Manning 2017), the copy mechanism (Gu et al. 2016; See, Liu, and Manning 2017), and convolutional neural networks (CNN) (Dauphin et al. 2017; Gehring et al. 2017).

Reinforcement learning is used to optimize non-differential metrics for language generation and mitigate exposure bias. Ranzato (et al. 2015) have applied reinforcement learning to train various RNN-based sequence generation task models, which resulted in significant improvements over previous supervised learning methods. Paulus, Xiong, and Socher (2017) use reinforcement learning algorithm policy gradient methods for abstractive summarization, Rennie (et al. 2017) designed a self-critical sequence training method for image captioning tasks.

4 Experiment

For all experiments, the dimension of the word vector is 128, the pre-trained word vector is not used, such as word2vec (Mikolov et al. 2013), word vector is learned from scratch during training, the internal state of LSTM is 256 dimensions, and the word list uses 50,000 words. The optimization method uses Adagrad (Duchi, Hazan, and Singer 2011), which was found to work best of Stochastic Gradient Descent, Adadelta, Momentum, Adam and RMSProp, with a learning rate of 0.15 and an initial accumulator value of 0.1.

We use the CNN/Daily Mail dataset for training and validation, which online news articles and multiple-sentence summaries, averaging an article with 781 tokens, each article matching an average of 3.75 sentences, with an average of 56 tokens. We used scripts supplied by (Nallapati and Xiang 2016) to obtain the same version of the data, which has 287,226 training pairs, 13,368 validation pairs and 11,490 test pairs. Following (See, Liu, and Manning 2017) we choose the non-anonymized version of the dataset.

On CNN/Daily Mail dataset, we report the full-length F-1 score of the ROUGE-1, ROUGE-2 and ROUGE-L metrics (which respectively measure the word-overlap, bigram-overlap, and longest common sequence between the reference summary and the summary to be evaluated), calculated using PyRouge package. For ML+RL training, we use the ROUGE-L score as a reinforcement reward.

5 Results

Our results for the CNN/Daily Mail dataset are shown in Table 1. We compare the performance of many recent approaches with our model. Our full model scores are shown in the last line of the table. Compared with other models, we can find that there are some improvements in the scores of the three evaluation indicators. Compared with the best performing inconsistency loss (Hsu et al. 2018), our model has a slight improvement in ROUGE-1 and ROUGE-2 scores, and the ROUGE-L score is more obvious. This is due to the fact that we set ROUGE-L as reward for training.

As shown in the last four lines of Table 1, we study the ablation of our model variables to analyze the importance of each component. We use three ablation models for the experiments. The first model is just to store attention; The second model uses improved attention; And the third model is to use RL based on improved attention. By comparing the first two models, using improved attention can be 2.16 average ROUGE higher than storing attention, indicating that improved attention provides effective help to the model. Comparing the latter two models, we observe that full model outperforms by 1.11 on average ROUGE, indicating that RL has an effect on the model. Ablation studies have shown that each module is necessary for our complete model, and that improvements on all indicators are statistically significant.

6 Conclusion and Future Work

In this work, we propose an improved attention model with reinforcement learning for abstractive text summarization. We evaluate our model on CNN/Daily Mail dataset, the experimental results show that compared to previous systems our approach effectively improves performance.

Note that the model in this paper mainly uses the basic reinforcement learning algorithm. In the future, our goal is to use more advanced reinforcement learning algorithm to achieve better results.

Acknowledgments

The financial support for this work is provided by the National Science Foundation of China (No.61502082, No.61502080).

References

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Barzilay, Regina, and Kathleen R McKeown. 2005. Sentence fusion for multidocument news summarization. Computational Linguistics 31(3), 297–328.

Berg-Kirkpatrick, Taylor, Dan Gillick, and Dan Klein. 2011. Jointly learning to extract and compress, Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pp. 481–90.

Chen, Qian, Xiaodan Zhu, Zhenhua Ling, Si Wei, and Hui Jiang. 2016. Distraction-based neural networks for modeling documents. IJCAI International Joint Conference on Artificial Intelligence 2016-January, 2754–60.

Chen, Qian, Xiao-Dan Zhu, Zhen-Hua Ling, Si Wei, and Hui Jiang. 2016. Distraction-Based Neural Networks for Modeling Document. IJCAI, pp. 2754–60.

Cho, Kyunghyun, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Chopra, Sumit, Michael Auli, and Alexander M Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks, Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 93–98.

Dauphin, Yann N, Angela Fan, Michael Auli, and David Grangier. 2017. Language modeling with gated convolutional networks, Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 933–41.

Dorr, Bonnie, David Zajic, and Richard Schwartz. 2003. Hedge trimmer: A parse-and-trim approach to headline generation, Proceedings of the HLT-NAACL 03 on Text summarization workshop-Volume 5, pp. 1–8.

Duchi, J, E Hazan, and Y Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. The Journal of Machine Learning Research 12, 2121–59.

Duchi, John, Elad Hazan, and Yoram Singer. 2011. Adaptive subgradient methods for online learning and stochastic optimization. Journal of Machine Learning Research 12(Jul), 2121–59.

Gehring, Jonas, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning, Proceedings of the 34th International Conference on Machine Learning-Volume 70, pp. 1243–52.

Gehrmann, Sebastian, Yuntian Deng, and Alexander M Rush. 2018. Bottom-up abstractive summarization. arXiv preprint arXiv:1808.10792.

Gu, Jiatao, Zhengdong Lu, Hang Li, and Victor OK Li. 2016. Incorporating copying mechanism in sequence-

to-sequence learning. arXiv preprint arXiv:1603.06393.

Gulcehre, Caglar, Sungjin Ahn, Ramesh Nallapati, Bowen Zhou, and Yoshua Bengio. 2016. Pointing the unknown words. arXiv preprint arXiv:1603.08148.

Hochreiter, Sepp, and Jurgen Schmidhuber. 1997. Long short term memory. Neural computation. Neural Computation 9(8), 1735–80.

Hochreiter, Sepp, and Jürgen Schmidhuber. 1997. Long short-term memory. Neural computation 9(8), 1735–80.

Hsu, Wan-Ting, Chieh-Kai Lin, Ming-Ying Lee, Kerui Min, Jing Tang, and Min Sun. 2018. A unified model for extractive and abstractive summarization using inconsistency loss. arXiv preprint arXiv:1805.06266.

Kraaij, W, M Spitters, and A Hulth. 2002. Headline extraction based on a combination of uni- and multidocument summarization techniques. Duc 2002.

Kraaij, Wessel, Martijn Spitters, and Anette Hulth. 2002. Headline extraction based on a combination of uni-and multidocument summarization techniques, Proceedings of the ACL workshop on Automatic Summarization/Document Understanding Conference (DUC 2002). ACL.

Lin, CY. 2004. Rouge: A package for automatic evaluation of summaries, Proceedings of the workshop on text summarization branches out (WAS 2004), pp. 25–26.

Lin, Chin-Yew. 2004. Rouge: A package for automatic evaluation of summaries, Text summarization branches out, pp. 74–81.

Martins, André FT, and Noah A Smith. 2009. Summarization with a joint model for sentence extraction and compression, Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing, pp. 1–9.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

Mnih, Volodymyr, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. 2015. Human-level control through deep reinforcement learning. Nature 518, 529.

Nallapati, Ramesh, and Bing Xiang. 2016. Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond Cicero dos Santos, 280–90.

Nallapati, Ramesh, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents, Thirty-First AAAI Conference on Artificial Intelligence.

Nallapati, Ramesh, Bowen Zhou, Caglar Gulcehre, Bing Xiang, and others. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. arXiv preprint arXiv:1602.06023.

Neto, Joel Larocca, Alex A Freitas, and Celso AA Kaestner. 2002. Automatic text summarization using a machine learning approach, Brazilian Symposium on Artificial Intelligence, pp. 205–15.

Paulus, Romain, Caiming Xiong, and Richard Socher. 2017. A deep reinforced model for abstractive summarization. arXiv preprint arXiv:1705.04304.

Ranzato, Marc'Aurelio, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2015. Sequence level training with recurrent neural networks. arXiv preprint arXiv:1511.06732.

Rennie, Steven J, Etienne Marcheret, Youssef Mroueh, Jerret Ross, and Vaibhava Goel. 2017. Self-critical sequence training for image captioning, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 7008–24.

Sankaran, Baskaran, Haitao Mi, Yaser Al-Onaizan, and Abe Ittycheriah. 2016. Temporal attention model for neural machine translation. arXiv preprint arXiv:1608.02927.

See, Abigail, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. arXiv preprint arXiv:1704.04368.

Sutskever, Ilya, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks, Advances in neural information processing systems, pp. 3104–12.

Thomas, Philip S, and Emma Brunskill. 2017. Policy gradient methods for reinforcement learning with function approximation and action-dependent baselines. arXiv preprint arXiv:1706.06643.

Vinyals, Oriol, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks, Advances in Neural Information Processing Systems, pp. 2692–2700.

Wang, Li, Junlin Yao, Yunzhe Tao, Li Zhong, Wei Liu, and Qiang Du. 2018. A reinforced topic-aware convolutional sequence-to-sequence model for

abstractive text summarization. arXiv preprint arXiv:1805.03616.

Williams, Ronald J, and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. Neural computation 1(2), 270–80.

Zhou, Qingyu, Nan Yang, Furu Wei, Shaohan Huang, Ming Zhou, and Tiejun Zhao. 2018. Neural document summarization by jointly learning to score and select sentences. arXiv preprint arXiv:1807.02305.

Semantic Distance and Creativity in Linguistic Synaesthesia

Emmanuele Chersoni
The Hong Kong Polytechnic University
Chinese and Bilingual Studies
emmanuelechersoni@polyu.edu.hk

Francesca Strik Lievers
University of Genoa
Lingue e Culture Moderne
francesca.striklievers@unige.it

Chu-Ren Huang
The Hong Kong Polytechnic University
Chinese and Bilingual Studies
churen.huang@polyu.edu.hk

Abstract

In this work, we aim at quantitatively assessing the *creativity of linguistic combinations* in terms of semantic distance, using synaesthetic metaphors (e.g. *bitter voice*) as a case study. We created an evaluation dataset containing examples of synaesthesia that are actually occurring in corpora and automatically generated synaesthetic metaphors, together with a control set of non-synaesthetic adjective-noun combinations. Then, we tested on the dataset three quantitative models of linguistic creativity that have been proposed in the NLP and in the cognitive science literature, and we compared their performance in discriminating between creative and non-creative, directional and non-directional synaesthetic metaphors, and between synaesthetic metaphors and non-synaesthetic phrases.

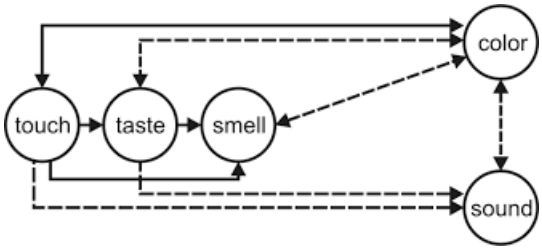


Figure 1: An example illustration of the directionality of synaesthetic transfers, graphically represented by the arrows that connect the senses (the image is taken from Werning et al., 2006), with the arrows indicating the directionality of the transfers. The details of intersensory connections may change depending on the specific language and on the specific study, but some common features are: a) transfers generally go from lower senses (touch, taste, smell) to the higher ones (sound and vision, which is sometimes identified with color, and sometimes divided between dimension and color); b) touch is the most common source and hearing the most common target; c) transfers between sound and vision are bidirectional.

1 Introduction

According to classical definitions, linguistic synaesthesia is a type of metaphor in which an experience related to a sensory modality (e.g. touch, hearing, etc.) is described through lexical means that are typically associated to a different sensory modality (Strik Lievers, 2015; Huang and Xiong, 2019). This figure is often discussed in studies on poetic and more generally literary texts (Ullmann, 1957; Shen and Cohen, 1998; Bretones-Callejas, 2001). On the one hand, synaesthesia has played an important role for literary poetics since the 19th century: see for instance the key role played by intersensory experiences in the works of symbolist poets such as Baudelaire and Rimbaud. On the other hand, research on synaesthesia as a clinical condition (Ramachandran and Hubbard, 2001; Simner and Hubbard, 2013) started a new trend of cognitively-inspired studies, putting the phenomenon in relation with the development of creative skills in the individuals.

More recent contributions focused instead on synaesthetic metaphors in ordinary language, on the basis of corpus-based analysis (see, for example, the studies of Marotta, 2012 on Italian; of Strik Lievers, 2015 on English and Italian; and of Jo, 2017; 2018 on Korean). A common point of agreement among most scholars is the observation that - both in literary and in ordinary language - synaesthetic transfers

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 370-378, Hakodate, Japan, September 13-15, 2019

are *directional* (among others, Ullmann, 1957, Shen and Cohen, 1998; for a critical discussion of this notion, see Strik Lievers, 2015; Winter, 2019a). That is, the synaesthetic transfers typically go from the "lower" senses (touch, smell and taste), which are the most common sources, to the "higher" senses (sight and sound), which are the most common targets (see Figure 1). For example, a synaesthetic metaphor like *sweet silence* is much more likely to occur than *silent sweetness* (Shen and Cohen, 1998).

In the present work, we adopt a different perspective on synaesthesia, since we are interested in the general notion of *linguistic creativity* and its quantitative assessment, as proposed in the recent cognitive science literature (Heinen and Johnson, 2017; Kenett, 2018a). [1] According to this view, the creativity of linguistic combinations can be seen as a function of *semantic distance*: the most creative combinations are those linking together concepts whose representations are far apart in the semantic memory space (Kenett, 2018a). Metaphors fit well this definition, as they typically link together concepts belonging to different conceptual domains (Lakoff and Johnson, 1980), and we believe this is the also the case for synaesthetic associations. [2] The first research question of our study is the following: *can models of semantic distance distinguish between synaesthetic metaphors and non-synaesthetic usages of sensory lexical items?* Secondly, are they able to detect *different degrees of creativity* in synaesthesia? And finally, *is semantic distance related to the directional tendency that has been observed by previous studies in synaesthesia*? In other words, can the rarity of some transfer types be explained by higher distances between concepts in the semantic memory?

Our paper is organized as follows. The computational models of creativity and semantic distance are briefly reviewed in Section 2, together with the literature on the creation of the sensory lexicon that we will use for querying synaesthetic metaphors in corpora. In Section 3, we present our procedure for generating a dataset including synaesthetic expressions with different degrees of creativity, and we describe the parameters of our experimental settings. In Section 4 we report the results of our experiments which are summarized and discussed in Section 5.

2 Related Work

2.1 Computational Models of Creativity

The traditional associative theory of creativity is probably due to Mednick (Mednick, 1962), according to which the notion involves the connection of remote, or weakly-related concepts. A common feature unifying this theory to modern research is the importance of the structure of human semantic memory in defining the distance between concepts (Kenett, 2018a). In the classic *Spreading Activation Model* (Collins and Loftus, 1975), the concepts in memory are organized in a network, and their proximity depends on their semantic similarity: concepts sharing many semantic properties will be connected by many links. Once a node in the network graph is activated, the activation spreads to all the direct neighbors, then decaying over time and space. In this model, semantic distance can be seen as the *length of the shortest path* connecting two concepts. This idea has been inspirational for some recent computational studies, which proposed to formalize the notion of semantic distance as the path length in a semantic network, and to conceive creative associations as new connections between distant nodes (Kenett, 2018b).

The most popular model for representing semantic distance in cognitive psychology is probably Latent Semantic Analysis (Landauer and Dumais, 1997). LSA models represent lexical items as vectors in a high-dimensional semantic space, on the basis of their distributional behavior in corpora. Vectors that are close correspond to semantically similar words, and the cosine between their angles is the most common similarity metric. LSA is a technique widely used also in the research field known as *Distributional Semantics* (see Lenci, 2008 for an overview), and such distributional models have been recently proposed, among the other things, to measure the semantic distance between the concepts involved in visual and linguistic metaphors (Bolognesi and Aina, 2019) and to account for the novelty and appropriateness of human noun-verb associations (Heinen and Johnson, 2017).

[1] For an overview of the different theories of linguistic creativity, see Veale (2012) and Jones (2016).

[2] Interestingly, conventional and creative metaphorical associations have been shown to activate different brain regions during sentence processing (Ahrens et al., 2007).

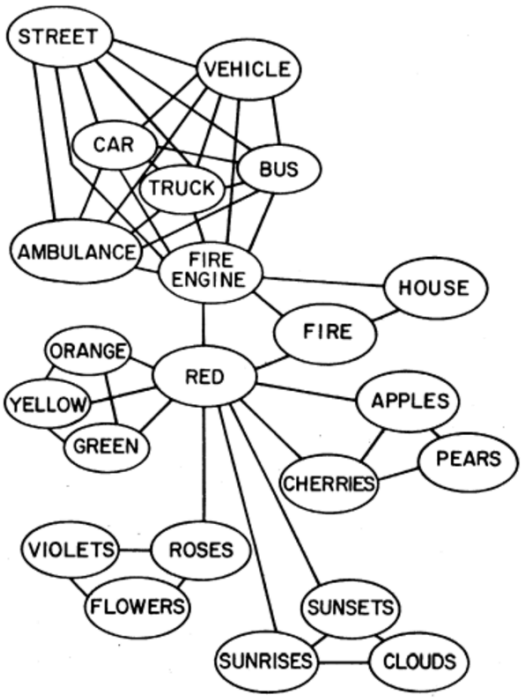


Figure 2: A schematic illustration from Collins and Loftus (1975) of the semantic memory structure. In the network graph, the shorter the line connecting the concepts, the higher their semantic relatedness.

A third metric that can be used to assess creativity, here adopted with this goal for the first time, is thematic fit (McRae and Matsuki, 2009; Lenci, 2011). Thematic fit can be described as the degree of compatibility, based on our event knowledge, between a verb and a given argument, but the notion can also be extended to adjective-noun combinations to measure the typicality of a given attribute for an entity (see the model of semantic anomaly for attributive adjective-noun pairs by Vecchi et al., 2011). In distributional models, the concept is often operationalized by means of *prototypes*: given a noun like *sound*, a representation of its typical attribute is built by averaging the vectors of the most typical co-occurring adjectives (e.g. *pleasant, nice, annoying, loud* etc.) and measuring the similarity between this prototype and a new, candidate attribute (e.g. *sweet*). Thus, for a synaesthetic combination like *sweet sound*, the degree of creativity will be an inverse function of the similarity between *sweet* and the typical adjectival modifiers of *sound*.

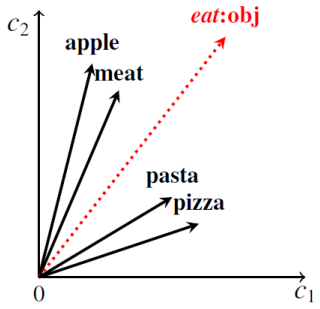


Figure 3: An illustration of a thematic fit model taken from Lenci, 2017 (oral presentation in Barcelona). In this case, the thematic fit for potential objects of *to eat* would be measured as their similarity with the prototype vector of the patient of the verb (in red).

2.2 Sensory Lexica for the Identification of Synaesthesic Metaphors

Two important requirements for building our evaluation dataset for synaesthesia are:

1. a methodology for automatically extracting synaesthetic metaphor candidates from corpora;

2. a list of words associated to different senses and annotated for part-of-speech, typically nouns and adjectives since noun phrases with an adjectival modifier are the most common form of synaesthesia [3] .

Adjective-noun combinations are the form of synaesthetic metaphor on which we will be focusing in the current study. As for the methodology, we adopt the dependency-based search proposed by Strik Lievers and Huang, 2016: given a parsed corpus, we look for all the adjective-noun phrases in which the adjective and the noun are typically associated with different sensory modalities.

Concerning the sensory lexicon, the first resource made available were probably the norms collected by Lynott and Connell (Lynott and Connell, 2009; Lynott and Connell, 2013), in which the association between words and sensory modality were generated on the basis of human ratings. An interesting

[3] For a systematic study on the distribution of lexical categories across different sensory modalities in the English sensory lexicon, see Strik Lievers and Winter, 2018.

feature of these datasets, respectively including 423 adjectives and 400 nouns, is that they also provide ratings reflecting to what extent each word tends to be associated with a single sensory modality. [4]

Another important resource is the sensory lexicon that has been built in a semi-automatic way by Tekiroglu et al., 2014, by using a list of words extracted from WordNet and expanded by means of NPMI association scores (Bouma, 2009). In terms of size, Sensicon is by far the biggest sensory dataset currently available, with associations for more that 22,000 words. However, being a large semi-automatically built resources, association scores have not been manually checked, and therefore it contains noisy data.

Finally, a wordlist annotated with sense associations was manually compiled by Strik Lievers, 2015, for a corpus-study on synaesthetic metaphors. Since the item selection was more controlled, we chose to use this wordlist as a reference for extracting a list of synaesthetic metaphors from corpora. However, for the purposes of our evalution we also wanted to generate a second list of *creative* synaesthetic metaphors that are unlikely to be found in corpora. For this task, we decided to rely on Lynott and Connell's dataset, which comes with "monoesthesia" scores: the idea is that, the stronger the association of a word with a single sense, the less likely it will be that it enters into a synaesthetic association.

3 Experiments

3.1 Dataset Creation

The first step for us was to extract a first set of synaesthetic metaphor candidates from a parsed corpus: we chose the British National Corpus, since it is a balanced corpus containing a wide variety of textual genres (Leech, 1992). As we said, we only took into account adjective-noun phrases, and we used as a seed set the nouns and the adjectives manually annotated by Strik Lievers, 2015. The set includes 119 nouns (13 for smell, 22 for taste, 5 for touch, 59 for hearing, 20 for sight) and 190 adjectives (10 for smell, 28 for taste, 43 for touch, 30 for hearing, 79 for sight).

Source / Target	Sight	Sound	Taste	Smell	Touch
Sight	-	272	**16**	**13**	**23**
Sound	22	-	**3**	**1**	-
Taste	20	89	-	233	**1**
Smell	-	-	-	-	-
Touch	174	476	98	130	-

Table 1: Summary of the transfers types for the candidate synaesthetic metaphors from the BNC. For each cell, the sense in the row is the source, while the sense in the column is the target. In bold, the transfers that contradict the directionality principle of the classical sense hierarchies.

As a result of the extraction, we obtained 1571 occurrences of synaesthetic combinations from the BNC (after manually filtering out some noise, mainly due to metonymic expressions such as *black music*). The different types of transfer between senses are shown in Table 1: at a glance, it is clear that they mostly follow the principle of directionality of classical sense hierarchies, with some exceptions (in most cases, adjective-noun phrases with sight as the source sensory modality). As for the total number of synaesthetic *types*, we found 471 of them: we call this set SYN. That is, the SYN set includes types of **synaesthetic metaphors that occur in the British National Corpus**. Secondly, we generated two other sets of phrases, to be used for comparison with our newly-found synaesthetic metaphors:

- **control collocations**: for all words in the synaesthetic expressions of the SYN set, we aimed at extracting a common collocate. This means that, for nouns, we generated new adjective-noun phrases by combining nouns with their most typical adjectival modifiers (e.g. for *colour*: *bright colour*, *dark colour*, etc.). For adjectives, we did the same by combining them with the nouns that they typically modify (e.g. for *bitter*: *bitter disappointment*, *bitter taste*, etc.);

- **new synaesthetic metaphors**: in order to create new, creative synaesthetic metaphors, we adopted the view that the more creative lexical associations are those linking concepts that are very distant. Thus, in the case of synaesthesia, they are more likely to involve concepts

[4] We recently became aware that Lynott and colleagues have released a new and bigger sensory norms dataset on Psyarxiv in May 2019 (Lynott et al., 2019), including almost 40,000 words: it was unfortunately too late to use it for the present studies, but it will certainly be an important resource for future works.

that typically do not co-occur with more than one sense. By using the word-sense association scores of the datasets by Lynott and Connell, we randomly combined adjectives and nouns with very high degrees of monoestheticity (i.e. a score quantifying the tendency of being associated with a single sense only).

For the control collocations, the typical collocates have been extracted on the basis of Positive Local Mutual Information ($PLMI$) (Evert, 2004), measured from the co-occurrences of adjectives and nouns in the Wacky corpus (Baroni et al., 2009). This metric can be seen as a slightly modified version of the more common PMI (Church and Hanks, 1990), less biased towards rare events.

Given an adjective adj and a noun n, the $PLMI$ is computed as follows:

$$LMI(adj, n) = log\left(\frac{f_{adj,n} * C}{f_{adj} * f_n}\right) * f_{adj,n} \quad (1)$$

$$PLMI(adj, n) = max(LMI(adj, n), 0) \quad (2)$$

where f_{adj} and f_n are the respective frequencies of adj and n, $f_{adj,n}$ is the frequency of their joint co-occurrence, and C the number of observed word pairs in the corpus. In other words, the $PLMI$ measures the statistical association between adjectives and nouns by comparing their observed co-occurrence with the expected co-occurrence under the assumption of statistical independence between the two. Each adjective and noun of the Strik Lievers list has been combined with the top $PLMI$-scoring word for the other POS, to create examples of the usage of those words in "standard", non-figurative expressions. Notice that this methodology does not guarantee that collocates will be retrieved for all words appearing in the synaesthetic metaphors, as some of them might be rare in Wacky and/or might not have collocates with enough statistical association strength. Out of the 156 different words composing the expressions in SYN, we could retrieve collocates only for 132 of them to form the adjective-noun phrases of the CONTROL set.

Finally, for the set of the new synaesthetic metaphors, we adopted the following procedure: for each adjective or noun in the set SYN, we generated a new combination with a word of the other category included in the data by Lynott and Connell, provided that it has a monoestheticity score equal or superior to 6.[5] Then, we randomly sampled 471 of these expressions, in order to have a set of the same size of SYN. We refer to this new set as NEW_SYN.

3.2 Models and Experimental Settings

The three models of creativity that we will test in this study are *Latent Semantic Analysis*, *Thematic Fit* and *shortest path length*. All of them have been trained on the Wacky corpus (Baroni et al., 2009)

The models represent as targets all the words included in our datasets. As contexts for the target words, we use the 30,000 more frequent words in the Wacky corpus (only considering nouns, verbs and adjectives).

We trained LSA models in two different versions: an unweighted version, and a version where co-occurrences between targets and contexts have been weighted via Positive Pointwise Mutual Information ($PPMI$).

$$PMI(adj, n) = log\left(\frac{f_{adj,n} * C}{f_{adj} * f_n}\right) \quad (3)$$

$$PPMI(adj, n) = max(PMI(adj, n), 0) \quad (4)$$

We refer to these two versions of the model as LSA and LSA_PPMI. For both of them, we reduced the word-context matrix by setting the parameter of the SVD components to 300 (the only difference being that, with LSA_PPMI, the frequencies are weighted before the dimensionality reduction step).

The thematic fit models use dependency-based contexts: that is, each dimension of the semantic space is a combination of 1) one of the 30,000 words; and 2) a syntactic dependency relation linking the word with the target (for example, *ADJ_MOD:loud* might be a context for the target

[5] The threshold was empirically selected: with a higher threshold, the number of candidate words for creating new combinations was too small.

noise). We trained two different thematic fit models: one assigning a score to an adjective-noun phrase by measuring the similarity between the adjective and the prototype of the noun modifiers; and the other measuring the similarity between the noun and the prototype of the nouns modified by the adjective. We call these models $TFIT\_ADJ$ and $TFIT\_N$.

For building these prototypes, we averaged the vectors of the 10 most strongly $PLMI$-associated collocates for each relation (that is, the 10 adjectives with the highest $PLMI$ association score as modifiers of the noun, and the 10 nouns having the highest $PLMI$ association score with the adjective as a modifier). In case one of the top nouns or adjectives had been previously used for generating the CONTROL set, it was excluded from the list for generating the prototype.

As for the shortest path length model ($PATH$), we use the same $PPMI$ matrix of the weighted LSA model to generate an undirected graph. In this graph, the nodes correspond to the target words. Two words are linked by an edge if they co-occur and their $PPMI$ score is $>= 0$. The weight for each edge is equal to 1 divided by the $PPMI$ score for the two words, which means, edges connecting strongly associated words will have a lower traveling cost. In order to compute the scores for this model, for each adjective-noun phrase in our dataset we computed the shortest path length between the adjective and the noun in the graph, by using the classical Dijkstra algorithm (Dijkstra, 1959).

The output of each model will be a score of the semantic similarity between the two words in the adjective-noun phrase. As we explained in the introductory section, the more creative combinations are those linking together more distant concepts. Thus, the scores of the model have to be read as *inverse indexes of creativity*: the more two items are similar, the lower their semantic distance. If a model is doing well, we expect the scores to pattern in the following way: 1) the CONTROL items get the highest scores; 2) the NEW_SYN ones get the lowest scores and 3) SYN items score in the middle.

4 Results

A quick check of the Spearman correlations between the different metrics reveals that they are weakly correlated, the only exceptions being LSA and LSA_PPMI, with $\rho = 0.48$ (as expected, as they are just different versions of the same metric), and LSA_PPMI and $TFIT\_N$, with $\rho = 0.37$.

After computing the scores for all the four models, we ran statistical tests to see how good are the models in discriminating between the different experimental conditions. According to the Kruskal-Wallis test, all models can find a highly significant difference between conditions (for all of them, $p < 0.001$): the boxplots for $TFIT\_ADJ$ and LSA_PPMI are shown, as an example, in Figures 4 and 5.

We came then to the post-hoc tests: for the Wilcoxon rank sum test, again all models find a significant difference between CONTROL and NEW_SYN (for all of them, $p < 0.001$), between CONTROL and SYN ($p < 0.01$), and between SYN and NEW_SYN ($p < 0.01$). This is an interesting result: we expected the models to be able to easily discriminate between CONTROL and the other two conditions, but given the rarity of synaesthetic metaphors, it is surprising that they also manage to distinguish between those that were actually found in the BNC and the automatically generated ones. We should recall here that the latter ones were generated in order to be more "creative", by combining the words of the SYN set with words i) of a different sensory modality and ii) that are typically not associated with words of a different sensory modality. From this point of view, all models did a good job in recognizing different levels of creativity between the two sets of synaesthetic combinations.

Interestingly, there was a partial exception in our results: the simple $TFIT\_N$ model only found a marginally significant difference between NEW_SYN and SYN ($W = 70619$, $p < 0.05$), while the difference for the $TFIT\_ADJ$ was much larger ($W = 14368$, $p < 0.001$). Thus, the model based on the similarity between the actual and the prototypical modifier of the noun seems to have a higher discriminative power.

To address the last question of our exploratory study, i.e. whether semantic distance is related to directionality observed in the literature on synaesthetic metaphors, we assigned a class to each of the expressions in the SYN and in the NEW_SYN set, depending on the type of synaesthetic transfer: we

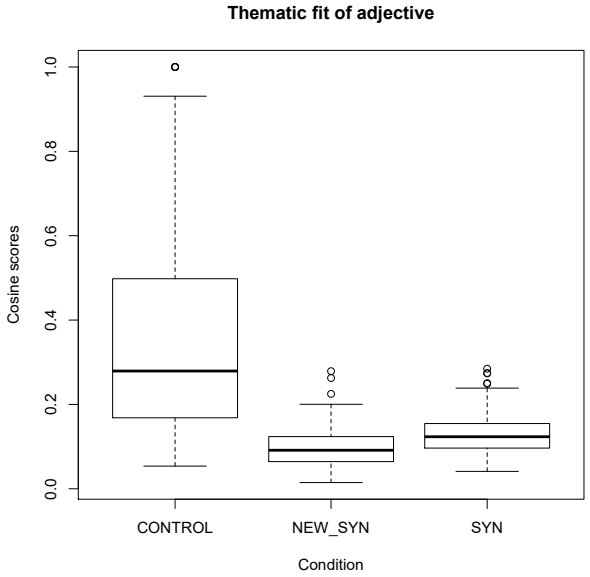


Figure 4: Cosine similarity scores assigned by $TFIT\_ADJ$ to the items in the three conditions.

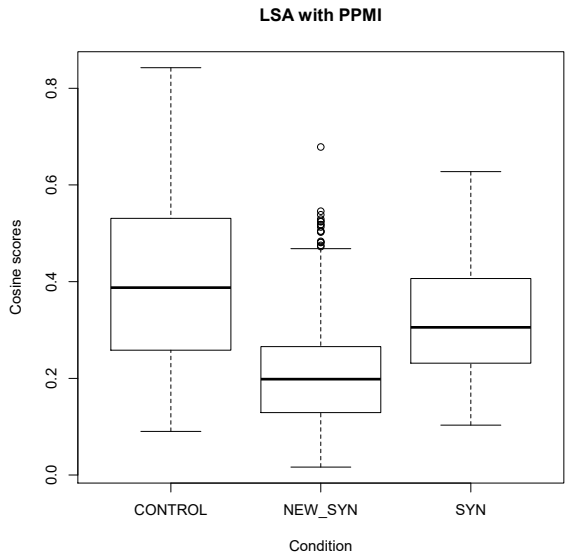


Figure 5: Cosine similarity scores assigned by LSA_PPMI to the items in the three conditions.

assigned a positive label to those expressions that are coherent with the directionality (e.g. pairs with a *taste* adjective and a *sight* noun) and a negative label to those that are not (e.g. pairs with a *sound* adjective and a *smell* noun). In total, we had 834 coherent and 108 non-coherent transfers. Again, we test our models in their ability of discriminating between the conditions, in order to check if semantic distance metrics confirm that transfers not respecting the common directionality are more creative. In this case, semantic distance scores assigned to the pairs with a negative label should be significantly higher (i.e. lower cosines for the distributional models, longer paths for the $PATH$ one).

It turns out that most models are not able to make the distinction: for the Wilcoxon rank sum test, $PATH$ ($W = 40450$, $p > 0.05$) and LSA ($W = 45644$, $p > 0.05$) failed to find a significant difference between directional and non-directional combinations. At a closer inspection, we found that a highly significant difference is found by LSA_PPMI ($W = 53937$, $p < 0.001$), but not in the expected direction: the cosines for the non-directional combinations are significantly higher, instead of being lower. On the other hand, thematic fit models struggle for the low coverage: out of the 108 items in the non-coherent set, only 40 take non-zero values for $TFIT\_ADJ$ and only 60 for $TFIT\_N$. By running the Wilcoxon test on the remaining phrases, both models fail to find a significant difference ($p > 0.1$). It should be noticed that we used a classical version of the thematic fit model, based on dependencies (Baroni and Lenci, 2010), which suffers by definition of more sparsity. If two vectors do not share any dependency-based context, their similarity will be zero, and from this point of view, the result could make sense, since the adjective-noun combinations of the non-coherent phrases are extremely unlikely. Actually, by taking into account also the phrases with a similarity score of zero, both $TFIT\_ADJ$ and $TFIT\_N$ assign significantly lower scores to the non-coherent combinations ($p < 0.01$ for both of them).

5 Conclusion

In this study, we tested three models of semantic distance to assess the creativity of linguistic

combinations, taking over the task of distinguishing between synaesthetic metaphors and control expressions, and between the former and some automatically-generated, more creative synaesthetic combinations. We found that all models are able to find significant differences and to properly distinguish between the conditions.

Then, we also tested our models on the task of distinguishing between those combinations that are consistent with the directionality tendency (from the lower to the higher senses) observed in the studies on synaesthesia, and those that are not. We found that this task is much more difficult: thematic fit models might be the closest to identify this distinction, as their similarity assessment is based on the direction of the dependency relation between the adjective and the noun. Thus, they could implicitly incorporate some notion of the directionality of synaesthetic metaphors (i.e. what are the typical source and target domains). On the other hand, they suffer from data sparsity: most of the non-coherent phrases of our dataset got assigned a similarity score of zero, and we found a significant difference between the two conditions only by including these latter phrases.

To the best of our knowledge, this is the first study in computational modeling on the topic of linguistic synaesthesia, and the first trying to account for its combinatory patterns in terms of semantic distance. In a recent contribution, Jo (2018) pointed out that there was almost no connection between the corpus- and the cognitive science-oriented perspectives of research on the phenomenon. We believe that the notion of semantic distance, seen as a possible factor influencing the likelihood of sensory words combinations as observed in natural language corpora, could provide a link between these two trends of studies. On the one hand, the notion has been proposed by the modern research in cognitive science, but on the other hand it can be modeled in a straightforward way by means of corpus-based models of meaning.

Some promising models for our future tests include a thematic fit model based on dense spaces, in order to overcome the sparsity problem, or shortest path length models based on directed graphs, which should also be better at modeling the directionality of synaesthetic metaphors. Another possible direction is repeating the experiments with other languages for which large-scale modality exclusivity norms have been made available, such as Mandarin Chinese (Chen et al., 2019) and Italian (Morucci et al., 2019).

References

Kathleen Ahrens, Ho-Ling Liu, Chia-Ying Lee, Shu-Ping Gong, Shin-Yi Fang, and Yuan-Yu Hsu. 2007. Functional MRI of Conventional and Anomalous Metaphors in Mandarin Chinese. In *Brain and Language*, vol. 100, no. 2, pp. 163-171. Elsevier.

Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora. In *Language, Resources and Evaluation*, vol. 43, no. 3, pp. 209-226. Springer.

Marco Baroni and Alessandro Lenci. 2010. Distributional Memory: A General Framework for Corpus-Based Semantics. In *Computational Linguistics*, vol. 36, no. 4, pp. 673-721.

Marianna Bolognesi and Laura Aina. 2019. Similarity Is Closeness: Using Distributional Semantic Spaces to Model Similarity in Visual and Linguistic Metaphors. In *Corpus Linguistics and Linguistic Theory*, vol. 15, no. 1, pp. 101-137. De Gruyter.

Gerlof Bouma. 2009. Normalized Pointwise Mutual Information in Collocation Extraction. In *Proceedings of GSCL.*

Carmen Bretones-Callejas. 2001. Synaesthetic Metaphors in English. Technical Reports. University of California at Berkeley and International Computer Science Institute.

I-Hsuan Chen, Qingqing Zhao, Yunfei Long, Qin Lu, and Chu-Ren Huang. 2019. Mandarin Chinese Modality Exclusivity Norms. In *PLOS One*, vol. 14, no. 2.

Kenneth Ward Church and Patrick Hanks. 1990. Word Association Norms, Mutual Information, and Lexicography. In *Computational Linguistics*, vol. 16, no. 1. MIT Press.

Allan Collins and Elizabeth Loftus. 1975. A Spreading-Activation Theory of Semantic Processing. In *Psychological Review*, vol. 82, no. 6. American Psychological Association.

Edsger W Dijkstra. 1959. A Note on Two Problems in Connexion with Graphs. In *Numerische Mathematik*, vol. 1, no. 1, pp. 269–271. Springer.

Stefan Evert. 2004. The Statistics of Word Cooccurrences: Word Pairs and Collocations. PhD Thesis, University of Stuttgart.

David JP Heinen and Dan R Johnson. 2017. Semantic Distance: An Automated Measure of Creativity That Is Novel and Appropriate. *Psychology of Aesthetics*

Creativity and the Arts. Amerian Psychological Association.

Chu-Ren Huang and Jiajuan Xiong. 2019. Linguistic Synaesthesia: An Introduction. *The Routledge Handbook of Chinese Applied Linguistics*. Routledge.

Charmhun Jo. 2017. A Corpus-Based Study on Synesthesia in Korean Ordinary Language. In *Proceedings of PACLIC*.

Charmhun Jo. 2018. Synaesthetic Metaphors in Korean Compound Words. In *Proceedings of the LREC Workshop on Linguistic and Neurocognitive Resources*.

Rodney H Jones. 2016. The Routledge Handbook of Language and Creativity. Routledge.

Yoed N Kenett. 2018. What Can Quantitative Measures of Semantic Distance Tell Us About Creativity? In *Current Opinions in Behavioral Sciences*, volume 27, pp. 11–16.

Yoed N Kenett. 2018. Investigating Creativity from a Semantic Network Perspective. In *Exploring Transdisciplinarity in Art and Science*, pp. 49-75. Springer.

George Lakoff and Mark Johnson. 1980. Metaphors We Live By. University of Chicago Press.

Thomas K Landauer and Susan Dumais. 1997. A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. In *Psychological Review*, vol. 104, no. 2. American Psychological Association.

Geoffrey Neil Leech. 1992. 100 Million Words of English: The British National Corpus (BNC).

Alessandro Lenci. 2008. Distributional Semantics in Linguistic and Cognitive Research. In *Italian Journal of Linguistics*, vol. 20, no. 1, pp. 1–31.

Alessandro Lenci. 2011. Composing and Updating Verb Argument Expectations: A Distributional Semantic Model. In *Proceedings of the ACL Workshop on Cognitive Modeling and Computational Linguistics*.

Dermot Lynott and Louise Connell. 2009. Modality Exclusivity Norms for 423 Object Properties. In *Behavior Research Methods*, vol. 41, no. 2, pp. 558–564. Springer.

Dermot Lynott and Louise Connell. 2013. Modality Exclusivity Norms for 400 Nouns: The Relationship Between Perceptual Experience and Surface Word Form. In *Behavior Research Methods*, vol. 45, no. 2, pp. 516–526. Springer.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2019. Lancaster Sensorimotor Norms. In *PsyArxiv*.

Giovanna Marotta. 2012. Sinestesie tra Vista, Udito e Dintorni. Un'Analisi Semantica Distribuzionale. In *Sinestesie e Monoestesie*, edited by Marina Catricalà. Milano: Franco Angeli.

Ken McRae and Kazunaga Matsuki. 2009. People Use Their Common Knowledge to Understand Language, and Do So as Quickly as Possible. In *Language and Linguistics Compass*, vol. 3, no. 6, pp. 1417–1429.

Sarnoff Mednick. 1962. The Associative Basis of the Creative Process. In *Psychological Review*, vol. 69: pp. 220-232. American Psychological Association.

Piermatteo Morucci, Roberto Bottini, and Davide Crepaldi. 2019. Augmented Modality Exclusivity Norms for Concrete and Abstract Italian Property Words. In *OSF Preprints*.

Vilayanur S Ramachandran and Edward Hubbard. 2001. Synaesthesia–A Window into Perception, Thought and Language. In *Journal of Consciousness Studies*, vol. 8, no. 2, pp. 3-34.

Yeshayahu Shen and Michael Cohen. 1998. How Come Silence Is Sweet but Sweetness Is Not Silent: A Cognitive Account of Directionality in Poetic Synaesthesia. In *Language and Literature*, volume 7, no. 2, pp. 123–140. Sage Publications Sage CA: Thousand Oaks, CA.

Francesca Strik Lievers. 2015. Synaesthesia: A Corpus-Based Study of Cross-Modal Directionality. In *Functions of Language*, volume 22, no. 1, pp. 69–95. John Benjamins.

Francesca Strik Lievers and Chu-Ren Huang. 2016. A Lexicon of Perception for the Identification of Synaesthetic Metaphors in Corpora. In *Proceedings of LREC*.

Francesca Strik Lievers and Bodo Winter. 2018. Sensory Language Across Lexical Categories In *Lingua*, pp. 45–61. Elsevier.

Julia Simner and Edward Hubbard. 2013. Oxford Handbook of Synaesthesia. Oxford University Press.

Serra Sinem Tekiroglu, Gözde Özbal, and Carlo Strapparava. 2014. Sensicon: An Automatically Constructed Sensorial Lexicon. In *Proceedings of EMNLP*.

Stephen Ullmann. 1957. The Principles of Semantics. Glasgow: Jackson.

Tony Veale. 2012. Exploding the Creativity Myth: The Computational Foundations of Linguistic Creativity. A&C Black.

Eva Maria Vecchi, Marco Baroni, and Roberto Zamparelli. 2011. (Linear) Maps of the Impossible: Capturing Semantic Anomalies in Distributional Space In *Proceedings of the ACL Workshop on Distributional Spaces and Compositionality*.

Markus Werning, Jens Fleischhauer, and Hakan Beseoglu. 2006. The Cognitive Accessibility of Synaesthetic Metaphors. In *Proceedings of CogSci*.

Bodo Winter. 2019. Sensory Linguistics. Amsterdam: Jonh Benjamins.

Bodo Winter. 2019. Synaesthetic Metaphors Are Neither Synaesthetic Nor Metaphorical. In *Perception Metaphors*. Amsterdam: Jonh Benjamins, pp. 105–126.

Investigating Mandarin Negative Terms: An Evaluation of Semantic-Pragmatic Meanings and Metaphorical Mechanisms

Siaw-Fong Chung
Department of English
National Chengchi University
National Chengchi University
No. 64, ZhiNan Rd. Sec. 2, Taipei
sfchung@nccu.edu.tw

Yi-Ling Tseng
Master's Program in Teaching
Chinese as a Second Language
National Chengchi University
No. 64, ZhiNan Rd. Sec. 2, Taipei
105161014@nccu.edu.tw

Heng-Chia Liao
Master's Program in Teaching
Chinese as a Second Language
National Chengchi University
No. 64, ZhiNan Rd. Sec. 2, Taipei
106161011@nccu.edu.tw

Man-Hua Huang
Master's Program in Teaching
Chinese as a Second Language
National Chengchi University
No. 64, ZhiNan Rd. Sec. 2, Taipei
106161010@nccu.edu.tw

Abstract

We collected lists of negative terms from past literature and from dictionary searches. From the lists, we selected thirteen negative terms for further observation. These terms were selected because they are not entirely negative although they were categorized as negative terms by others. By analyzing each instance of these terms in the news articles within ten years, we found the proportions of positive, negative, and neutral uses of these terms. Each term has various degrees of positive or negative uses. These different semantic prosodies were explained using metaphorical mechanisms we have established for each item. This study is one of the few studies that look into the significance of negative terms in discourse. Most previous studies aimed to identify or ascribe a label once a term is found. The look into the semantic and pragmatic uses of negative terms is rare, but needed.

1 Introduction

Negative language could be seen from several perspectives. One of the ways is through identifying negative terms, which is often accompanied by negative evaluation. For evaluation, Hunston (2011:23) proposed a "three-move evaluation act", namely:

(1) (a) "identification and classification of an object to be evaluated";
(b) "ascribing a value to that object"; and
(c) "identifying the significance of the information".

The object being evaluated could be any forms, but in Hunston's term, it refers to "discoursal and epistemic […] forms of propositions", meaning that the evaluation is found in discourse.

The study of negative terms has many applicational uses. For example, 彭宣維 et al. (2011; 2015) created a database called the *Chinese-English Parallel Corpus of Appraisal Meanings* (漢英對應評價意義語料庫) which provided a tagged version of data according to Martin and White's (2005) Appraisal Theory. The theory, although included only the word 'appraisal', has both appraisal and non-appraisal sides. The identification of these appraisal terms fell mostly in the first (or maybe second) step of Hunston's three moves. In addition to the list, there is also the *Taiwan Corpora of Chinese Emotions and Relevant Psychophysiological Data* (臺灣地區華人情緒與相關心理生理資料庫) (EemotioNeT).[1]

[1] http://ssnre.psy.ntu.edu.tw (requires application to use the database)

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 379-387, Hakodate, Japan, September 13-15, 2019

This database was used mainly for psychophysiological research. For instance, 陳學志、詹雨臻 and 馮彥茹 (2013), as well as 卓淑玲、陳學志 and 鄭昭明(2013) used the database as a reference for experimental material designs.

Although many intended to create a database of negative terms or of emotive language, most of these studies did not go beyond the second moves – i.e., to identify the significance of the information in discourse. Ascribing and classifying a value to a term (step two) seems possible in both databases mentioned above. However, interpreting the significance of the data was not the main concerns when creating the databases. For instance, interpreting why a certain semantic prosody exists in a negative term requires contextual information. Instead of ascribing a single label to a lexical item, we looked into their possible negative and non-negative meanings in different occurrences. More importantly, we tried to establish the relation between the negative and non-negative meanings of the same term.

In this paper, we will look at words with negative meaning and evaluate their negative meanings by examining their use in news articles. We selected negative terms that could have both negative and non-negative meanings, and examined how they were presented in news discourse. We also analyzed the metaphorical extensions of these negative terms in order to understand the connotation difference carried by a similar term. We will answer the following research questions:

(2) (a) To what extent a potentially negative term will carry non-negative meanings?
(b) How can we explain the different semantic prosodies possessed by a similar item?

By semantic prosody, we refer to the two main two issues brought up by Hunston (2007:265): "the discourse function of an extended unit of meaning, and the attitudinal meanings". For both, we looked into the actual occurrences of the selected negative terms in discourse. In the next section, we provide the literature review of the current work.

2 Negative Words

The discussion of negativity in literature often surrounded 'negative-polarity items' (NPIs) such that it is unacceptable to say a positive-polarity item in a negative sentence (**I'm not pretty please with it*) but it is acceptable to say it in a positive sentence (*I am pretty pleased with it*) (Linebarger, 1980: 7). Such a discussion is not what we intend to pursue in this paper.

We are interested in the semantic and pragmatic use of the negative terms in discourse. In one of the studies we found, 傅奇珅 (2010) examined negative words in news reports about a particular company in Taiwan and computed the number of positive and negative words used in the reports. The determination of positive and negative words was based on whether or not good things or bad things were said about the company. Following 郭先珍 (1996), 傅奇珅's (2010:11) definitions of positive and negative terms are as follows (translated by the current authors):

(3) (a) Positive terms: Terms that contain appraisal (讚許), and confirmation emotions (肯定感情);
(b) Negative terms: Terms that contain demotion (貶斥), denial (否定), resentment(憎恨), disdain (輕蔑) emotions

郭先珍's series of work has become a major reference for many studies as they are the few sources that provided the explicit lists of positive and negative keywords.

In one study, Giora et al. (2004) used a scale of 1 to 7 to test the meaning of positive and negative of adjectives in a continuum: *ugly* (negative), *not pretty* (negative positive), *fairly pretty* (hedged positive), and *pretty* (positive). The salient meaning of a negated concept would not be eradicated by the negation marker (*not*), but be mitigated to a certain degree. The researchers thus suggested that negation markers should be classified as modifiers instead of suppressors. They also tested the pragmatic function of negation. Participants were shown 'nonnegatived negative item (*What you said was a lie*) versus *What you said was not true* and were asked to decide which sentence they would prefer to be polite. The result showed that negated semantically positive element was preferred when people wanted to show politeness or express undesirable state, while non-negated semantically negative constituent was avoided. The finding of this work was among the few that examined the relation between the negative and positive elements of a sentence in a discourse.

In one other study, Xiao and McEnery (2006:113-114) compared the semantic prosody of *result* and the Mandarin equivalent *jie2guo3* (結果). The authors found that "the six near synonyms of *jie2guo3* in Chinese can be arranged on a semantic continuum, from positive to negative, as follows: *shuo4guo3* (碩果), *cheng2guo3* (成果), *jie2guo3* (結果), *hou4guo3* (後果), and *ku3guo3* (苦果)/*e4guo3* (惡果)." Compared to English, these words are divided clearly into negative and positive meanings, with the first two being more positive, and the latter two being negative. The middle *jie2guo3* (結果) seems more neutral.

As there is a continuum between positive and negative meanings, Sobieraj and Berry (2011), on the other hand, found that insulting words could mean different things in different situations. Words such as *idiotic* or *pompous*, when referring to a person or group's behavior, are part of insulting language. However, if these words are used to call person or group, they are name-calling, and thus, a personal attack. Therefore, there is a fine line between evaluation and personal attack. Although not mentioned in Sobieraj and Berry, some negative words could be used positively. For instance, the following example is taken from the *Corpus of Contemporary American English*.[2]

(4) *His face immediately broke into a bright* ***idiotic*** *smile.*

Bednarek (2008:130) believed that "it is important to distinguish between the nature of collocates (negative/positive collocates) and the connotation of a lexical item (negative/positive prosody)." In this case, the word *idiotic* is in itself negative but it could have a non-negative meaning if it modifies a word such as *smile*. It would mean 'foolish' but not 'stupid'. In this paper, our aims are to observe some selected negative words to see if different degrees of differentiation can be found in the uses of the negative terms.

3 Metaphorical Extensions

In literature, one of the ways to analyze metaphors is by the use of source and target domains. The main reference of this model is Lakoff and Johnson's (1980:10) Conceptual Metaphor Theory, in that it mentions that, in metaphorical mappings, highlighting and hiding are parts of the "systematicity" of metaphors. In metaphor creation, the 'highlighting aspects' become the foci of a metaphorical concept, whereas the 'hiding aspects' are "other aspects of the concept that are inconsistent with that metaphor." For instance, the metaphorical concept of *speed* in TIME (source) could not be mapped onto the metaphorical concept of *budget* in MONEY (target). Therefore, unrelated concepts are not usually mapped, according to this principle. Only concepts with related variance will be mapped. From this, examples such as *you're* ***wasting*** *my time* and *this gadget will* ***save*** *you hours*, in which both have a feature that is valuable, especially that of a certain commodity, are mapped between TIME and MONEY (Lakoff and Johnson, 1980:8).

Therefore, for metaphorical extensions that occur within a word, a certain kind of metaphor mapping should be observed. Many people have looked at source and target domains from mapped domains, but we would like to apply this to look at metaphorical meanings that may derive from a source meaning, especially when the meanings could change from positive to negative, and vice versa. Murphy (2003) discussed the relations between semantics and lexicon; in this paper, the relations between the co-existence of positive and negative senses, if any, will be scrutinized in fine-grained analyses.

The following shows our steps in selected our keywords of interest.

4 Methodology

First, we collected a list of negative terms from previous literature, including the list of negative terms from the *Dictionary of Frequently Used Positive and Negative Terms* (常用褒貶義詞語詳解詞典) complied by 郭先珍, 張偉, 劉縉 and 王玲 (1996), as well as Zhang's (2014) sadness expressions in Mandarin.

Zhang's list consists of (a) English expressions of sadness taken from the Bank of English; (b) Chinese expressions taken from the Chinese Corpus from the Center for Chinese Linguistics (CCL); and (c) a match of the two lists by checking them in the Babel English-Chinese Parallel Corpus. The list was divided according to 'Extroversion', 'Introversion', 'Verb', 'Modifier', 'Feeling', 'People involved', 'Cause', 'Modifies', and 'Others'.

[2] https://www.english-corpora.org/coca/

From the list of expressions, we identified 92 expressions that are potentially negative in Mandarin.

Category	English	Chinese
Verb	*induce/ give rise to/ lead to/ invite*	惹
		鬧
Modifier	*heavy*	沉重
	with grief	沉痛
Feeling	*extremely*	悲痛欲絕
	grief	悲痛
	sorrow	悲傷
		悲哀
	sad	悲
	sad	難過
		難受
	sad	傷心
	immensely sad/ heartbroken	心碎
	mourn	哀悼
	depression	抑鬱
		抑鬱症
		憂鬱症
	melancholy	憂鬱
	depressed	壓抑
	melancholy	憂傷
	gloomy	陰鬱
	gloomy	悶悶不樂
	dejected	沮喪
		喪氣
		垂頭喪氣
	disheartened	心灰意冷
		灰心喪氣
	low-spirited	情緒低落
	despondent	灰心
	anguish	痛苦
	low-spirited	消沉
	discouraged	氣餒
	frustrated	失意
	pessimistic	悲觀
	indignation	憤怒
	grief and indignation	悲憤
	annoyance/ anger	生氣
	agitated/ perturbed	煩
		煩躁
		心煩
	anxious	焦慮
	dreary	沉悶
	worry	憂愁
		憂
		愁
	woebegone	愁眉苦臉
		愁容滿面
	gloomy	苦悶
	grievance/ feel wronged	委屈
	uneasiness	不安
	vexation	煩惱
	worried	苦惱
	fear	恐懼
		害怕
	despair	絕望
	disappointment	失望
	loneliness	孤獨
		寂寞
		孤寂
	shame	羞愧
	regret	後悔
People involved	*victim*	死難者
		罹難者
		遇害者
		受難者
Cause	*news about the death of a beloved person*	噩耗
	failure	失敗
	setback	挫折
	hardship	艱難
	misfortune	不幸
	loss	失去
	disaster	災難
	die/ death	死
		死去
		死亡
		逝世
		去世
	die of disease	病逝
	be killed in a disaster	罹難
		遇難
		遇害
		身亡
Others	*(of sorrow) be blocked*	鬱結
		鬱積
	intestine+broken	腸斷
	gall bladder+crack	膽裂

	torment	折磨
	torture	煎熬
	suicide	自殺
	desolate	淒涼
	sad	悲慘
	chained	束縛

Table 1. List of Negative Terms Organized from Zhang (2014)

In addition to this list, we also went over the list of negative terms taken from the *Dictionary of Frequently Used Positive and Negative Terms*. There were altogether 506 positive terms and 587 negative terms in their dictionary. The negative terms were categorized according to 'Resentment' (憎惡), 'Blame' (貶責), 'Criticism' (批評), and 'Distain' (鄙夷). From the list, we selected three from the first three categories and one from the final categories for further analysis. Three other terms were added in the process from searches in dictionary (placed in 'Others' in Table 2). In total, there were thirteen terms to be investigated in details.

'Resentment' (憎惡)	慣技	抱頭鼠竄	保護傘
'Blame' (貶責)	奇貨可居	無所不至	貨色
'Criticism' (批評)	瞻前顧後	八面玲瓏	兩面光
'Distain' (鄙夷)	平庸		
Others	卵翼	冒泡	暴發戶

Table 2. Selected Terms for Analyses

Terms	**Definitions**[3]
1. 平庸	平凡 'ordinary; mediocre'
2. 慣技	時常用的方法、手段 'customary tactic'
3.卵翼	鳥以羽翼護卵，孵出小鳥。比喻養育或庇護。'Birds hatching eggs with wings – a metaphor of fostering or shielding'
4. 暴發戶	稱突然發跡，得財或得勢的人。'parvenu; someone who has suddenly become rich or powerful'
5. 抱頭鼠竄	形容像鼠懼人一般，狼狽逃走的樣子。'To descibe a rat-like timid behavior, running to hide'
6. 瞻前顧後	(1)比喻做事謹慎周密。'Cautious and careful' (2) 形容做事猶豫不決，顧慮太多。'People who are hesitant to do things, worry too much'
7. 保護傘	比喻賴以不受傷害的資本 'To not harm as the basic principle' [Literally, 'protective umbrella']
8. 八面玲瓏	形容人處世圓滑，面面俱到。'People who are sleek and cover all dimensions'
9. 貨色	(1)商品的種類及質料 'Different kinds of goods' (2)財貨、女色'money; women'
10.兩面光	比喻做人處事老練成熟，兩方面討好。'People who are experienced and sophisticated , always please both sides'
11.奇貨可居	珍異的貨品，可以收藏聚集起來，等候高價出售。後比喻仗持某種專長或有利用價值的東西作為資本以謀利。'Precious goods that are kept so that they could be sold at a higher price later on; Used to describe someone who earn profit by selling precious goods'
12.無所不至	(1)沒有到達不了的地方 'Nowhere is unreachable' (2)形容細心周到：'Careful and thoughtful' (3)形容才藝精通：'Talented' (4)比喻什麼壞事都做得出來：'Capable of any crime' (5)沒有不會發生的：'Nothing won't happen'
13. 冒泡	由下往上或往外透出氣泡。'Effervesce' *美得冒泡：嘲諷的話。諷刺人自以為是，想得太美。'To sneer at someone who believe him/herself infallible'

Table 3.Translation of the Negative Terms

These thirteen terms were numbered so that they could be matched to their results in the next section. Some words have more than one sense (#6, #9, #12). Among the senses, we could see that some

[3] Chinese definitions taken from http://dict.revised.moe.edu.tw/cbdic/index.html

senses carry a positive meaning. Therefore, it is not right to provide a 'negative' tag to these terms. The next section will provide the analysis of news articles.

5 Results

For all thirteen terms, they were searched in the *UDN News* in the *United Daily News Database* for the period of ten years from October 1, 2008 through October 1, 2018. The number of news articles found is shown in Table 4 below, with the highest being #7 'protective umbrella' (保護傘), followed by #1 'ordinary; mediocre'(平庸). A total of 1460 articles were collected.

Terms	**No. of Articles**	**%**
7. 保護傘	405	27.57
1. 平庸	384	26.14
6. 瞻前顧後	167	11.37
9. 貨色	136	9.26
11. 奇貨可居	133	9.05
4. 暴發戶	111	7.56
8. 八面玲瓏	51	3.47
10. 兩面光	26	1.77
13. 冒泡	17	1.16
5. 抱頭鼠竄	12	0.82
2. 慣技	11	0.75
3. 卵翼	10	0.68
12. 無所不至	6	0.41
Total	**1469**	**100.00**

Table 4. Number of UDN New Articles within 1o years

From 1469 news articles, a total of 1543 occurrences of these terms were found. For each occurrence, annotated was made by identifying if the thirteen terms were indeed negative, positive, or neutral. Using 平庸 *píngyōng* 'ordinary; mediocre' (#1) as example, we provide the following examples.

(4) Negative

好 作家 謙虛， **平庸** 作家 驕狂，
hǎo zuòjiā qiānxū, píngyōng zuòjiā jiāokuáng,
good writer humble mediocre writer arrogant
明明 文字 不佳 卻 也 意見 最多。
míngmíng wénzì bùjiā què yě yìjiàn zuìduō
obviously word not.good but also opinion most
'Good writers are humble while <u>mediocre</u> writers are arrogant. Mediocre writers' works are not so good but they have too many opinions.'

(5) Positive

你 現在 沒 了 膽、 少 了 腎，
nǐ xiànzài méi le dǎn, shǎo le shèn,
you now Neg. LE gallbladder lack LE kidney
從 今天 起 要 多 一點 **平庸**、
cóng jīntiān qǐ yào duō yīdiǎn píngyōng,
from today start want more little ordinary
多 一點 微笑， 以 自己 為 優先。
duō yīdiǎn wéixiào, yǐ zìjǐ wèi yōuxiān.
more little smile for yourself as priority
'Now, you lose your gallbladder and kidney. From today, you should become more <u>simple</u> [<u>ordinary</u>] with more smile; make yourself the priority.'

(6) Neutral

老天 保佑， 資質 **平庸** 的 兒子
lǎotiān bǎoyòu, zīzhì píngyōng de érzi
god bless talent ordinary DE son
終於 可以 免 受 基測
zhōngyú kěyǐ miǎn shòu jīcè
finally can avoid suffer competence.test
之 苦， 好好 享受 國中 生活
zhī kǔ hǎohǎo xiǎngshòu guózhōng shēnghuó
ZHI pain well enjoy junior.high.school life
'God bless my son. My <u>ordinary</u> boy finally can avoid the suffering of the competence test and enjoy his junior high school life.'

Based on the above criteria, we annotated all thirteen terms and the results are provided in Table 5.

Terms	**Negative**	**Positive**	**Neutral**	**Total**
1. 平庸	396	6	49	451
2. 慣技	9	0	2	11
3. 卵翼	8	2	0	10
4. 暴發戶	92	8	17	117
5. 抱頭鼠竄	9	0	3	12

6. 瞻前顧後	120	19	28	167
7. 保護傘	202	176	27	405
8. 八面玲瓏	13	21	17	51
9. 貨色	30	5	101	136
10. 兩面光	5	10	11	26
11. 奇貨可居	23	28	82	133
12. 無所不至	1	4	1	6
13. 冒泡	2	11	5	18
Total	**910**	**291**	**343**	**1543**

Table 5. Negative, Positive, and Neutral Use

The results in Table 4 can be clearly compared in Figure 1 when converted to percentages.

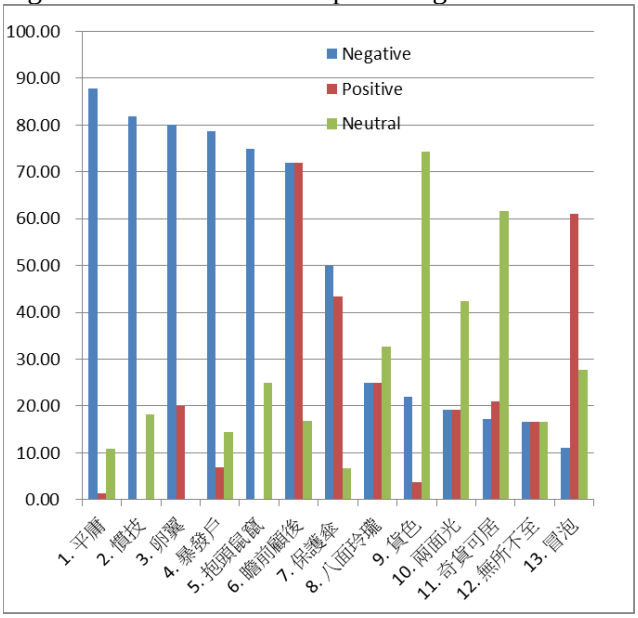


Figure 1. Proportions of Negative, Positive, and Neutral Uses

From Figure 1, we can clearly see that, the top seven terms are highly negative, although #6 (瞻前顧後) has similar proportion of positive and negative use. If we look at Table 3, we could see that this word has two sense -- 'Cautious and careful' and 'People who are hesitant to do things, worry too much' -- between the two senses, the first one is considered more positive than the second sense. From the results, we could also see that both senses are equally used.

It is worth noting that, we took these terms from a list of words listed under negative terms by other dictionaries and resources. The results we found showed that these terms are not entirely negative. In another example, a potentially negative term 保護傘 'protective umbrella', a metaphorical use, which means 'to not harm as the basic principle', is slightly high in negative use, although its positive use is also frequent. Examples are given in (7) and (8).

(8) Negative

法治 底線 撤守，
fǎzhì *dǐxiàn* *chèshǒu*
rule.of.law bottom.line withdrawn
成 了 鬧事 者 的
chéng le nàoshì zhě de
become LE make trouble person DE
保護傘
bǎohùsǎn
protective.umbrella
'The failure of defending the rule of laws [literally, the bottom line of the rule of laws withdrew] has become the protective umbrella [excuses] of the troublemakers.'

(9) Positive

柯文哲 昨天 受訪 表示，
kēwénzhé *zuótiān* *shòufǎng* *biǎoshì*
KeWenzhe yesterday interviewed indicated
大 巨蛋 公開 透明，
dà *jùdàn* *gōngkāi* *tòumíng*
Big dome open transparent
反而 是 **保護傘**，
fǎnér *shì* *bǎohùsǎn*
instead is protective.umbrella
後續 處理 就是 安全 合法
hòuxù *chùlǐ* *jiùshì* *ānquán* *héfǎ*
follow-up procedure is safe legal
'Wenzhe Ke [Mayor of Taipei] said yesterday that the [procedures of building] Big Dome is open and transparent, and because of this, it has become a protective umbrella. Therefore, the follow-up procedures will be safe and legal.'

The use of 'protective umbrella' is not entirely negative too. Other terms from #9 to #13 are more neutral or positive than the other terms on the left of Figure 1. Terms that are more positive than negative are #11(奇貨可居) and #13(冒泡), between which the latter is highly positive.

Examples of positive and negative use of 冒泡 'effervesce' re given in (10) below.

(10) 但 「央行 不 樂見 這些 地區
dàn `yāngháng bù lèjiàn zhèxiē dìqū
but Central.bank Neg. glad.see these districts
一直 **冒泡**」，
yīzhí màopào
continuously effervesce
因為 房價 不斷 飆高，
yīnwèi fángjià bùduàn biāogāo,
because house.price coninou rise.high
日後 跌下來也會 很 慘
rìhòu diē xiàlái yě huì hěn cǎn
future fall down also want very disastrous
'However, the central bank is not willing to see the house transaction keeps happening[to effervesce]. If the housing price keeps rising rapidly, when it falls down in an unexpected way, it will be disasterous. '

(11) Positive
比起 往年， 花 開 得
bǐqǐ wǎngnián, huā kāi de
compare previous.year flowers open DE
又多、又 漂亮！ 令 民眾
yòu duō yòu piàoliang lìng mínzhòng
so many so pretty let people
直呼 「真的 美得 **冒泡**！」
zhí hū `zhēnde měi dé màopào
directly.say really pretty DE effervesce
"Compared with the last few years, more blossom of flowers and more beautiful flowers are seen. People sighed that 'the flowers are really beautiful [until effervesce].'"

From the above, we can see that, some terms which are potentially negative, or may have been collected as part of negative words, can still be used positively, or non-negatively. In Table 5 below, we provide the possible metaphorical extension from negative to positive meanings. The metaphorical concepts are in lower capitals. The translation is underlined.

Terms	**Metaphorical Extensions**
1. 平庸	Being ordinary is BAD, but being ordinary brings simplicity and therefore could become GOOD.
2. 慣技	No positive meaning is found.
3.卵翼	To provide a shield is BAD when used for people with bad intent; but for people who gain benefits from it, it is a GOOD thing.
4. 暴發戶	A parvenu refers to someone usually from a LOWER STATUS who has become rich but his or her manner does not improve. It is later used to mean the QUANTITY THAT EXPLODES which is seen as a good thing.
5. 抱頭鼠竄	No positive meaning is found.
6. 瞻前顧後	To do things with caution is a VIRTUE; but over-caution could mean TIMIDITY.
7. 保護傘	Like #3, to be a protective umbrella is BAD when used for people with bad intent; but for people who gain benefits from it, it is a GOOD thing.
8. 八面玲瓏	Someone who is sleek and cover all dimensions is HARD TO TRUST; but someone who can do this well could SUCCEED IN SOCIALIZATION.
9. 貨色	People, especially women, who have been objectified is referred to as different kinds of goods, which is a BAD LABEL FOR WOMEN. However, some skilled people or goods that can use this label to mean they are of RARE GOOD QUALITY.
10.兩面光	Like, #8, someone who is experienced and sophisticated , always please both sides is HARD TO TRUST; but someone who can do this well could SUCCEED IN SOCIALIZATION.
11.奇貨可居	Someone who keeps precious goods so that to sell a t higher price is HARD TO TRUST for their ultimate goal is profit; It is later used to refer to people who know SUCCEED IN BUSINESS for they how to do business by selling at the right time.
12.無所不至	When someone or something is able to reach all corners, it is something GOOD for things are well taken care of. But it could become BAD when the reaching is for a bad purpose.
13. 冒泡	Literally, 'to effervesce' is neutral. But if something happens too rapidly like the formation of bubles, it is seen as a BAD SIGN. A good side of this use is not

	found except in describing beauty. Originally to describe someone's beauty to the extent that it effervesces, it is OVER-CONFIDENCE of one's beauty, but it is later used to mean beauty that kills.

Table 6. Metaphorical Mechanisms of Each Term

From the above, we can see the semantic pragmatics meanings of potentially negative terms, and the possible metaphorical mechanisms at play when a potentially negative meaning could become positive, or vice versa.

6 Conclusion and Limitations

We started out by collecting lists of negative terms and ended up focusing on thirteen that we have observed to serve both positive and negative meanings. We then evaluated the negative meanings of these thirteen words by examining their use in news articles. By looking into the context of each of the occurrence, we found the proportions of positive, negative, and neutral uses. We then summarized the metaphorical mechanisms that may have caused the change or meanings in each term. This study, however, did not analyze the different senses of each item, if any, in the separate context. For words that have more than one sense, it is possible that a particular sense if predominantly positive or negative. We leave this for future study.

Acknowledgments

This research was supported by MOST project 106-2410-H-004-109-MY2 and NCCU Grant 108H112-01.

References

Bednarek, M. (2008). Semantic preference and semantic prosody re-examined. *Corpus Linguistics and Linguistic Theory*, *4*(2), 119-139.

Giora, R., Balaban, N., Fein, O., & Alkabets, I. (2004). Negation as Positivity in Disguise. In Colston, H. L., & Katz, A. (Eds.), *Figurative Language Comprehension: Social and Cultural Influences* (pp. 233-258). Hillsdale, NJ: Lawrence Erlbaum Associates.

Hunston, S. (2007). Semantic prosody revisited. *International Journal of Corpus Linguistics*, *12*(2), 249-268. DOI: 10.1075/ijcl.12.2.

Hunston, S. (2011). *Corpus Approaches to Evaluation: Phraseology and Evaluative Language*. New York: Routledge.

Lakoff, G. (1993). The contemporary theory of metaphor. In Ortony, Andrew (Ed.). Metaphor and thought, 202-251, Cambridge: Cambridge University Press.

Lakoff, G., & Johnson, M. (1980). Metaphors we live by. Chicago: University of Chicago press.

Linebarger, M. C. (1980). *The grammar of negative polarity* (Doctoral dissertation, Massachusetts Institute of Technology).

Martin, J. R., & White, P. R. R. (2005). *The language of evaluation: Appraisal in English*. London: Palgrave Macmillan.

Murphy, M. L. (2003). *Semantic relations and the lexicon: Antonymy, synonymy and other paradigms*. Cambridge University Press.

Sobieraj, S., & Berry, J. M. (2011). From incivility to outrage: Political discourse in blogs, talk radio, and cable news. *Political Communication*, *28*(1), 19-41.

Xiao, R., & McEnery, T. (2006). Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective. *Applied linguistics*, *27*(1), 103-129.

Zhang, R. 2014. Sadness Expressions in English and Chinese. London, New York: Bloomsbury.

卓淑玲、陳學志和鄭昭明。2013。台灣地區華人情緒與相關心理生理資料庫—中文情緒詞常模研究。*中華心理學刊*，55(4)，493-523。

郭先珍、王玲玲。1991。褒義、貶義詞在搭配中的方向性。中國人民大學學(6)，96-100。

郭先珍、張偉、劉縉、王玲玲編。1996。常用褒貶義詞語詳解詞典。北京：商務印書館。

陳學志、詹雨臻和馮彥茹。2013。台灣地區華人情緒與相關心理生理資料庫—中文情緒隱喻的刺激常模。*中華心理學刊*，55(4)，525-553。

傅奇珅。2010。文字背後的意含：資訊的量化測量公司基本面與股價（以中鋼為例）。台灣：國立政治大學碩士論文。

彭宣維、楊曉軍和何中清。2011。漢英對應評價意義語料庫簡介。*外語電化教學*(5)，3-10。

彭宣維、劉玉潔、張冉冉、陳玉娟、談仙芳、王玉英和楊曉軍著。2015。漢英評價意義分析手冊：評價語料庫的語料處理原則與研制方案。北京：北京大学出版社。

Mapping distributional to model-theoretic semantic spaces: a baseline

Franck Dernoncourt
Adobe Research
franck.dernoncourt@adobe.com

Abstract

Word embeddings have been shown to be useful across state-of-the-art systems in many natural language processing tasks, ranging from question answering systems to dependency parsing. Herbelot and Vecchi (2015) explored word embeddings and their utility for modeling language semantics. In particular, they presented an approach to automatically map a standard distributional semantic space onto a set-theoretic model using partial least squares regression. We show in this paper that a simple baseline achieves a +51% relative improvement compared to their model on one of the two datasets they used, and yields comparable results on the second dataset.

1 Introduction

Word embeddings are one of the main components in many state-of-the-art systems for natural language processing (NLP), such as language modeling (Mikolov et al., 2010), text classification (Socher et al., 2013; Kim, 2014; Blunsom et al., 2014; Lee and Dernoncourt, 2016), question answering (Weston et al., 2015; Wang and Nyberg, 2015), machine translation (Bahdanau et al., 2014; Tamura et al., 2014; Sundermeyer et al., 2014), as well as named entity recognition (Collobert et al., 2011; Dernoncourt et al., 2016; Lample et al., 2016; Labeau et al., 2015).

Word embeddings can be pre-trained using large unlabeled datasets typically based on token co-occurrences (Mikolov et al., 2013; Collobert et al., 2011; Pennington et al., 2014). They can also be jointly learned with the task.

Understanding what information word embeddings contain is subsequently of high interest. Herbelot and Vecchi (2015) investigated a method to map word embeddings to formal semantics, which is the center of interest of this paper. Specifically, given a feature and a word vector of a concept, they tried to automatically find how often the given concept has the given feature. For example, the concept *yam* is always a *vegetable*, the concept *cat* has a coat most of the time, the concept *plug* has sometimes 3 prongs, and the concept *dog* never has wings.

The method they used was based on partial least squares regression (PLSR). We propose a simple baseline that outperforms their model.

2 Task

In this section, we summarize the task presented in (Herbelot and Vecchi, 2015). The following is an example of a concept along with some of its features, as formatted in one of the two datasets used to evaluate the model:

yam	a_vegetable	all	all	all
yam	eaten_by_cooking	all	most	most
yam	grows_in_the_ground	all	all	all
yam	is_edible	all	most	all
yam	is_orange	some	most	most
yam	like_a_potato	all	all	all

The concept *yam* has six features (*a_vegetable*, *eaten_by_cooking*, *grows_in_the_ground*, *is_edible*, *is_orange*, and *like_a_potato*). Each feature in this dataset is annotated by three different humans. The annotation is a quantifier that reflects how frequently

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 388-394, Hakodate, Japan, September 13-15, 2019

the concept has a feature. Five quantifiers are used: *no*, *few*, *some*, *most*, and *all*. In this example, the concept *yam* has been annotated as *some*, *most* and *most* for the feature *is_orange*.

Each of the five quantifiers is converted into a numerical format with the following (somehow arbitrary) mapping: no $\mapsto$ 0; few $\mapsto$ 0.05; some $\mapsto$ 0.35; most $\mapsto$ 0.95; all $\mapsto$ 1. The value is averaged over the three annotators. Using this mapping, we can map a concept into a "model-theoretic vector" (also called feature vector). If a feature has not been annotated for a concept, then the element in the model-theoretic vector corresponding to the feature will have value 0. As a result, any element of a model-theoretic vector that has value 0 may correspond to a feature that has either been annotated as *no* by the three annotators, or not been annotated (presumed *no*). Given that there can be many features and it is possible that only some of them are annotated for each concept, the model-theoretic vector may be quite sparse.

In the yam example, if we only included features annotated with yam, the model-theoretic vector would be as follows:

$$\begin{bmatrix} \frac{\text{all+all+all}}{3} \\ \frac{\text{all+most+most}}{3} \\ \frac{\text{all+all+all}}{3} \\ \frac{\text{all+most+all}}{3} \\ \frac{\text{some+most+most}}{3} \\ \frac{1+1+1}{3} \end{bmatrix} = \begin{bmatrix} \frac{1+1+1}{3} \\ \frac{1+0.95+0.95}{3} \\ \frac{1+1+1}{3} \\ \frac{1+0.95+1}{3} \\ \frac{0.35+0.95+0.95}{3} \\ \frac{1+1+1}{3} \end{bmatrix} \approx \begin{bmatrix} 1 \\ 0.967 \\ 1 \\ 0.983 \\ 0.75 \\ 1 \end{bmatrix}$$

The additional coordinates corresponding to all the remaining features would be zero. Each concept word will have a vector of the same dimension (number of unique features) in the same dataset. The coordinates mean the same from one concept to another. For example, the feature `is_vegetable` appears in the same coordinate position in all the vectors.

3 Datasets

Two datasets are used:

- The Animal Dataset (AD) (Herbelot, 2013) contains 73 concepts and 54 features. All concepts are animals, and for each concept all features are annotated by 1 human annotator. There are 3942 annotated pairs of concept-feature ($73 * 54 = 3942$). The dimension of the model-theoretic vectors will therefore be 54.
- TheMcRae norms (QMR) (McRae et al., 2005) contains 541 concepts covering living and non-living entities (e.g., alligator, chair, accordion), as well as 2201 features. One concept is annotated with 11.4 features on average by 3 human annotators. There are 6187 annotated pairs of concept-feature ($541 * 11.4 \approx 6187$). The dimension of the model-theoretic vectors will therefore be 2201, and each model-theoretic vector will have on average $2201 - 11.4 = 2189.6$ elements set to 0 due to unannotated features.

4 Model

In the previous section, we have seen how to convert a concept into a model-theoretic vector based on human annotations. The goal of Herbelot and Vecchi (2015) is to analyze whether there exists a transformation from the word embedding of a concept to its model-theoretic vector, the gold standard being the human annotations. The word embeddings are taken from the word embeddings pre-trained with word2vec *GoogleNews-vectors-negative300*[1] (300 dimensions), which were trained on part of the Google News dataset, consisting of approximately 100 billion words.

The transformation used in (Herbelot and Vecchi, 2015) is based on Partial Least Squares Regression (PLSR). The PLSR is fitted on the training set: the inputs are the word embeddings for each concept, and the outputs are the model-theoretic vectors for each concept.

To assess the quality of the predictions, the Spearman rank-order correlation coefficient is computed between the predictions and the gold model-theoretic vectors, ignoring all features for which a concept has not been annotated. The idea is that some of the features might be present but not given as options during annotation. The method should therefore not be penalized for not suggesting them. Figure 1 illustrates the model.

[1] `https://code.google.com/p/word2vec/`

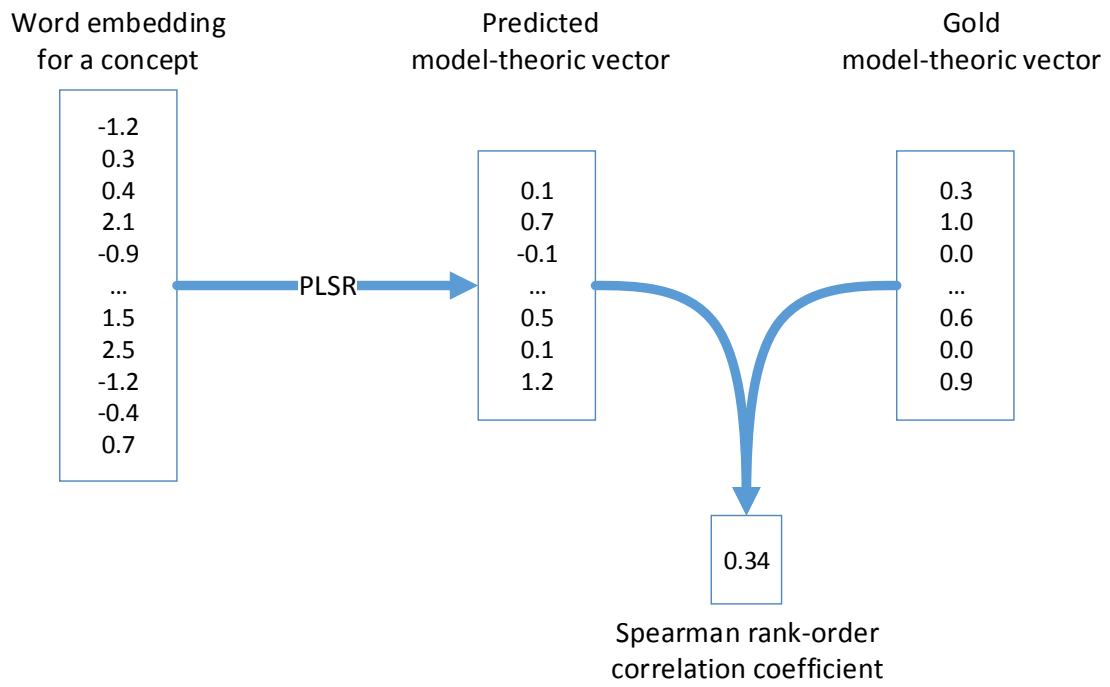


Figure 1: Overview of Herbelot and Vecchi (2015)'s system. The word embedding of a concept is transformed to a model-theoric vector via a PLSR. The quality of the predicted model-theoric vector is assessed with the Spearman rank-order correlation coefficient between the predictions and the gold model-theoretic vectors. Note that some of the elements that equal 0 in the gold model-theoretic vector may correspond to features that are not annotated for the concept. Such features are omitted when evaluating the Spearman rank-order correlation coefficient. Also, the dimension of the model-theoretic vectors could be larger or smaller than the dimension of the word embedding. Since the word embeddings we use have 300 dimensions, the model-theoretic vectors will be smaller than the word embeddings in the AD dataset, and larger in the QMR dataset.

5 Experiments

We compare Herbelot and Vecchi (2015)'s model (PLSR + word2vec) against three baselines: mode, nearest neighbor and random vectors.

- *Mode*: A predictor that outputs, for each feature, the most common feature value (i.e., the mode) in the training set. For example, if a feature is annotated as *all* for most concepts, then the predictor will always output *all* for this feature. When finding the most common value of a feature, we ignore all the concepts for which the feature is not annotated. The resulting predictor does not take any concept into account when making a prediction. Indeed, the predicted values are always the same, regardless of the concept. If a feature has the same value for most concepts, the predictor may perform reasonably well.
- *Nearest neighbor (NN)*: A predictor that outputs for any concept the model-theoretic vector from the training set corresponding to the most similar concept in the training set. Similarity is based on the cosine similarity of the word vectors. This is a simple nearest neighbor predictor.
- *Random vectors*: Herbelot and Vecchi (2015) used pre-trained word embeddings as input to the PLSR, we instead simply use random vectors of same dimension (300, continuous uniform distribution between 0 and 1).

We also apply retrofitting (Faruqui et al., 2014) on the word embeddings in order to leverage relational information from semantic lexicons by encouraging linked words to have similar vector representations. Using (Faruqui et al., 2014)'s retrofitting tool[2], we retrofit the word embeddings (*GoogleNews-vectors-negative300*) on each of the 4 datasets present in the retrofitting tool (*framenet*, *ppdb-xl*, *wordnet-synonyms+*, and *wordnet-synonyms*.

[2] https://github.com/mfaruqui/retrofitting

	AD			QMR		
	Min	Average	Max	Min	Average	Max
PLSR + word2vec (our implementation)	0.435	0.572	0.713	0.244	0.332	0.407
PLSR + word2vec + framenet	0.423	0.577	0.710	0.236	0.331	0.410
PLSR + word2vec + ppdb	**0.455**	0.583	0.688	0.247	0.332	0.421
PLSR + word2vec + wordnet	0.429	0.583	0.713	0.252	0.339	0.444
PLSR + word2vec + wordnet+	0.453	**0.604**	0.724	0.261	0.344	0.428
PLSR + random vectors	0.253	0.419	0.550	−0.017	0.087	0.178
NN + word2vec	0.338	0.524	0.751	0.109	0.215	0.291
NN + word2vec + framenet	0.321	0.516	0.673	0.108	0.204	0.288
NN + word2vec + ppdb	0.360	0.531	0.730	0.114	0.213	0.300
NN + word2vec + wordnet	0.384	0.551	0.708	0.115	0.208	0.297
NN + word2vec + wordnet+	0.390	0.597	**0.806**	0.138	0.235	0.324
NN + random vectors	0.244	0.400	0.597	−0.063	0.029	0.107
mode	0.432	0.554	0.643	**0.420**	**0.522**	**0.605**
true-mode	0.419	0.551	0.637	0.379	0.466	0.551
Herbelot and Vecchi (2015) (PLSR + word2vec)	?	0.634	?	?	0.346	?

Table 1: All the presented results are averaged over 1000 runs, except for the results of Herbelot and Vecchi (2015) in the last row. PLSR stands for partial least squares regression, NN for nearest neighbor, ppdb for the Paraphrase Database (Ganitkevitch et al., 2013). There are two ways to compute the mode: either taking the mode of the means of the 3 annotations (*mode*), or the mode for all annotations (*true-mode*). QMR has 3 potentially different annotations for each concept-feature pair, while AD has 3 only one annotation for each concept-feature pair: as a result, *mode* and *true-mode* have similar results for AD, but potentially different results for QMR. For each run, a train/test split was randomly chosen (60 training samples for AD, 400 for QMR, in order to have the same number of training samples as in Herbelot and Vecchi (2015)'s Table 2).

6 Results and discussion

Table 1 presents the results, using the Spearman correlation as the performance metric. The experiment was coded in Python using scikit-learn (Pedregosa et al., 2011) and the source as well as the complete result log and the two datasets are available online[3]

Furthermore, the *mode* baseline yields results that are good on the AD dataset (0.554, vs. in 0.634 (Herbelot and Vecchi, 2015) vs. 0.572 in our PLSR + word2vec implementation), and significantly better than all other models for the QMR dataset (0.522, vs. 0.346 in (Herbelot and Vecchi, 2015), i.e. +51% improvement). To get an intuition of why the mode baseline works well, Figures 2 and 3 show that most features tend to have one clearly dominant quantifier in the AD dataset. A similar trend can be found in the QMR dataset. In the AD dataset, there are 54 features, each of them being annotated for all 73 concepts. In the QMR dataset, there are 2201 features, each of them being annotated for only $\frac{6187}{2201} \approx 2.81$ concepts on average. As a result, it is much more difficult for the PLSR to

[3] https://github.com/Franck-Dernoncourt/model-theoretic

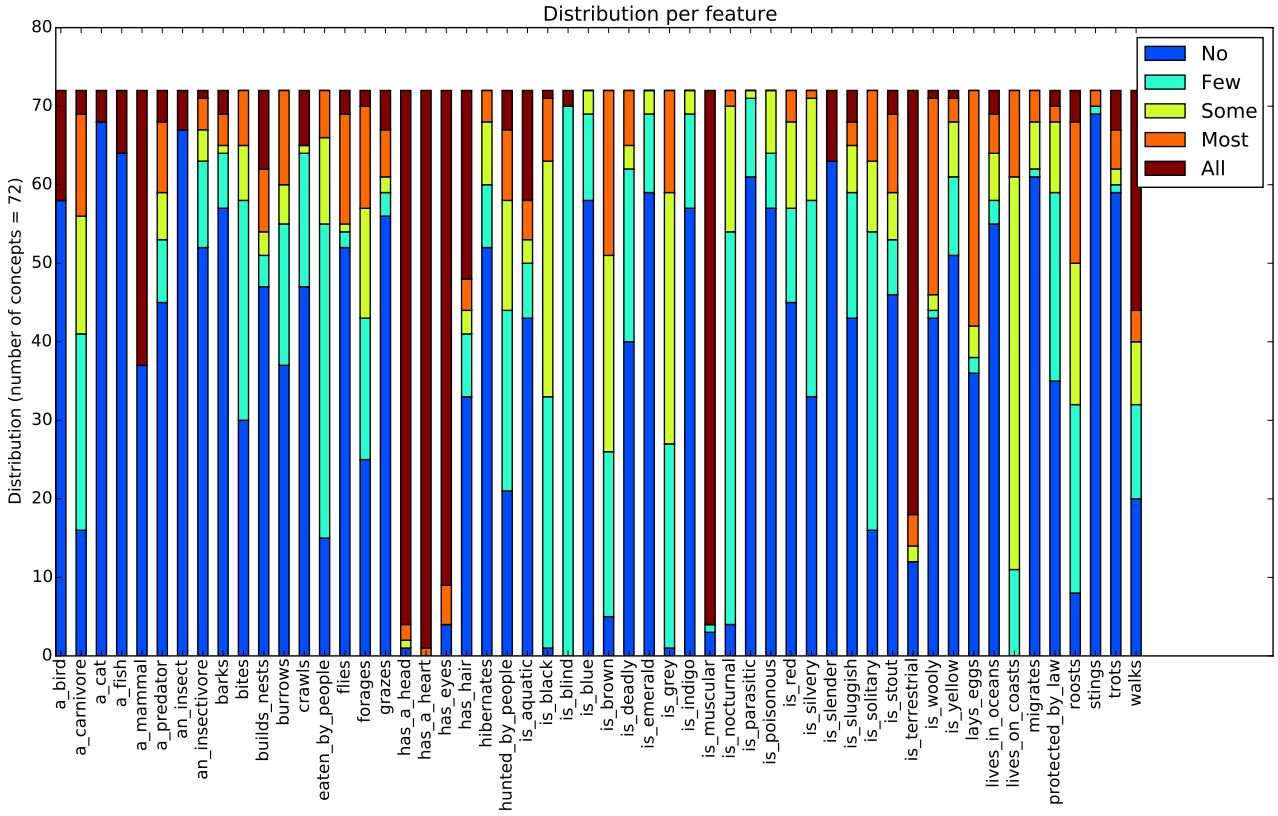


Figure 2: Stacked bars showing the distribution of quantifiers among features in the AD dataset: most features tend to have one clearly dominant quantifier. For example, the feature *a_cat* is almost always annotated with the qualifier *no*.

learn the mapping from word embeddings to model-theoretic vectors in the QMR dataset than in the AD dataset. This explains why the *mode* baseline outperforms PLSR in the QMR dataset but not in the AD dataset.

The *random vector* baseline with PLSR performs mediocrely on the AD dataset, and very poorly on the QMR dataset. The *nearest neighbor* baseline yields some competitive results on the AD dataset, but lower results on the QMR dataset. Lastly, using retrofitting increases the performances on both AD and QMR datasets. This is expected as applying retrofitting to word embeddings leverages relational information from semantic lexicons by encouraging linked words to have similar vector representations.

7 Conclusion

In this paper we have presented several baselines for mapping distributional to model-theoretic semantic spaces. The *mode* baseline significantly outperforms Herbelot and Vecchi (2015)'s model on the QMR dataset, and yields comparable results on the AD dataset. This indicates that state-of-the-art models do not efficiently map word embeddings to model-theoretic vectors in these datasets.

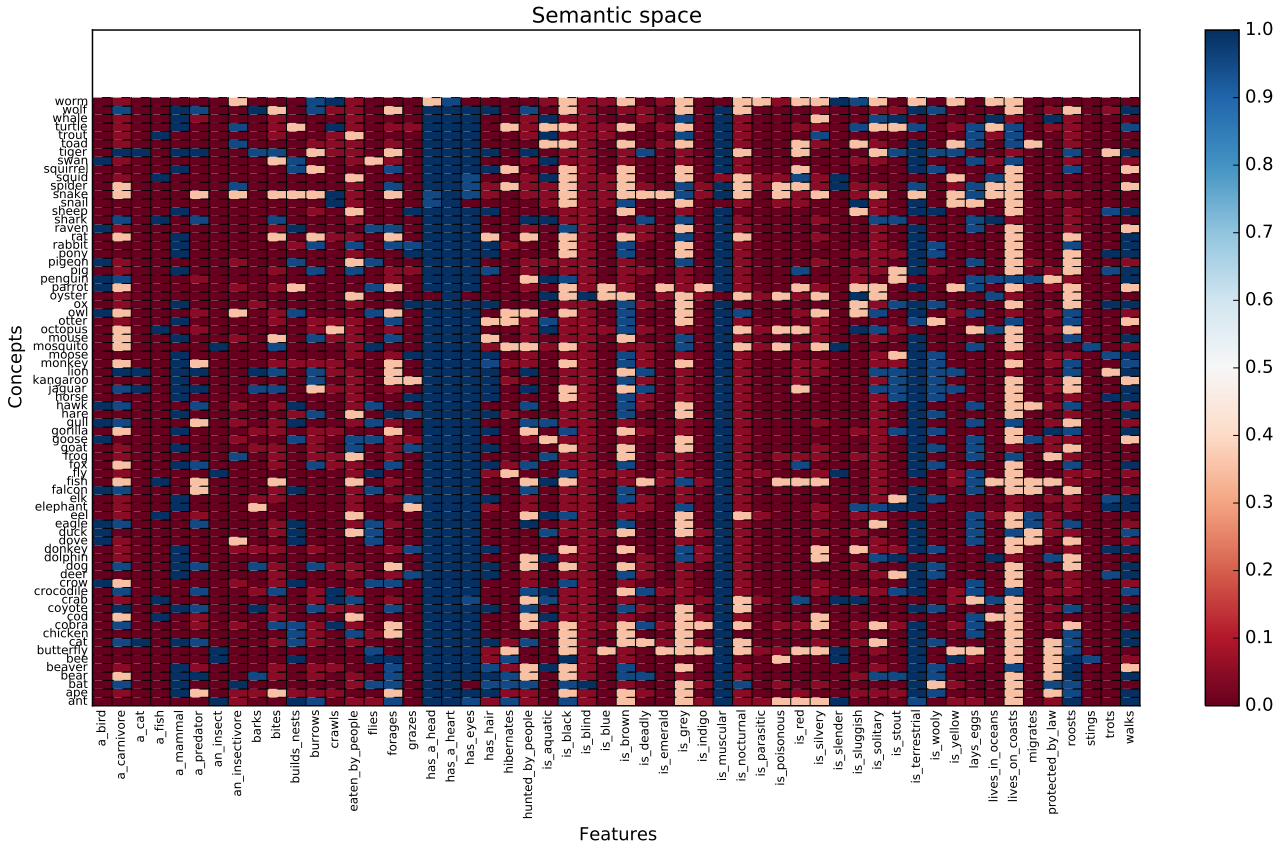


Figure 3: Heatmap showing the distribution of quantifiers among features in the AD dataset: most features tend to have one clearly dominant quantifier.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Phil Blunsom, Edward Grefenstette, Nal Kalchbrenner, et al. 2014. A convolutional neural network for modelling sentences. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*. Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *The Journal of Machine Learning Research*, 12:2493–2537.

Franck Dernoncourt, Ji Young Lee, Ozlem Uzuner, and Peter Szolovits. 2016. De-identification of patient notes with recurrent neural networks. *arXiv preprint arXiv:1606.03475*.

Manaal Faruqui, Jesse Dodge, Sujay K Jauhar, Chris Dyer, Eduard Hovy, and Noah A Smith. 2014. Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166*.

Juri Ganitkevitch, Benjamin Van Durme, and Chris Callison-Burch. 2013. PPDB: The paraphrase database. In *Proceedings of NAACL-HLT*, pages 758–764, Atlanta, Georgia, June. Association for Computational Linguistics.

Aurélie Herbelot and Eva Maria Vecchi. 2015. Building a shared world: Mapping distributional to model-theoretic semantic spaces. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 22–32.

Aurelie Herbelot. 2013. What is in a text, what isnt, and what this has to do with lexical semantics. *Proceedings of the Tenth International Conference on Computational Semantics (IWCS2013)*.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1746–1751. Association for Computational Linguistics.

Matthieu Labeau, Kevin Löser, and Alexandre Allauzen. 2015. Non-lexical neural architecture for fine-grained

POS tagging. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 232–237, Lisbon, Portugal, September. Association for Computational Linguistics.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. *arXiv preprint arXiv:1603.01360*.

Ji Young Lee and Franck Dernoncourt. 2016. Sequential short-text classification with recurrent and convolutional neural networks. In *Human Language Technologies 2016: The Conference of the North American Chapter of the Association for Computational Linguistics, NAACL HLT 2016*.

Ken McRae, George S Cree, Mark S Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior research methods*, 37(4):547–559.

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH*, volume 2, page 3.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. 2011. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research*, 12(Oct):2825–2830.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. GloVe: global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)*, 12:1532–1543.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer.

Martin Sundermeyer, Tamer Alkhouli, Joern Wuebker, and Hermann Ney. 2014. Translation modeling with bidirectional recurrent neural networks. In *EMNLP*, pages 14–25.

Akihiro Tamura, Taro Watanabe, and Eiichiro Sumita. 2014. Recurrent neural networks for word alignment model. In *ACL (1)*, pages 1470–1480.

Di Wang and Eric Nyberg. 2015. A long short-term memory model for answer sentence selection in question answering. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 707–712, Beijing, China, July. Association for Computational Linguistics.

Jason Weston, Antoine Bordes, Sumit Chopra, and Tomas Mikolov. 2015. Towards AI-complete question answering: A set of prerequisite toy tasks. *arXiv preprint arXiv:1502.05698*.

Intrinsic Evaluation of Grammatical Information within Word Embeddings

Daniel Edmiston
Department of Linguistics
University of Chicago
Chicago, IL, USA
danedmiston@uchicago.edu

Taeuk Kim
Department of Computer Science and Engineering
Seoul National University
Seoul, Korea
taeuk@europa.snu.ac.kr

Abstract

This work presents a proof-of-concept study for a framework of intrinsic evaluation of continuous embeddings as used in NLP tasks. This evaluation method compares the geometry of such embeddings with ground-truth embeddings in a linguistically-inspired, discrete feature space. Using model distillation (Hinton et al., 2015) as a means of extracting morphological information from models with no explicit morphological awareness (e.g. word-atomic models), we train multiple learner networks which do model morpheme composition so as to compare the amount of grammatical information different models capture. We use Korean affixes as a case-study, as they encode multiple types of linguistic information (phonological, syntactic, semantic, and pragmatic), and allow us to investigate specific types of linguistic generalizations models may or may not be sensitive to.

1 Introduction

While NLP systems built with neural network architectures have dominated the field in recent years, it is often lamented that their improved performance has come at the cost of understanding the models. Furthermore, recent work on natural language inference (McCoy et al., 2019) has cast doubt on the ability of such models to generalize linguistic patterns effectively. Particularly, carefully selected examples are shown to fool these systems, suggesting they learn heuristics for performing well on data sets rather than truly capturing linguistic information. As such, methods of probing the information within these models are of growing importance. This work proposes such a method by comparing the geometry of ground-truth, discrete-space embeddings against continuous embeddings learnt by neural networks. To do this, we extract morphological information from different embedding models via transparent model distillation (Tan et al., 2017) and examine the resulting morpheme embeddings for grammatical information. By investigating the embeddings directy this falls under the rubric of intrinsic evaluation. This complements extrinsic evaluation methods, where embeddings' ability to serve as input for classifiers on linguistic tasks is seen as a proxy for their linguistic content.

As is, intrinsic evaluation methods in the literature most often test for lexical semantics. For instance, the word analogy task tests similarities between pairs of words, e.g. *king* is to *queen* as *man* is to what? Here, since our ground-truth embeddings reflect grammatical properties, the similarities between them and the continuous representations can be taken to reflect *grammatical*, rather than *lexical* content in the continuous representations.

As an initial investigation, we compare the morphological information from multiple methods of embedding Korean words into continuous space, using three learner networks to distill morpheme representations for comparison. These are the StaffNet architecture of Edmiston and Stratos (2018), the morphological recursive neural network MRNN model of Luong et al. (2013), and a TreeLSTM (Tai et al., 2015) over morphological parses. By distilling into these networks, we extract explicit morpheme representations. We focus on the *affix* repre-

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 395-404, Hakodate, Japan, September 13-15, 2019

sentations, as they house the grammatical information we are testing for.

Korean affixes are the choice for this pilot study for multiple reasons. (i) Korean is an agglutinative language, meaning there is (largely) a one-to-one correspondence between affixes and meanings. (ii) The morphology is highly regular, which facilitates high-fidelity morphological parsing. (iii) Korean affixes display at least four distinct types of linguistic information. **Phonolgical**: Korean exhibits phonologically-driven allomorphy. **Syntactic:** Korean affixes contain syntactic information as they attach to different syntactic units. **Semantic:** affixes perform different semantic functions, e.g. logical operators vs. focus markers. **Pragmatic:** Certain affixes are indicative of formal language, and others display honorific features. This allows us to run focused experiments which probe what type of information different models are sensitive to.

Having distilled affix embeddings from various ground-truth models, we run our evaluation task by comparing the distilled representations of different models with ground-truth morpheme representations embedded in a discrete, linguistically-inspired feature space, and show that indeed neural models are picking up on at least some forms of *grammatical* meaning. Results suggest that semantic relationships are more difficult to capture than syntactic, and models appear insensitive to phonological/pragmatic information.

Contributions: **(i)** We introduce a new linguistically-inspired intrinsic evaluation task which probes for *grammatical* meaning, rather than *lexical* meaning. **(ii)** We show results that different types of linguistic information are captured to differing degrees, and may be of a differing nature from one another. **(iii)** To the authors' knowledge, this is the first application of transparent model distillation for interpretation in an NLP setting. **(iv)** We focus on an under-studied language in NLP, which also happens to be typologically far removed from those usually studied in the literature.

2 Related Work

This work falls in the context of the emerging literature on the interpretability of neural network models using model distillation, and on the interpretability of linguistic representations learnt for NLP tasks.

In the broader context of interpreting the behavior of neural network models, model distillation has emerged as a viable method. Tan et al. (2018b) use so-called transparent model distillation to audit black-box risk scoring models. By distilling a black-box model into a transparent learner model, and comparing this learner model with a non-distilled transparent model trained on ground-truth data, they are able to gain insights into black-box models. Likewise, Zhang et al. (2018) use model distillation (which they call *knowledge distillation*) to extract human-interpretable features from the middle layers of convolutional networks trained on computer vision tasks. While similar, our method differs slightly from these approaches in that we use model distillation to induce representations for linguistic units which would otherwise be unavailable (affixes).

As for the interpretability of linguistic representations learnt for NLP tasks, extrinsic evaluation methods have focused both on morphological information (e.g. (Belinkov et al., 2017)) and syntactic information (e.g. (Adi et al., 2016)). As these are extrinsic methods, the task consists of learning representations, and then training classifiers on carefully designed supervised learning tasks using these representations as input. The tasks are meant to probe for linguistic properties (e.g. negation (Ettinger et al., 2016)), and performance on these tasks can be interpreted as a proxy for the amount of linguistic information contained in the original representations. Work in the intrinsic evaluation domain has largely focused on testing lexical semantics, as by comparing relations between e.g. countries and capitals (Mikolov et al., 2013a). By comparison, here we intrinsically test for what we call *grammatical information*, to be made specific in Section 3.3.

3 Background

3.1 Model Distillation

Model distillation (Hinton et al., 2015) is the technique of training one neural network (the *learner network*) to approximate the output of another (the *ground-truth network*). While the technique was originally designed to train relatively lightweight networks to approximate the outputs of larger, more cumbersome networks or ensembles, model distillation has recently been used to facilitate the interpretation of so-called "black-box" neural network models (Tan et al., 2017, 2018a,b; Zhang

et al., 2018).

The idea behind using model distillation for the interpretation of word-embeddings in this study is to train a learner network amenable to morphological interpretation to approximate a model which is otherwise not straight-forward to analyze morphologically. In this case, this includes word-atomic word-embedding models (such as *Word2Vec*), as well as word-embedding models which compose embeddings from sub-word constituents (e.g. syllable-embedding models). Assuming successful distillation, the morphological representations of the learner network can be seen as a proxy for the morphological information captured by the original ground-truth network. Here, model distillation of a ground-truth model into a learner network proceeds by iterating over the corpus the ground-truth model was originally trained on. For each word, the ground-truth embedding $\mathbf{y}$ is calculated, as is the learner network's embedding $\hat{\mathbf{y}}$. Training proceeds by optimizing the learner network's parameters to minimize the squared distance between $\mathbf{y}$ and $\hat{\mathbf{y}}$.

3.2 Learner Networks

3.2.1 StaffNet

The first learner network is the StaffNet architecture. Introduced in Edmiston and Stratos (2018), StaffNet ('Stem-Affix Net') is a dynamic neural network architecture designed to compose morpheme representations into full word-embeddings in a linguistically plausible way. The defining feature of StaffNet is its distinguishing of morphemes into stems and affixes, treating the former as vectors and the latter as functions over vectors. Here, we treat affixes as linear transformations and represent them as matrices in $\mathbb{R}^{d \times d}$.

StaffNet is a dynamic architecture, whose composition of a word-embedding varies with the morphological parse of the word. It composes a word-embedding in a three-step process. First, a word is decomposed into its constituent stems and affixes.[2] Second, the (potentially compound) stem representation is calculated as the convex combination of the outputs of a BiLSTM into which the stems are fed. Third, any affixes are iteratively applied to the compound stem representation—the convex combination from the previous step. Figure 1 illustrates this for the word *cheese-burger*.PL.NOM, or *cheeseburgers* marked with nominative case.

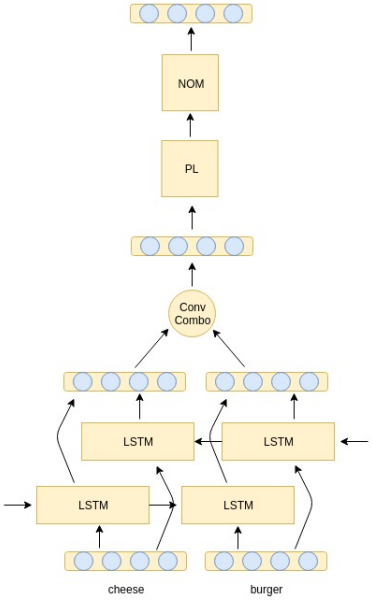


Figure 1: StaffNet architecture showing the dynamic composition of *cheese-burger*.PL.NOM.

3.2.2 MRNN

The second learner network is the morphological RNN model of Luong et al. (2013), which constructs a word's embedding from constituent morpheme embeddings by means of a recursive neural network over the binary tree of the word's morphological parse. Parent nodes in the network are functions of their children nodes, and are calculated as $p = f(W[c_1; c_2] + b)$. That is, they are the result of a non-linearity (here *tanh*) applied to the output of an affine transformation over the concatenated children embeddings. An example is as in Figure 2.

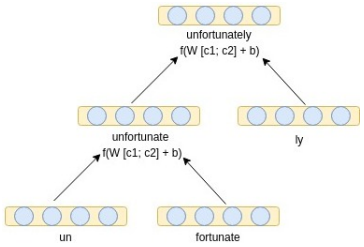


Figure 2: Morphological RNN architecture showing the composition of *un-fortunate-ly*

[2] All morphological parsing done with the *Komoran* POS-tagger available with the KoNLPy package. http://konlpy.org/en/latest/

3.2.3 TREELSTM

The final learner network is the N-ary Tree-LSTM of Tai et al. (2015). The transition equations follow the original paper, where k indexes the kth child of parent node p.

$$i_p = \sigma(W^{(i)}x_p + \sum_{l=1}^{N} U_l^{(i)} h_{pl} + b^{(i)})$$

$$f_{pk} = \sigma(W^{(f)}x_p + \sum_{l=1}^{N} U_{kl}^{(f)} h_{pl} + b^{(f)})$$

$$o_p = \sigma(W^{(o)}x_p + \sum_{l=1}^{N} U_l^{(o)} h_{pl} + b^{(o)})$$

$$u_p = tanh(W^{(u)}x_p + \sum_{l=1}^{N} U_l^{(u)} h_{pl} + b^{(u)})$$

$$c_p = i_p \odot u_p + \sum_{l=1}^{N} f_{pl} \odot c_{pl}$$

$$h_p = o_p \odot tanh(c_p)$$

Here, N is restricted to 2, as this model also operates over binary morphological parses.

3.3 Korean Affixes as a Case Study

As mentioned above, Korean affixes exhibit at least four different types of information: phonological, syntactic, semantic, and pragmatic. Phonological information is shown through phonologically-driven allomorphy. For example, the accusative marker -을 '*-eul*' attaches to nominals ending in a coda, while its allomorph -를 '*-leul*' attaches to nominals ending in vowels. A familiar analogy from English would be the choice between '*an*' and '*a*', e.g. '*an animal*' vs. '*a dog*.' Syntactic information is shown through place of attachment (Cho and Sells, 1995). There are many semantic dimensions along which affixes vary, e.g. some contribute focus semantics, others serve as logical operators. Multiple pragmatic dimensions of meaning are also evident in Korean affixation. Some affixes are reserved for written usage, others indicate the relationship between speaker and hearer, such as honorifics.

Having formalized the feature set of 107 Korean affixes, we embed each affix into the binary feature space $\{0, 1\}^n$, the dimensions of which are interpreted as linguistic features (e.g. [+CODA] vs. [-CODA]) with 1 indicating the presence of that feature, 0 otherwise. These embeddings serve as 'ground-truth' representations of Korean affixes, and geometrically can be interpreted as the corners of an $n-$dimensional hypercube. We define distance in this binary feature space with Hamming distance, where $d_H(x, y)$ is the number of dimensions along which x and y differ.

4 Methodology

4.1 Ground-truth Models and Distillation

We propose a method of intrinsically evaluating the extent to which different word-embedding models capture the meaning of affixes in Korean, and therefore how well they capture phonological, syntactic, semantic, and pragmatic information. The models we distill and compare are as follows. **Character-level:** *fastText* (Bojanowski et al., 2017),[3] Naver's *kor2vec* model (based on Kim et al. (2016))[4], **Syllable:** The model of Choi et al. (2017), **Word:** *Skip-gram* and *CBOW* models of Mikolov et al. (2013a), and *GloVe* Pennington et al. (2014).

All models were trained on a Korean Wikidump with vocabulary size limited to 10,000, and were trained to produce embeddings in 300 dimensions, as suggested by Choi et al. (2016). Other hyperparameters followed the suggestions of the original publications where applicable. Each model was then distilled into each of the learner architectures, embedding affixes as vectors in $\mathbb{R}^{300}$ or in STAFFNET's case, as matrices in $\mathbb{R}^{300\times300}$.

4.2 Comparing Affix Representations for Intrinsic Evaluation

One standard method of performing intrinsic evaluation of high-dimensional word-embeddings is analogy tests (Mikolov et al., 2013b). In such a test, word-sets are assembled of the form *a:b, c:d*, where d is withheld. Given an embedding model, an estimation of withheld d is given by $\hat{d} = \vec{b} - \vec{a} + \vec{c}$. An example is counted as correct if the vector $\vec{d}$ is the closest—via cosine similarity—in the embedding space to the model's calculated $\hat{d}$, and incorrect otherwise.

In the case where affixes are represented by vectors, here too we make use of cosine similarity. Where affixes are represented as matrices, cosine similarity is not applicable, and instead we compare them via subspace similarity, as described in Algorithm 1 (Mu et al., 2017).

$\mathbf{M}_{aff1}$ and $\mathbf{M}_{aff2}$ represent the matrix representations of the affixes to be compared. N is an inte-

[3] https://fasttext.cc/docs/en/pretrained-vectors.html

[4] https://github.com/naver/kor2vec

Algorithm 1 Subspace Similarity

Input: $\mathbf{M}_{aff1}$, $\mathbf{M}_{aff2}$, N
Output: $score \in [0, 1]$

$X \leftarrow [pc(\mathbf{M}_{aff1})_1; ...; pc(\mathbf{M}_{aff1})_N]$
$Y \leftarrow [pc(\mathbf{M}_{aff2})_1; ...; pc(\mathbf{M}_{aff2})_N]$
$Z \leftarrow X^\top Y$
$score \leftarrow \sqrt{\sum_{t=1}^{N} \sigma_t^2}/N$
return(*score*)

ger signifying the number of principal directions to use. σ_t represents the t^{th} singular value of Z. The algorithm proceeds by performing PCA on the matrix inputs, and stacking the first N principal directions into matrices in $\mathbb{R}^{d \times N}$. The sum of squares of the N singular values of the product of these matrices, divided by N, results in a similarity score in $[0, 1]$. The geometric intuition behind this metric is that similar affixes should map stems to similar subspaces.

To evaluate the distilled representations, we compare the continuous embeddings against the discrete embeddings in the following way. Given the set of all affixes $\mathcal{A}$, consider the subset $\mathcal{Y}_{\text{Aff}} = \underset{x \in \mathcal{A}, x \neq \text{Aff}}{argmin}\, d_H(x, \text{Aff})$, the set of closest affixes to any given affix in the ground truth discrete feature space, where $|\mathcal{Y}_{\text{Aff}}| = k$. In the continuous space, we define the $k-$closest affixes to any given affix with $\hat{\mathcal{Y}}_{\text{Aff}} = \underset{x \in \mathcal{A}, x \neq \text{Aff}}{k\text{-}argmax}\, sim(x, \text{Aff})$, where $sim(x, y)$ is cosine similarity or subspace similarity, depending on architecture. By design $\mathcal{Y}_{\text{Aff}}$ and $\hat{\mathcal{Y}}_{\text{Aff}}$ are of the same cardinality.

Given $\mathcal{Y}_{\text{Aff}}$ and $\hat{\mathcal{Y}}_{\text{Aff}}$, we define two scores. The first is the percentage of overlap—the percentage of the k-closest affixes in the continuous representation which are k-closest with regard to the ground-truth embeddings.[5] The second measures error; $avg(\{d_H(x, \text{Aff}) \mid x \in \hat{\mathcal{Y}}_{\text{Aff}}\}) - \underset{x \in \mathcal{A}, x \neq \text{Aff}}{min}\, d_H(x, \text{Aff})$; that is, the true error with regard to an affix, as calculated by the average hamming distance between Aff and x for $x \in \hat{\mathcal{Y}}_{\text{Aff}}$, minus the minimum possible error. We label this penalty the *Hamming offset*. Note that percentage of overlap and the Hamming offset provide two graded measures of success with regard to an affix, unlike all-or-nothing diagnostics like e.g. analogy tests.

[5] Since it is always the case that $|\mathcal{Y}_{\text{Aff}}| = |\hat{\mathcal{Y}}_{\text{Aff}}|$ this amounts to the F_1 measure on cluster analysis.

4.3 Data Sets

This study makes use of two data sets. The first is a Korean WikiDump used to train the original models, and also used as the corpus for model distillation. The second data set was hand-constructed for this study, and constitutes the ground-truth representations used in the experiments below. It is the embedding of 107 Korean affixes into a discrete, binary feature space, consisting of 62 dimensions. These 62 dimensions correspond to linguistic features along which Korean affixes vary, and include features such as the aforementioned [+CODA], and [+HONORIFIC]. We divided the linguistic features into four distinct feature subsets, one for each of phonological, syntactic, semantic, and pragmatic features, so as to run tests on feature subsets. Over the affixes, there are five distinct phonological configurations, eight distinct syntactic configurations, 55 distinct semantic configurations, and five distinct pragmatic configurations.

5 Experiments

5.1 Verifying Distillations

As we are testing models based on their distilled representations rather than their original representations, the first question to ask is whether the distilled embeddings are a faithful recreation of the models they are meant to emulate.

MODEL	STAFFNET	MRNN	TREELSTM
KOR2VEC	0.971	0.840	0.952
FASTTEXT	0.949	0.824	0.946
SYLLABLE	0.967	0.869	0.979
W2V-SG	0.953	0.755	0.920
W2V-CBOW	0.941	0.786	0.878
GLOVE	0.935	0.690	0.875

Table 1: Average cosine similarity between ground-truth embeddings and distilled embeddings over 10k vocabulary.

The results in Table 1 show that each of the models was able to reproduce the original embeddings to a very high degree of accuracy. This is especially true given that the overwhelming majority of volume in $\mathbb{R}^{300}$ is orthogonal to any given point. We take this to mean that the models have been successfully distilled, and our distilled representations can serve as faithful representatives of their ground-truth models.

5.2 Closest Affix

This section tabulates the scores for each of the models—as well as a random baseline—on the intrinsic task described in Section 4.2. We test each model distilled into each architecture, and for the StaffNet architecture we also test for different values of subspace rank N. We chose the rank of the subspace to be the minimum number of principal components which accounted for 25%, 50%, and 75% of the average variance of the affixes for each model. As mentioned above, we calculate the average percent of overlap between the k-closest in the continuous and discrete spaces, and also calculate the average Hamming offset. The results are in Table 2, where the figures represent the scores for each model averaged over performance on all 107 affixes.

As can be seen in the results, no model was able to achieve a high level of accuracy on the task, but all models significantly outperform the random baseline, and the MRNN and TreeLSTM models fare better than the StaffNet distillations with regard to both average percentage of overlap and average hamming offset.

5.3 Feature Subsets

Given that Korean affixes contain linguistic information of different sorts, we can perform a variant of our experiment from above using only subsets of features. For example, given the feature-makeup described above, affixes fall into one of five phonological configurations, which we can describe as +CODA, -CODA, +LOW, -LOW, or NONE. For these experiments, similar to before define $\mathcal{Y}_{\text{Aff}} = \underset{x \in \mathcal{A}, x \neq \text{Aff}}{argmin}\, d_H(phon(x), phon(\text{Aff}))$, or the set of affixes which are closest to a certain affix in the discrete space considering only phonological features. We then define $\hat{\mathcal{Y}}_{\text{Aff}}$ as before, $\hat{\mathcal{Y}}_{\text{Aff}} = \underset{x \in \mathcal{A}, x \neq \text{Aff}}{argmax}\, sim(x, \text{Aff})$. Percentage of overlap and Hamming offset are as before. Results can be interpreted as: for the k-closest affixes in the continuous space, how many behave the same with regard to, e.g. phonological features? This should help identify what type of linguistic information models are sensitive to.

The results for the phonological, syntactic, semantic, and pragmatic subset tests are in Tables 3-6.

6 Discussion

Before discussing individual test results, it is noteworthy that the scores of the tables vary significantly from one another, and the random baseline shows particularly strong results for certain subsets, particularly the pragmatic subset. This is the result of fluctuations in the average k (i.e. cluster size) and average *Hamming offset* for each subset. Table 7 lists these figuers. Dividing the average k by the total number of affixes $|\mathcal{A}|$ gives the expected random score for percentage of overlap. For each subset, the random baselines roughly reflect these figures. For interpreting the results, performance relative to the random baseline is what is important, not the percentage figure or Hamming offset figures themselves.

Examining Table 2, the MRNN and TreeLSTM models outperformed the StaffNet models by a large margin. A plausible hypothesis would be to attribute this difference to the lack of non-linearity in the StaffNet-derived affixes. While StaffNet does derive stem representations via back-propagation through a BiLSTM, the affix representations always apply post-non-linearity during forward propagation, and as such are unable to learn any potentially non-linear relationships.

Regarding Table 3, while nearly all models outperformed the random baseline, none did so significantly. It is furthermore surprising that character and syllable-level models did not significantly outperform word-atomic models, to which phonological information of the kind driving allomorphy in Korean should be unavailable. This may suggest that the models examined here are not sensitive to phonological information in any significant way.

The syntactic results in Table 4 show all models outperforming the random baseline, suggesting it is possible for models to capture syntactic information. Furthermore, the distilling architectures performed relatively similarly, with a StaffNet distillation achieving the highest score. This suggests that non-linearity may not be necessary to capture syntactic relations.

For the semantic subset results in Table 5, while StaffNet distillations performed similar to the random baseline, the models deriving affix representations via non-linearity showed relatively strong results. This suggests two things. (i) It is possible for modern neural network models to capture *grammatical* meaning of a semantic nature,

Architecture	StAffNet-25% variance		StAffNet-50% variance		StAffNet-75% variance		MRNN		TreeLSTM	
Model	%Overlap	Offset$_{d_H}$	%Overlap	Offset$_{d_H}$	%Overlap	Offset$_{d_H}$	%Overlap	Offset$_{d_H}$	%Overlap	Offset$_{d_H}$
kor2vec	12%	3.35	11%	3.49	9%	3.39	16%	2.82	**19%**	2.62
fastText	13%	3.23	12%	3.22	9%	3.31	18%	**2.59**	16%	2.63
Syllable	13%	3.18	11%	3.42	8%	3.36	17%	2.88	17%	3.29
w2v-sg	12%	3.17	8%	3.57	7%	3.40	18%	**2.59**	15%	3.01
w2v-cbow	13%	3.19	9%	3.46	7%	3.56	18%	2.63	16%	2.86
GloVe	10%	3.30	10%	3.34	7%	3.47	16%	2.80	11%	3.13
Random	2%	4.23	3%	4.38	3%	4.47	2%	4.32	1%	4.35

Table 2: Percent correct and average Hamming Distance for models mapping to subspaces of different sizes. Best scores in bold, worst scores in red.

Architecture	StAffNet-25% variance		StAffNet-50% variance		StAffNet-75% variance		MRNN		TreeLSTM	
Model	%Overlap	Offset$_{d_H}$	%Overlap	Offset$_{d_H}$	%Overlap	Offset$_{d_H}$	%Overlap	Offset$_{d_H}$	%Overlap	Offset$_{d_H}$
kor2vec	31%	0.89	30%	0.91	30%	0.91	**32%**	0.91	**32%**	0.90
fastText	**32%**	0.87	31%	0.89	31%	0.89	29%	0.95	29%	0.95
Syllable	**32%**	0.86	31%	0.90	31%	0.90	31%	0.88	29%	0.96
w2v-sg	**32%**	**0.85**	30%	0.89	30%	0.89	29%	0.95	28%	0.96
w2v-cbow	31%	0.88	30%	0.91	30%	0.90	31%	0.92	31%	0.93
GloVe	32%	0.87	29%	0.91	29%	0.90	29%	0.95	28%	0.94
Random	26%	0.97	26%	0.99	27%	0.97	28%	0.96	29%	0.96

Table 3: Percent correct and Average Hamming offset: Restricting to **Phonological** features.

Architecture	StAffNet-25% variance		StAffNet-50% variance		StAffNet-75% variance		MRNN		TreeLSTM	
Model	%Overlap	Offset$_{d_H}$	%Overlap	Offset$_{d_H}$	%Overlap	Offset$_{d_H}$	%Overlap	Offset$_{d_H}$	%Overlap	Offset$_{d_H}$
kor2vec	36%	1.27	38%	1.24	39%	1.22	37%	1.27	39%	1.23
fastText	37%	1.27	40%	1.21	**41%**	**1.19**	38%	1.23	37%	1.25
Syllable	35%	1.30	37%	1.26	39%	1.22	35%	1.29	38%	1.25
w2v-sg	37%	1.26	37%	1.26	39%	1.21	37%	1.27	36%	1.29
w2v-cbow	36%	1.29	38%	1.24	38%	1.24	38%	1.23	39%	1.23
GloVe	36%	1.28	37%	1.26	38%	1.24	35%	1.30	34%	1.32
Random	28%	1.43	28%	1.45	29%	1.43	29%	1.42	30%	1.41

Table 4: Percent correct and Average Hamming offset: Restricting to **Syntacitc** features.

Architecture	StAffNet-25% variance		StAffNet-50% variance		StAffNet-75% variance		MRNN		TreeLSTM	
Model	%Overlap	Offset$_{d_H}$	%Overlap	Offset$_{d_H}$	%Overlap	Offset$_{d_H}$	%Overlap	Offset$_{d_H}$	%Overlap	Offset$_{d_H}$
kor2vec	6%	2.54	5%	2.52	5%	2.53	23%	1.85	25%	1.80
fastText	6%	2.39	5%	2.40	5%	2.43	**33%**	**1.65**	24%	1.72
Syllable	5%	2.47	6%	2.51	5%	2.64	18%	1.96	19%	2.04
w2v-sg	5%	2.48	5%	2.46	5%	2.48	26%	1.77	29%	1.78
w2v-cbow	6%	2.56	5%	2.54	5%	2.6	24%	1.89	19%	2.04
GloVe	5%	2.49	5%	2.49	4%	2.57	24%	1.89	23%	1.92
Random	3%	2.79	3%	2.70	2%	2.71	2%	2.72	2%	2.67

Table 5: Percent correct and Average Hamming offset: Restricting to **Semantic** features.

Architecture	StaffNet-25% variance		StaffNet-50% variance		StaffNet-75% variance		MRNN		TreeLSTM	
Model	%Overlap	Offset$_{d_H}$	%Overlap	Offset$_{d_H}$	%Overlap	Offset$_{d_H}$	%Overlap	Offset$_{d_H}$	%Overlap	Offset$_{d_H}$
kor2vec	69%	0.33	69%	0.33	70%	0.31	71%	0.31	**72%**	**0.29**
fastText	69%	0.32	69%	0.33	70%	0.32	71%	0.31	69%	0.32
Syllable	71%	0.31	69%	0.33	70%	0.33	70%	0.32	70%	0.31
w2v-sg	70%	0.33	69%	0.34	70%	0.32	70%	0.32	69%	0.33
w2v-cbow	68%	0.33	68%	0.34	69%	0.34	70%	0.32	70%	0.32
GloVe	68%	0.34	68%	0.34	68%	0.34	70%	0.32	69%	0.32
Random	71%	0.31	70%	0.34	69%	0.33	70%	0.31	69%	0.34

Table 6: Percent correct and Average Hamming offset: Restricting to **Pragmatic** features.

Subset	Avg. k	Avg. k / $\lvert\mathcal{A}\rvert$	Avg. d_H
All	2.21	0.02	5.56
Phon	29.93	0.28	0.96
Syn	30.92	0.29	1.42
Sem	3.25	0.03	3.00
Prag	73.91	0.69	0.32

Table 7: Average size of nearest k affixes for each feature subset.

and (ii) non-linearity is required to capture these meanings.

For the pragmatic results in Table 6, all models perform virtually indistinguishably from the random baseline. This suggests that the word-embedding models examined here are not sensitive to pragmatic features, at least not when distilled into another model.[6]

As this is a proof-of-concept study, the principal takeaway is that this method of intrinsic evaluation is possible, and reveals interesting characteristics of neural representations. Specifically, the underlying geometry of ground-truth discrete embeddings are at least to some extent being captured in the geometry of the continuous representations learnt by different neural-network models. Furthermore, the results in this study show that not only are neural network models sensitive to grammatical information—as distinct from lexical information—they are sensitive to different types of grammatical information to differing degrees. Syntactic information can apparently be captured by linear relations, while semantic information requires non-linearities. This result is perhaps not surprising, as even in the theoretical linguistics literature semantic analyses often require more complex algebraic structures built on top of syntactic parses. Finally, neural models seem insensitive to phonological and pragmatic information, with all models here performing virtually the same as the random baseline on these sub-tests.

Finally, in addition to the test results themselves, the fact that any models significantly outperformed the random baseline shows that transparent model distillation can serve as a viable means of extracting sub-atomic information from otherwise atomic representations.

7 Conclusion

This study has been a proof-of-concept for an intrinsic evaluation method which probes grammatical, rather than lexical information in word-embeddings. To do this, we studied the case of Korean affixes, which display multiple types of grammatical information. In order to derive affix representations, we used Transparent Model Distillation to extract morpheme representations where they otherwise did not exist. Our study has shown that neural network models are sensitive to grammatical information, and that the geometry of the continuous representations learnt by neural network models reflects to some degree the geometry of ground-truth discrete embeddings.

We put forward such a process as a framework for the fine-grained intrinsic analysis of the high-dimensional continuous embeddings which are often used to help solve natural language processing tasks.

[6] Though it is of note that all models were trained on text which is ostensibly academic in character (a WikiDump file), and which therefore is almost devoid of pragmatically-marked language like honorifics.

References

Yossi Adi, Einat Kermany, Yonatan Belinkov, Ofer Lavi, and Yoav Goldberg. 2016. Fine-Grained Analysis of Sentence Embeddings using Auxiliary Prediction Tasts. ArXiv preprint arXiv:1608.04207.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do Neural Machine Translation Models Learn about Morphology. ArXiv preprint arXiv:1704.03471.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Young-Mee Yu Cho and Peter Sells. 1995. A Lexical Account of Inflectional Suffixes in Korean. *Journal of East Asian Linguistics*, 4(2):119–174.

Sanghyuk Choi, Taeuk Kim, Jinseok Seol, and Sanggoo Lee. 2017. A Syllable-based Technique for Word Embeddings of Korean Words. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 36–40.

Sanghyuk Choi, Jinseok Seol, and Sang-goo Lee. 2016. On Word Embedding Models and Parameters Optimized for Korean. In *Proceedings of the 28th Annual Conference on Human and Cognitive Language Technology (In Korean)*.

Daniel Edmiston and Karl Stratos. 2018. Compositional Morpheme Embeddings with Affixes as Functions and Stems as Arguments. In *Proceedings of the Workshop on the Relevance of Linguistic Structure in Neural Architectures for NLP*, pages 1–5.

Allyson Ettinger, Ahmed Elgohary, and Philip Resnik. 2016. Probing for Semantic Evidence of Composition by Means of Simple Classification Tasks. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 134–139.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. 2015. Distilling the Knowledge in a Neural Network. In *NIPS Deep Learning and Representation Learning Workshop*.

Yoon Kim, Yacine Jernite, David Sontag, and Alexander Rush. 2016. Character-Aware Neural Language Models. In *Proceedings of AAAI*, pages 2741–2749.

Thang Luong, Richard Socher, and Christopher Manning. 2013. Better Word Representations with Recursive Neural Networks for Morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 104–113.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. ArXiv preprint arXiv:1902.01007.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeff Dean. 2013a. Distributed Representations of Words and Phrases and their Compositionality. In *Advances in Neural Information Processing Systems*.

Tomas Mikolov, Wen-Tau Yih, and Geoffrey Zweig. 2013b. Linguistic Regularities in Continuous Space Word Representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 764–751.

Jiaqi Mu, Suma Bhat, and Pramod Viswanath. 2017. Representing Sentences as Low-rank Subspaces. ArXiv preprint arXiv:1704.05358.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*.

Kai Sheng Tai, Richard Socher, and Christopher Manning. 2015. Improved Semantic Representations from Tree-Structured Long Short-Term Memory Networks. ArXiv preprint arXiv:1503.00075.

Sarah Tan, Rich Caruana, Giles Hooker, Paul Koch, and Albert Gordo. 2018a. Learning Global Additive Explanations for Neural Nets Using Model Distillation. ArXiv preprint arXiv:1801.08640.

Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. 2017. Detecting Bias in Black-box Models Using Transparent Model Distillation. ArXiv preprint arXiv:1710.06169.

Sarah Tan, Rich Caruana, Giles Hooker, and Yin Lou. 2018b. Distill-and-Compare: Auditing Black-Box Models Using Transparent Model Distillation. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 303–310.

Quanshi Zhang, Yu Yang, Yuchen Liu, Ying-Nian Wu, and Zhu Song-Chun. 2018. Unsupervised Learning of Neural Networks to Explain Neural Networks. ArXiv preprint arXiv:1805.07468.

Appendix A: Feature subsets

Table 8 displays the feature values which make up the dimensions of the discrete space. The combination of all feature values from the subsets comprise the full discrete space, upon which the experiment in Table 2 was run.

SUBSET	FEATURE VALUES
PHONOLOGICAL FEATURES	[+CODA], [-CODA], [+LOW], [-LOW]
SYNTACTIC FEATURES	[+POSTPOSITION], [+CONJUNCTIVE], [+X-LIMITER], [+Z-LIMITER], [+V1], [+V2], [+V3], [+V4]
SEMANTIC FEATURES	[+DECLARATIVE], [+NOMINATIVE], [+TOPIC], [+LOCATIVE], [+DIRECTIVE], [+GENATIVE], [+ACCUSATIVE], [+PAST], [+CONJUNCTION], [+ADVERBIAL], [+ALSO], [+TAG], [+NOMINALIZER], [+FROM], [+CONDITIONAL], [+LIKE], [+ESSIVE], [+COMPARATOR], [+COMPLEMENTIZER], [+RELATIVIZER], [+PAST], [+RETROSPECTIVE], [+FUTURE], [+PLURAL], [+PRESENT], [+BECAUSE], [+COPULA], [+QUOTATIVE], [+ABLATIVE], [+INTENT], [+MUST], [+RESULT], [-ANIMATE], [+ANIMATE], [+INSTRUMENTAL], [+INTERROGATIVE], [+DATIVE], [+GOAL], [+EVEN], [+COHORTATIVE], [+ONLY], [+DISJUNCTIVE], [+EACH], [+DURATION]
PRAGMATIC FEATURES	[+FORMAL], [-FORMAL], [+HONORIFIC], [+FAMILIAR]

Table 8: Description of feature subsets.

A Continuation-based Analysis of Contrastive *Wa* in Japanese

Hitomi Hirayama
Kyushu Institute of Technology / 1-1 Sensuicho, Tobata Ward, Kitakyushu, Fukuoka 804-8550, Japan
hirayama@dhs.kyutech.ac.jp

Abstract

This paper proposes an analysis of contrastive *wa* in Japanese using continuations. In this paper *wa* is treated as a type-shifter, which "continuizes" the element attached to it. Semantically and pragmatically, *wa* does not do anything when it is used as a thematic *wa*. However, it gives a special focus semantic value when it is used as so-called contrastive *wa*: a set of sets of propositions. The proposed analysis can also handle multiple contrastive topics (CTs) and *wa*-phrases appearing in the designated topic position.

1 Introduction

This paper proposes an analysis of contrastive *wa* in Japanese (Kuno, 1973) using continuations (Barker, 2001; Barker & Shan, 2015). The particle *wa* is treated as a type-shifter that works to separate a sentence into two parts: the *wa*-phrase and the rest of the sentence. This continuation-based analysis is not only useful for deriving the special focus semantic value associated with contrastive *wa* used as a contrastive topic (CT), but also has several merits in explaining empirical facts observed about this particular item.

The rest of the paper is structured as follows. In the next section, I will describe the basic behavior of the particle of our interest, *wa*. Then, some basic concepts of continuations will be introduced. Given that, the semantic calculation of the sentence with contrastive *wa* will be examined. In Section 3, I will demonstrate how the proposed analysis can account for some of the unique behaviors of contrastive *wa*. Section 4 offers conclusions and how this approach can be extended further.

1.1 Introduction – Contrastive *wa*

The particle *wa* is a well-known topic marker in Japanese. When used as a thematic topic (i.e., without accompanying an F-marked element), it usually refers back to a referent that is already introduced in the discourse, as shown in (1).

(1) Taroo-wa kita.
T-TOP came
'Taro came.'

This thematic *wa* cannot occur with new information. Therefore, it cannot mark a phrase that corresponds to an answer to the question as shown in (2b). However, when the phrase *wa* is attached to bears phonological focus, the *wa*-phrase can be used as an answer to the question, as in (2c).[1]

(2) a. Dare-ga kita?
who-NOM came
'Who came?'

b. Taroo-(ga/??wa) kita.
T-(NOM/TOP) came
'Taro came.'

[1] More concretely, when a *wa*-phrase is used contrastively as would be the case in (2c), we can observe post-focal reduction (Ishihara, 2003). I appreciate an anonymous reviewer's pointing out that just referring to bearing phonological focus is not sufficient.

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 405-413, Hakodate, Japan, September 13-15, 2019

c. [$_F$ Taroo]-wa kita.
T-TOP came
'(At least) Taro came'
$\rightsquigarrow$ I'm sure Taro came but not sure about other people.
$\rightsquigarrow$ Taro came but there are people who didn't come.

Note that (2c) has *at least* in the translation. As an answer to question (2a), (2c) is marked and conveys more information than the literal meaning of the sentence does. Depending on context, the addressee of the utterance in (2c) can draw different inferences. For instance, the speaker could have limited knowledge about who actually came (i.e., ignorance inferences). It is also possible that the speaker is suggesting that there are people who did not come but worth mentioning.

The extra information conveyed by contrastive *wa* has been keenly discussed in the literature (Hara, 2006; Kuroda, 2005; Oshima, 2002; Tomioka, 2009; Yabushita, 2017). These analyses vary in what kind of extra information is focused on and how the contribution of *wa* is treated. For the sake of space, the extensive review of all alternative analyses cannot be done here. The extra information conveyed by contrastive *wa* itself is not the main focus of the analysis given in this paper. Rather, the focus will be the special focus semantic value of this lexical item, assuming that the function of contrastive *wa* is just like the CT.[2] A CT is treated as a strategy that interlocutors can employ and that refers to the discourse structure that they entertain in the immediate context (Büring, 2003; Constant, 2014).

The questions to be addressed about the behavior of contrastive *wa* can be largely divided into two: (i) How is the discourse structure entertained by the interlocutors built? and (ii) How can inferences available with the use of this lexical item be explained? This paper will be mainly concerned about the first question by proposing an analysis using continuations. The next subsection is intended to provide a brief overview of the system used in the analysis.

[2]Recently Oshima (in press) discusses the reasons not to treat contrastive *wa* as a contrastive topic in Büring's sense. I agree with him in that contrastive *wa* cannot be treated as functioning exactly in the same way as CTs in English. Nevertheless, I would argue that contrastive *wa* is a realization of a CT in Japanese and refers to the discourse structure that is entertained by the interlocutors. This approach is useful to see how the structured discourse is utilized in a case of questions involving contrastive *wa*. See Chapter 4 of Hirayama (2019) for the discussion.

1.2 Introduction – Continuations

Ordinary Semantic Value

The analysis given in this paper is based on the continuation hypothesis (Barker & Shan, 2015) given below.

(3) *The continuation hypothesis*
Some natural language expressions denote functions on their continuations, i.e., functions that take their own semantic context as an argument.

For instance, we can treat quantifiers as functions that take their surrounding context as their argument and give us a truth value. Let us examine how it works with a simple example in (4a). This sentence has a quantifier, *everybody*. If we treat *everybody* as a function on its surrounding context, it will take the boldfaced part as its argument. Such a surrounding context is called a continuation. The continuation of *everyone*, which is the boldfaced part in (4a), should semantically be of type et given that we treat an NP argument as of type e. As shown in (4b), the boldfaced part lacks an NP to become an S and provide a truth value. (4b) is an argument of *everybody*. After taking this argument, *everybody* needs to provide a truth value. As a result, we can see *everybody* ends up with being of type *(et)t*, as shown in (4c). This is the semantic essence of continuations.

(4) a. **Taro saw** $\boxed{\text{everybody}}$ **yesterday**: *t*
b. **Taro saw** ___ **yesterday**: *et*
c. $\boxed{\text{everybody}}$: *(et)t*

Quantifiers are not only categories that can be thought of in terms of continuations. We can continuize any category. Continuized categories (XP) take their surrounding context (i.e., continuations, c_{XP}) and give back a truth value. (5) offers a list of some continuized words.

S → NP VP:	$\lambda c_{\text{S}}.\underline{\text{VP}}(\lambda P_{et}.\underline{\text{NP}}(\lambda x_e.c_{\text{S}}(P(x))))$	[The object takes scope over the subject]
S → NP VP:	$\lambda c_{\text{S}}.\underline{\text{NP}}(\lambda x_e.\underline{\text{VP}}(\lambda P_{et}.c_{\text{S}}(P(x))))$	[The subject takes scope over the object]

Table 1: Two possible rules for an S

VP→ NP Vt:	$\lambda c_{\text{VP}}.\underline{\text{NP}}(\lambda x.\ \underline{\text{Vt}}(\lambda R_{e(et)}.c_{\text{VP}}(R(x))))$

Table 2: Syntactic rule for a VP with a transitive verb (in Japanese)

[$_{\text{S}}$[$_{\text{NP}}$ Taro] [$_{\text{VP}}$ came]]	
$\rightsquigarrow$ (by a syntactic rule for an S)	$\lambda c_{\text{S}}.\underline{\text{VP}}(\lambda P_{et}.\underline{\text{NP}}(\lambda x.c_{\text{S}}(P(x))))$
$\rightsquigarrow$ ($\underline{\text{VP}}$=$\lambda c_{\text{VP}}.\ c_{\text{VP}}(\lambda x.\text{come}(x))$)	$\lambda c_{\text{S}}.[\lambda c_{\text{VP}}.\ c_{\text{VP}}(\lambda x.\text{come}(x))]\ (\lambda P_{et}.\underline{\text{NP}}(\lambda x.c_{\text{S}}(P(x))))$
$\rightsquigarrow$ (β-reduction)	$\lambda c_{\text{S}}.\underline{\text{NP}}(\lambda x.c_{\text{S}}(\lambda x.\text{come}(x)))$
$\rightsquigarrow$ ($\underline{\text{NP}}$=$\lambda c_{\text{NP}}.\ c_{\text{NP}}(t)$)	$\lambda c_{\text{S}}.\ [\lambda c_{\text{NP}}.\ c_{\text{NP}}(t)]\ (\lambda x.c_{\text{S}}(\lambda x.\text{come}(x)))$
$\rightsquigarrow$ (β-reduction)	$\lambda c_{\text{S}}.c_{\text{S}}(\text{come}(t))$

Table 3: The derivation of the ordinary semantic value of 'Taro came'

(5) *Continuized lexicon in the ordinary dimension*

a. Taro $\rightarrow \lambda c_{\text{NP}}.c_{\text{NP}}(t)$

b. everybody $\rightarrow \lambda c_{\text{NP}}.\forall x : c_{\text{NP}}(x)$

c. come $\rightarrow \lambda c_{\text{VP}}.c_{\text{VP}}(\lambda x.\text{come}(x))$

d. invite $\rightarrow \lambda c_{\text{Vt}}.c_{\text{Vt}}(\lambda x.\lambda y.\text{invite}(y,x))$

In terms of the composition of a sentence, we assume ordinary binary branching rules. A crucial difference between a usual binary branching rule and the rules used here is that the relation between the function and its argument is determined by syntactic rules as shown in Table 1. One possible way to see a binary branching S is to regard the VP as a function that takes the rest of the sentence (i.e., the subject) as its argument. In this case, the object ends up taking wide scope over the subject. If we switch the relation between the subject and object, the scope relation also changes. What is notable here is that there is no Quantifier Raising or covert movement necessary to derive the inverse scope reading. In the semantic computations given from now on, everything stays in situ. That is the same as the rule for a VP with a transitive verb, shown in Table 2. Now, given the lexicon in (5) and syntactic rules given in Tables 1–2, we can derive the ordinary semantic value of simple sentences such as *Taro came*.[3] The derivation is given in Table 3.[4]

Note that in the last step in Table 3, we have $\lambda c_{\text{S}}.c_{\text{S}}(\text{come}(t))$. This is because everything including a sentence is continuized. Therefore, at the last stage of the derivation we need to feed a trivial continuation of a sentence in order to obtain a usual semantic denotation for a sentence. In the ordinary dimension, such a trivial continuation of a sentence, c_{S}, is $\lambda p.p$ which is of type tt.

Focus Semantic Value

The basic notion of continuations is now in order. The process introduced above is, however, not enough to account for contrastive *wa* in Japanese. Recall that contrastive *wa* accompanies an F-marked element. As a result, the focus semantic value rather than the ordinary one is, in fact, a crucial component for the analysis. In this paper, I simply extend the mechanism introduced in the previous section to derive the focus semantic value using continuations.[5]

In the focus dimension, everything is to be treated as sets. As a result, everything has a higher type in the focus dimension. When an element is not F-marked, it is treated as a singleton set while when an element bears F-marking, it denotes a set of alternatives in the relevant domain. The lexicon in the

[3]There is no difference between syntactic rules for an S between English and Japanese. A rule for VP is different, however, due to the word order between the head and its complements.

[4]The semantic derivation of Japanese sentences will be shown using English words for the sake of readability.

[5]This is not the only way to achieve the same result, however. For instance, using monads (Charlow, 2014) would bring us the results of the same kind.

S→ NP VP: $\lambda c_{\text{S}}.\underline{\text{NP}}(\lambda X_{et}.\ \underline{\text{VP}}(\lambda \mathcal{P}_{(et)t}.\ c_{\text{S}}(\mathcal{P}(X))\ ))$
The shaded part is computed via PFA

Table 4: A rule for an S in the focus dimension

$[_{\text{S}}[_{\text{NP}}$ Taro$][_{\text{VP}}$ came$]]$	
$\rightsquigarrow$ (by a syntactic rule for an S)	$\lambda c_{\text{S}}.\underline{\text{NP}}(\lambda X_{et}.\ \underline{\text{VP}}(\lambda \mathcal{P}_{(et)t}.\ c_{\text{S}}(\mathcal{P}(X))\ ))$
$\rightsquigarrow$ ($\underline{\text{NP}}$=$\lambda c_{\text{NP}}.c_{\text{NP}}(\{t\})$)	$\lambda c_{\text{S}}.\ [\lambda c_{\text{NP}}.c_{\text{NP}}(\{t\})]\ (\lambda X_{et}.\ \underline{\text{VP}}(\lambda \mathcal{P}_{(et)t}.\ c_{\text{S}}(\mathcal{P}(X))\ ))$
$\rightsquigarrow$ (by β-reduction)	$\lambda c_{\text{S}}.\underline{\text{VP}}(\lambda \mathcal{P}_{(et)t}.\ c_{\text{S}}(\mathcal{P}(\{t\}))\ )$
$\rightsquigarrow$ ($\underline{\text{VP}}$=$\lambda c_{\text{NP}}.c_{\text{NP}}(\{\lambda x.\text{come}(x)\})$)	$\lambda c_{\text{S}}.[\lambda c_{\text{VP}}.c_{\text{VP}}(\{\lambda x.\text{come}(x)\})]\ (\lambda \mathcal{P}_{(et)t}.\ c_{\text{S}}(\mathcal{P}(\{t\}))\ )$
$\rightsquigarrow$ (by β-reduction)	$\lambda c_{\text{S}}.\ c_{\text{S}}(\{\lambda x.\text{come}(x)\}(\{t\}))$
$\rightsquigarrow$ (by PFA)	$\lambda c_{\text{S}}.\ c_{\text{S}}(\{\text{come}(t)\})$

Table 5: The derivation of the focus semantic value of 'Taro came.'

focus dimension can be given as in (6). The shaded part is to be computed via Pointwise Functional Application (Rooth, 1985, 1996), given in (7).

(6) *Continuized lexicon in the focus dimension*

a. Taro $\rightarrow \lambda c_{\text{NP}}.\ c_{\text{NP}}(\{t\})$

b. Taro$_{\text{F}} \rightarrow \lambda c_{\text{NP}}.\ c_{\text{NP}}(\{x : x \in D_e\})$

c. come $\rightarrow \lambda c_{\text{VP}}.\ c_{\text{VP}}(\{\lambda x.\text{come}(x)\})$

d. invite $\rightarrow \lambda c_{\text{Vt}}.\ c_{\text{Vt}}(\{\lambda x.\lambda y.\text{invite}(y,x)\})$

(7) *Pointwise Functional Application (PFA)*:
If $\beta \subseteq D_{\sigma\tau}$ and $\gamma \subseteq D_\sigma$,
then $\beta(\gamma)=\{f(x) \in D_\tau : f \in \beta\ \&\ x \in \gamma\}$

The syntactic rules are to be defined in terms of sets as well. A possible rule for an S in the focus dimension is given in Table 4. Again, the shaded part is computed via PFA. Using the rule in Table 4 and the lexicon, we can derive the focus semantic value of the sentence *Taro came.* as in Table 5. In the last line, we have $\lambda c_{\text{S}}.c_{\text{S}}(\{\text{come}(t)\})$. The trivial continuation of a sentence in the focus dimension is $\{\lambda p.p\}$. Once this trivial continuation is fed to the final result in Table 5 via PFA, we can get $\{\text{come}(t)\}$, which is the semantic denotation desired as the focus semantic value.

What has been introduced above is very basic, but it allows us to proceed to an analysis of contrastive *wa* using continuations in the next section.

2 How the Continuation-based Analysis Works

2.1 A Rough Sketch of the Analysis

Before providing a full analysis, I will describe how the proposed analysis works to analyze the lexical item of our interest: *wa*. First, continuations are primarily used to derive a special "focus" semantic value of a sentence involving a contrastive *wa* phrase. The computation of the ordinary semantic value of a sentence can be done using continuations, but the denotation of *wa* does not play a special role. That is reflected in the denotation of *wa* that takes an NP, given below as (8).[6]

(8) *Wa with an NP in the ordinary dimension*
$[\![\, wa \,]\!]^o = \lambda x_e.\lambda c_{\text{NP}}.c_{\text{NP}}(x)$

This *wa* just type-shifts an NP (type e) so that it would have type $(et)t$, which is typically a type assigned to a quantifier. Notably, the type-shift triggered by *wa* does not change the final result of the computation but only the way of computation.

In the focus dimension, on the other hand, *wa* does do a special job. The denotation of contrastive *wa* is given below in (9).[7]

[6]For the sake of space, only *wa* attached to an NP is discussed in this paper. The particle *wa* can be used with other kinds of phrases such as quantifier phrases, and it is possible to have a denotation of *wa* that is generalized so that it could handle any category, XP: $[\![\, \text{wa} \,]\!]^o = \lambda x_\sigma.\lambda c_{\text{XP}}.c_{\text{XP}}(x)$, where c_{XP} is of type σt.

[7]The denotation of non-contrastive *wa* (i.e., *wa* used with an NP without F-marking) in the focus dimension looks exactly the same as (8) except that everything is treated as a set and

$[_{\text{S}}[_{\text{NP}}$ Taro$_{\text{F}}$-wa$][_{\text{VP}}$ came$]]$	
$\rightsquigarrow$ (by a syntactic rule for an S)	$\lambda c_{\text{S}}.\underline{\text{VP}}(\lambda\mathcal{P}_{(et)t}.\underline{\text{NP}}_{\text{F-wa}}(\lambda X_{et}.\ c_{\text{S}}(\mathcal{P}(X))\ ))$
$\rightsquigarrow$ ($\underline{\text{VP}}$=$\lambda c_{\text{VP}}.c_{\text{VP}}(\{\lambda x.\text{come}(x)\})$ and β-reduction)	$\lambda c_{\text{S}}.\underline{\text{NP}}_{\text{F-wa}}(\lambda X.\ c_{\text{S}}(\{\lambda x.\text{come}(x)\}(X))\ )$
$\rightsquigarrow$ ($\underline{\text{NP}}_{\text{F-wa}}$=$\lambda F_{(et)t}.\{F(\{x\})\|x \in D_e\}$ and β-reduction)	$\lambda c_{\text{S}}.\{\ c_{\text{S}}(\{\lambda x.\text{come}(x)\}(\{x\}))\ \|x \in D_e\}$
$\rightsquigarrow$(by PFA)	$\lambda c_{\text{S}}.\{\ c_{\text{S}}(\{\text{come}(x)\})\ \|x \in D_e\}$
$\rightsquigarrow$(by $c_{\text{S}} = \{\lambda p.p\}$ and PFA)	$\{\{\text{come}(x)\}\|x \in D_e\}$
	$\rightsquigarrow \{\{\text{come}(t)\}, \{\text{come}(j)\}, \{\text{come}(h)\}\}$

Table 6: The derivation of the focus semantic value of 'Taro$_{\text{F}}$-wa came.'

(9) *Semantics of Contrastive wa with an NP$_F$*
$[\![\,\text{NP}_{\text{F}}\text{-wa}\,]\!]^f = \lambda F_{(et)t} : \text{NP}_{\text{F}} \supset \{\ [\![\,\text{NP}\,]\!]^o\ \} \wedge F(\text{NP}_{\text{F}}) = S.\{\ F(\{x\})\ |x \in D_e\}$

The denotation in (9) has two parts. First, it has two presuppositions about the attached NP: (i) it needs to have an F-marking, as expressed by showing that the denotation of NP in the focus dimension needs to be a strict superset of that in the ordinary dimension ($\text{NP}_{\text{F}} \supset \{\ [\![\,\text{NP}\,]\!]^o\ \}$), and (ii) the result obtained by combining the NP and its continuation must match a strategy to be employed in context (S). Second, as the focus semantic value of a sentence, NP$_{\text{F}}$-*wa* produces a set of sets of propositions rather than a set of propositions. The operation in (9) is essentially the same as Topic Abstraction in (10).

(10) $[\![\,\text{CT-}\lambda_i\ \phi\,]\!]^f = \{\lambda x.\ [\![\,\phi\,]\!]^f_{g[i\rightarrow x]}\ \}$
(Constant, 2014)

Both *wa* and a CT project a structured discourse and indicate a particular strategy that the interlocutors are entertaining at the time of the utterance.

2.2 Special Focus Semantic Value for Contrastive *wa*

Let us assess how this continuation-based analysis works to derive a special focus semantic value. One of the most important factors to be captured is that contrastive *wa* or a CT indicates a particular strategy. Let us take up a simple example, *Taro$_F$-wa came.* When *wa* is used with Focus as seen in this case, what is indicated is that other alternatives such as *Jiro-wa came.* are possible answers that the speaker could have used. The derivation of the focus semantic value of the sentence *Taro$_F$-wa came.* is given in Table 6.

computation involves PFA.

The final result of the focus semantic value given in Table 6 is different from that of a sentence without *wa*. When *wa* is present, each individual proposition is packed in a set. In other words, we get a set of sets of propositions. By contrast, when *wa* is not used with the F-marked phrase (i.e., when phonological focus only indicates so-called information focus), the result is a set of propositions. This special focus value is the discourse effects associated with contrastive *wa*. Through its discourse effects, *wa* indicates that those alternatives could also be relevant to the Question under Discussion (QuD: Roberts (2012)) entertained at the time of utterance.

Another thing to be captured is the interaction between CTs and informational focus. For example, to answer an overarching QuD, "Who invited whom?," there are two ways to approach the answers, as shown in (11): (i) looking for answers by hosts and (ii) answering by guests. The assignment of a CT and an informational focus (henceforth Focus) varies depending on which strategy the speaker wants to adopt. In English, a different intonational contour is used to distinguish a CT and Focus (Jackendoff, 1972), while *wa* is used to indicate a CT in Japanese.

(11) Who invited whom?

a. A: What about Taro? Who did he invite?
B: Taroo$_{\text{CT}}$-wa Hanako$_{\text{F}}$-o yonda.
TARO$_{\text{CT}}$ invited HANAKO$_{\text{F}}$.

b. A: What about Hanako? Who invited her?
B: HANAKO-wa TAROO-ga yonda.
TARO$_{\text{F}}$ invited HANAKO$_{\text{CT}}$.

What we need to have here is two different focus semantic values for the two different ways of answering the question in (11).

Using a continuized grammar and the semantics of contrastive *wa* given in (9), it is possible to capture such a contrast. For the sake of space, only the final results after feeding a trivial continuation of a sentence are provided below as (12).

(12) a. TARO_{CT} invited HANAKO_{F}
$\{\{\text{invite}(x,y)|y \in D_e\}|x \in D_e\}$
$=\{\{\text{invite}(t,t), \text{invite}(t,j), \text{invite}(t,h)\},$
$\{\text{invite}(j,t), \text{invite}(j,j), \text{invite}(j,h)\},$
$\{\text{invite}(h,t), \text{invite}(h,j), \text{invite}(h,h)\}\}$

b. TARO_{F} invited $\text{HANAKO}_{\text{CT}}$.
$\{\{\text{invite}(x,y)|x \in D_e\}|y \in D_e\}$
$=\{\{\text{invite}(t,t), \text{invite}(j,t), \text{invite}(h,t)\},$
$\{\text{invite}(t,j), \text{invite}(j,j), \text{invite}(h,j)\},$
$\{\text{invite}(t,h), \text{invite}(j,h), \text{invite}(h,h)\}\}$

As we can see, the structures of the two focus semantic values are different. In (12a) it is organized by subject first and then object, while (12b) indicates that a guest-by-guest strategy is employed in the discourse. Note that we have a CT-marked phrase *Hanako* in the object position in (12b). The continuation-based approach pursued here requires no movement of this phrase and can derive the desired focus semantic value in-situ.

We have seen that this continuation-based analysis can give us the desired result — contrastive *wa*, which is a realization of a contrastive topic in Japanese, plays an important role in projecting a particular type of structured discourse. Any analysis needs to explain the behavior of this item. We have seen that the proposed analysis can derive the special focus semantics value without issue, but this analysis can explain more, as will be shown in the next section.

3 Empirical Facts Explained by the Analysis

3.1 Two Kinds of CTs

Typically contrastive *wa* phrases appear at the beginning of the sentence as shown in (11b). This kind of *wa*-phrase seems to occupy a designated topic position. However, this is not the only possible position in which a contrastive *wa*-phrase can appear. It can also appear in the middle of the sentence (in-situ). As Hoji (1985, 131) pointed out, these two kinds of *wa*-phrases present different behaviors when they contain *zibun*. In (13), a *wa*-phrase appears in the designated topic position, and the sentence is ungrammatical under the reading that *zibun* refers to John.

(13) * *sono zibun nituite-no hon-wa John-ga*
that self about book-*wa* John-NOM
suteta.
threw away
'As for that book about himself, John threw it away.'

The ungrammaticality of (13) indicates that the *wa*-phrase in (13) is base-generated in the topic position. Otherwise, the sentence would be grammatical thanks to reconstruction. In (13), it is reasonable to assume that a *pro* occupies the object position of *suteta* 'threw away' and the topic phrase binds the *pro*. Schematically we have two patterns of *wa*-phrases as shown below in (14).

(14) a. *Wa appears in the root clause*
$[_{\text{S}}$ ... XP_{F}-wa ...$]$

b. *Wa appears in the topic position*
$\text{XP}_{\text{F}i}$-wa $[_{\text{S}}$ ... pro_i ... $]$

We have seen continuations can handle (14a) without movement. Now, do we need to have a different lexical entry for the *wa*-phrase in the topic position (14b)? The answer is no, as long as we adopt the treatment of binding with continuations, as discussed in Barker & Shan (2015). It is possible to keep the lexical entry for NP-*wa* untouched by incorporating pronouns in the grammar and assuming that the presence of an unbound pronoun is reflected in the syntactic category. In (14b), the root clause involves *pro*. This sentence is an open proposition (Dowty, 2007), which requires an NP for a complete interpretation. Following Jacobson (1999), I assume this kind of clause has a different semantic type from a clause that does not have any pronouns (e.g., *Taro sneezed.*). An open proposition is of type et since it needs to take the referent of the pronoun for a full interpretation. Pronouns are expressed as identity functions as in (15).

(15) $\underline{\text{NP}} \rightarrow pro$: $\lambda c_{\text{NP}}.\lambda y.c_{\text{NP}}(y)$

With this semantics of pronouns, we can obtain an appropriate denotation for an open proposition. For instance, the ordinary semantic value of *Taro pro in-*

vited is computed as (16).

(16) $[_{S}[_{NP}$ Taro$][_{VP}[_{NP}$ *pro*$]$ invited$]]$
$\rightsquigarrow \lambda c_{S}.\lambda y.c_{S}(\text{invite}(t, y))$

To (16), we need to feed a trivial continuation of a sentence, $\lambda p.p$. As a result, we get $\lambda y.\text{invite}(t, y)$, which is exactly what we want — it is of type et. Remember that NPs with *wa* are of type $(et)t$, which is a function from a continuation of NP to the truth value. As a result, *wa*-NP ends up taking an open proposition as its argument and offers a truth value. The same mechanism works in the focus dimension, too. Overall, it is not necessary to have two different semantic denotations for *wa*-phrases in the topic position and those in-situ. As a result, we can capture the fact that these two kinds of *wa*-phrases function almost in the same way. I said "almost" because their behaviors are not exactly the same. The proposed analysis can provide an explanation of how they could be different. Before discussing it, let us present another relevant behavior of Japanese contrastive *wa*-phrases.

3.2 Multiple CTs

It is well known that contrastive *wa* phrases can appear multiple times in a sentence, as shown in (17) if appropriate context is set. This example is from Yabushita (2017, 25).[8]

(17) *John$_F$-wa* *Mary$_F$-wa* *Bob$_F$-ni-wa*
John-*wa* Mary-*wa* Bob-DAT-*wa*
syookai-si-ta.
introduce-do-PAST

'John$_{CT}$ introduced Mary$_{CT}$ to Bob$_{CT}$.'

An utterance in (17) would be possible when the speaker is asked who introduced whom to whom and trying to answer by looking at the list of people under discussion, for example.

However, this is not a unique characteristic of contrastive *wa* in Japanese. As Constant (2014, 76) pointed out, in English we can have multiple CTs in a sentence as well when appropriate context is set. Constant does offer an analysis of multiple CTs in a sentence, but the analysis requires modifying the basic operation he uses, Topic Abstraction. Furthermore, depending on the number of CTs in a sentence, different rules apply. The proposed analysis of contrastive *wa* in Japanese does not require such modified rules. As long as the multiple *wa* phrases occur in a sentence in canonical word order, it is possible to get a heavily nested focus semantic value without any ado. For the detailed discussion on this issue, please refer to Chapter 3 of Hirayama (2019).

In addition, if we adopt the assumption that evaluation order is left-to-right, we can also assume that the word order also reflects how the heavily nested focus semantic value is organized. In other words, in (17), questions are first ordered by the subject, then the direct object, and finally the indirect object. Such a focus semantic value of a sentence with multiple contrastive *wa* phrases can be derived without adding anything to our denotation of contrastive *wa*. However, our semantic mechanism only works when multiple contrastive *wa* phrases appear in canonical word order. Empirically, this seems to be the case. If we try to scramble the indirect object of (17), for example, the sentence becomes degraded. In the next subsection, I will demonstrate why multiple contrastive *wa* phrases only work in canonical word order from the difference of the semantic roles of two kinds of *wa* phrases discussed in the last subsection.

3.3 Type-mismatch in Split CTs

When we have multiple contrastive *wa*-phrases in a sentence and they appear in canonical word order, we can keep computing nested focus semantic value without encountering any type-mismatches. Contrastive *wa* phrases can successfully take $(et)t$ in the derivation. However, a type-mismatch can happen when a contrastive *wa* phrase is moved to the designated topic position, and the root clause has another CT. Remember that when a contrastive *wa* phrase occupies a designated topic position, the root clause is treated as an open proposition. When there is another CT in the root clause, the type of the root clause and what the contrastive *wa* phrase attempts to take as its argument do not match. For instance, imagine that we are trying to compute the focus se-

[8] If all of the three *wa*-phrases are contrastive as expected in the context given in the text, prosodic prominence would be observed at each *wa*-phrase and only the predicate would undergo post-focal reduction. Yabushita mentions that the first *wa*-phrase can be a thematic, but as an anonymous reviewer pointed out to me, the second *wa*-phrase can also be non-contrastive depending on context. I appreciate their feedback on this.

mantic value of the sentence in (18). The root clause that involves a CT and *pro* ends up with the focus semantic value given in (18a). This is of type $((et)t)t$. Recall that a *wa*-phrase in the focus dimension takes something of type $(et)t$ as its argument. This is how type-mismatch occurs.

(18) *Hanako$_F$-wa* [$_S$ *Taro-$_F$-wa pro invited.*]
Hanako$_{\text{CT}}$, Taro$_{\text{CT}}$ invited her.

a. The final result of the root clause after feeding c_{S}
$\{\lambda Y_{et}.\{\text{invite}(x, y)|y \in Y\}|x \in D_e\}$

b. The semantic denotation of *Hanako$_F$-wa*
$\underline{\text{NP}}_{\text{F-wa}}$=$\lambda F_{(et)t}.\{F(\{x\})|x \in D_e\}$

What does this type-mismatch tell us? A strong prediction is we cannot have split CTs. That is, we cannot have a CT in the topic position and other CTs in the root clause. Note that as far as the root clause does not contain any CTs, the computation can be carried out smoothly. For instance, having informational focus in the root clause is not a problem, for there would be no type-mismatch. Further empirical investigations are required to determine whether having split CTs is really impossible, but it seems that multiple CTs often occur in the canonical word order in a sentence. This empirical fact can be accounted for by the different semantic roles of *wa*-phrases in-situ and those in the topic position. The *wa*-phrase in-situ only works as a function on its continuation, while that in the topic position needs to bind a pronoun in the root clause, in addition to working as a function on its continuation.

However, the hypothesis given above might be too rough. It is true that the computation clashes if there is a type-mismatch. However, natural languages are also equipped with tools that can handle such a problem — type-shift. In fact, the type-shift that would be necessary here is not that complicated. The denotation of NP with contrastive *wa* after type-shift (NP-*wa*$_2$) can be given using what we have in (9) as in (19).

(19) $[\![\,\text{NP-wa}_2\,]\!]^f$ =$\lambda\mathcal{F}.\{\,[\![\,\text{NP-wa}\,]\!]^f\,(F) \mid F \in \mathcal{F}\}$
where $\mathcal{F}$ is of type $((et)t)t$

The result in (19) is mathematically related to the semantics of NP-*wa*; What is given in (19) can be characterized as an image of $\mathcal{F}$ under g, which is the function denoted by NP-*wa*. Generally, the image of a subset $A \subseteq X$ under f is defined as in (20). Using this notation, the function given in (19) can be expressed more simply using g, as in (21). In other words, it is an image of $\mathcal{F}$ under the semantics of contrastive *wa* in-situ.

(20) Image of a subset: $f[A] = \{f(a)|a \in A\}$

(21) $[\![\,\text{NP-wa}_2\,]\!]^f$ =$\lambda\mathcal{F}.g[\mathcal{F}]$ where $\mathcal{F}$ is of type $((et)t)t$ and g is $[\![\,\text{NP-wa}\,]\!]^f$

Remember that when we have multiple CTs, the first one is ambiguous between a CT and a thematic topic. Even when *wa*-phrases are split between the topic position and the root clause, it can always be interpreted as a thematic topic. In such a case, a type-shift such as that illustrated in (21) is necessary, too. As such, a more plausible reason for the ban of split CTs is not just a type-mismatch but the complexity of the operation after type-shift.

As mentioned earlier, an investigation into exactly how these split CT examples are bad is required. However, the proposed analysis has the potential to provide reasoning for the ban of split CTs.

4 Conclusions

In this paper, I showed how we can derive the special semantic values indicated by the use of contrastive *wa* using continuations. It was shown that not only can this continuation-based analysis give us the desired results so as to account for the basic behavior of *wa*-phrases but also it covers a wider range of empirical facts.

This analysis can be extended in order to handle interrogative sentences with contrastive *wa* as well. What to be done is to add words, operators, and syntactic rules necessary to form interrogative sentences. It is well known that Japanese contrastive *wa* can appear in various kinds of sentences (Tomioka, 2009) and sometimes has an important pragmatic effect (Schwarz & Shimoyama, 2010). This continuation-based analysis has potential in that it can be used to explain those other interesting behaviors of contrastive *wa* as well.

Acknowledgments

This paper is based on the content of Chapter 3 of Hirayama (2019). My sincere gratitude goes to my

advisors, Adrian Brasoveanu and Donka Farkas. I am also grateful for helpful comments by people at UC Santa Cruz and two anonymous reviewers of PACLIC 33. I appreciate Yu Tomita for his suggestions for formatting. All errors are my own.

References

Barker, Chris. (2001). Introducing Continuations. In R. Hastings, B. Jackson, & Z. Zvolenszky (Eds.), *Proceedings of salt ix* (p. 20-35). Ithaca, NY: Cornell University.

Barker, Chris, & Shan, Chung-Chieh. (2015). *Continuations and Natural Language*. Oxford University Press.

Büring, Daniel. (2003). On D-Trees, Beans, and B-Accents. *Linguistics and Philosophy*, *26*, 511-545.

Charlow, Simon. (2014). *On the semantics of exceptional scope*. PhD thesis, New York University.

Constant, Noah. (2014). *Contrastive Topics: Meanings and Realizations*. PhD thesis, University of Massachusetts Amherst.

Dowty, David. (2007). Compositionality as empirical problem. In Chris Barker & I. Pauline Jacobson (Eds.), (p. 14-23). Oxford University Press.

Hara, Yurie. (2006). *Japanese Discourse Items at Interfaces*. PhD thesis, University of Delaware.

Hirayama, Hitomi. (2019). *Asking and Answering Questions: Discourse Strategies in Japanese*. PhD thesis, University of California, Santa Cruz.

Hoji, Hajime. (1985). *Logical Form constraints and configurational structures of Japanese*. PhD thesis, University of Washington, Seattle.

Ishihara, Shinichiro. (2003). *Intonation and interface conditions*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Jackendoff, Ray. (1972). *Semantic Interpretation in Generative Grammar*. MIT Press.

Jacobson, Pauline. (1999). Towards a Variable-Free Semantics. *Linguistics and Philosophy*, *22*(2), 117-184.

Kuno, Susumu. (1973). *The Structure of the Japanese Language*. Cambridge: MIT Press.

Kuroda, S.-Y. (2005). Focusing on the matter of topic: a study of *wa* and *ga* in Japanese. *Journal of East Asian Linguistics*, *14*, 1–58.

Oshima, David Y. (2002). *Contrastive topic as a paradigmatic operator.* Workshop on Information Structure in Context, Stuttgart University.

Oshima, David Y. (in press). The English rise-fall-rise contour and the Japanese contrastive particle wa: A uniform account. In *Japanese/Korean Linguistics* (Vol. 26). Stanford: CSLI Publications.

Roberts, Craige. (2012). Information Structure in Discourse: Towards an Integrated Formal Theory of Pragmatics. *Semantics & Pragmatics*, *5*(6), 1-69.

Rooth, Mats. (1985). *Association with Focus*. Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Rooth, Mats. (1996). Focus. In S Lappin (Ed.), *The handbook of contemporary semantic theory* (pp. 271–297). Blackwell.

Schwarz, Bernhard, & Shimoyama, Junko. (2010). Negative Islands and Obviation by *Wa* in Japanese Degree Questions. *Proceedings of SALT*, *20*, 1-18.

Tomioka, Satoshi. (2009). Contrastive topics operate on speech acts. In Malte Zimmermann & Caroline Fèry (Eds.), *Information structure: Theoretical, typological, and experimental perspectives* (pp. 115–138). Oxford University Press.

Yabushita, Katsuhiko. (2017). Partition Semantics and Pragmatics of Contrastive Topic. In Chungmin Lee, Ferenc Kiefer, & Manfred Krifka (Eds.), *Contrastiveness in Information Structure, Alternatives and Scalar Implicatures* (Vol. 91, p. 23-45). Springer.

Effects of Prosodic Focus on Voice Onset Time (VOT) in Chongming Chinese

Yitian Hong
The Chinese University of Hong Kong,
Shatin, NT, Hong Kong SAR, China
hongyt@link.cuhk.edu.hk

Si Chen
The Hong Kong Polytechnic University,
Kowloon, Hong Kong SAR, China
sarahchen@polyu.edu.hk

Yike Yang
The Hong Kong Polytechnic University,
Kowloon, Hong Kong SAR, China
yi-ke.yang@connect.polyu.hk

Bei Li
The Hong Kong Polytechnic University,
Kowloon, Hong Kong SAR, China
benita.li@polyu.edu.hk

Abstract

Prosodic focus is phonetically realized by increasing intensity, extending duration, and expanding pitch rage of focused components (Xu et al., 2012). Previous studies have also found the effect of prosodic focus on enlarging two-way or three-way stop contrast by lengthening the VOT (voice onset time) of voiceless or aspirated consonants (e.g. Choi, 2003; Chen, 2011). The present study investigates the influence of prosodic focus on the realization of VOT of an under-studied language, Chongming Chinese. Twelve monosyllabic words were selected and embedded in carrier sentences with different discourse conditions: one baseline neutral focus condition and three focus conditions. Precursor questions were prepared to elicit production from native speakers of Chongming Chinese. Results showed the significant main effects of stop types and discourse conditions on VOT realization. VOTs were shortened in unaspirated and voiced stops in the on-focus condition, suggesting a different way of expanding the three-way difference in stops. VOT was also affected by other focus conditions, providing implications for the study of focus domain. The study suggests that VOT can serve as acoustic cue for stop contrasts in Chongming Chinese in different prosodic environment and contributes new data to the typology of prosodic focus study as well as stop contrast research.

Index Terms: prosodic focus, VOT, phonological contrast, Chongming Chinese

1 Introduction

Prosodic focus refers to the use of speech prosody to emphasize a specific part of an utterance (Xu et al., 2012). Research across world languages shows that prosodic focus is realized with various acoustic cues including raising intensity, elongating duration, and expanding pitch range of the focused elements (e.g., Jong, 2004; Chen, 2011; Alzaidi et al., 2019). Accumulative evidence has found post-focus realization, mostly refers to compressed pitch range after the focused elements (e.g., Chen et al., 2009; Lee and Xu, 2018; Xu, 2011). The reduction of intensity was also found in post-focus condition (Chen et al., 2009). Several studies also investigated the pre-focus region of the utterance. Pre-focus pitch compression was reported in a Japanese study in both information focus and contrastive focus without plausible explanation (Hwang, 2012). Tone 3 in Mandarin Chinese also demonstrated the increase of duration, intensity and raising pitch of pre-focused components (Lee, 2015). However, the observations of both post-focus and pre-focus encodings vary from languages to languages (e.g., Lee and Xu, 2018; Xu et al., 2012)

Voice Onset Time (VOT) has been defined as "the interval between the release of the stop and the onset of glottal vibration" (Lisker and Abramson, 1964: 389). It was found to be a relatively reliable acoustic measurement for differentiating phonemic categories (i.e., voicing, aspiration, and force of articulation) of stops in a language (Lisker and Abramson, 1964). Factors that are likely to affect

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 414-422, Hakodate, Japan, September 13-15, 2019

VOT include the following vowel, features of the stops, and the prosodic environment. VOT is sensitive to the duration of the following vowel (Ling and Liang, 2016). Tense vowels were more likely to be preceded by longer VOT and lax vowels by shorter VOT (Port and Rotunno, 1979). The place of articulation of stops also plays a role in VOT. The longest VOT is found in velar stops, while the difference of VOT between alveolar and bilabial stops varies across languages (Ling and Liang, 2016; Lisker and Abramson, 1964).

In addition to the place of articulation and the following vowel duration, previous studies also examined the effects of prosodic focus on VOT. In the study of English, Choi (2003) reported that both voiced and voiceless stops demonstrated an apparent increase of VOT under focus, particularly voiceless stops. Their difference was also enlarged, suggesting the enhancement of stop voicing contrast. However, Cho and McQueen (2005) reported different findings in Dutch. The VOT of voiced stops was increased under focus while the VOT of voiceless stops was shortened. Dutch has a shorter VOT in voiceless stops than English. The increase of VOT in English in an on-focus condition signaled the enhancement of a [+aspirated] feature, while the reduced VOT in Dutch indicated the strengthening of the [-aspirated] feature (Ling and Liang, 2016).

Similar studies were conducted in a few languages with three-way contrastive stops, especially in Shanghai Chinese, a Wu dialect. Chen (2011) found significant lengthening of VOT in aspirated stops under the influence of focus, but not in unaspirated stops and voiced stops. Ling and Liang (2016) reported a significant increase on VOT in both aspirated and voiced stops of Shanghai Chinese in an on-focus condition, which was explained by maximizing phonological contrast among the three stop types. Cross-linguistic differences on the encoding of voicing and aspiration contrast in focus conditions are observed. However, little is known about the realization of VOT in various focus conditions across languages. Therefore, the present study aims to contribute novel data to the study of prosodic focus and phonological contrast by investigating the encodings of VOT in a Chinese dialect under various focus conditions.

Chongming Chinese is a tonal language and also a variety of Wu dialect. It is spoken by people living in eastern China, including Chongming County, Haimen City, Qidong City, Shazhou County, etc. (S. Chen, 2014). Similar to other Wu dialects, the three-way contrasts in the onset stops are reported (i.e., voiceless aspirated, voiceless unaspirated, and voiced). Previous studies have shown the effects of focus on F0, duration and intensity in Chongming Chinese (Yang et al., 2018, 2019), indicating that Chongming Chinese is a language with noticeable encodings of prosodic focus. The current study will examine a new perspective of focus realization in Chongming Chinese to enrich the study of the Wu dialect.

Based on the above review, three research questions are raised:

1) What are the differences among the VOT of stop types in neutral focus condition and focus conditions?
2) What effects of focus on the VOT of target stops can be found?
3) Is manipulation of phonological contrast observed in the VOT of Chongming Chinese under focus conditions? If yes, how is it achieved?

Provided previous findings of the realization of prosodic focus in Chinese languages as reviewed above, our prediction is that there will be significant differences among three stop types between neutral focus and focus conditions in order to maximize their phonological contrast.

2 Methods

2.1 Subjects

Twelve Chongming Chinese native speakers (six males, six females), aged 38 to 57 (mean ± SD: 52.00 ± 4.53), were recruited for the current study. According to their self-reports, Chongming Chinese is their mother tongue and dominant language for daily communication. They have never received any formal musical training and none of them reported speaking, hearing or language difficulties.

2.2 Stimuli

Twelve monosyllabic words varying in tones and initial stop types were selected as stimuli, as is illustrated in Table 1. Only one vowel was embedded in the stimuli in order to control the effect of following vowel duration. The vowel /æ/ was selected because it was found to appear in most tones for every consonant onset (S. Chen, 2014).

	t	t^h	d
T1 (55)	tæ	t^hæ	
T2 (24)			dæ
T3 (424)	tæ	t^hæ	
T4 (242)			dæ
T5 (33)	tæ	t^hæ	
T6 (313)			dæ
T7 (5)	tæ	t^hæ	
T8 (2)			dæ

Table 1. *Target Stimuli*

Only alveolar stop was adopted to control the effect of place of articulation.

These stimuli were embedded in carrier sentences where contexts and discourse conditions were manipulated [illustrated in (1)]. The tones of the syllables before and after the target stimuli were controlled by selecting two syllables respectively before (i.e., Part C) and after (i.e., Part D) the target stimuli. Four combinations were generated (i.e., /ɦuo313/ and /ɦnø24/, /ɕiɑ424/ and /ɦnø24/, /ɕiɑ424/ and / ɦnø24/, /ɕiɑ424/ and /ɦnø24/).

(1)

ɦmei24ɦain24 *ɦuo313* *ɦuo313/ɕia424*
matchmaker say say/write
(A) (B) (C)
TARGET *ɦnø24/tu55* *ɦlɛ24tɕiɔ55kuæ55*
TARGET difficultly/much very
(D) (E)

'The matchmaker said that she said/wrote TARGET far difficultly/more.'

Four discourse conditions were employed in the carrier sentences: neutral focus condition, pre-focus condition, on-focus condition and post-focus condition, which signaled the position of the target stimuli, as shown in Table 2.

2.3 Procedure

All the subjects were recorded in a quiet room in Qidong City. One PC demonstrated stimuli in E-prime (Schneider et al., 2012) for subjects' reference and another PC were used for recording by Praat (Boersma and Weenink, 2001).

In the neutral focus condition, subjects were instructed to read the carrier sentences in natural and normal speech. In three focus conditions, precursor questions were prepared for eliciting responses (the carrier sentences in Table 2). There were 2304 total tokens produced (12 target stimuli * 4 contexts * 4 focus conditions * 12 speakers).

Discourse conditions	Carrier sentences
neutral focus	(A) (B) (C) TARGET (D) (E)
pre-focus	(A) (B) ***(C)*** TARGET (D) (E)
on-focus	(A) (B) (C) ***TARGET*** (D) (E)
post-focus	(A) (B) (C) TARGET ***(D)*** (E)

Table 2. *Carrier sentences designed in different discourse conditions (The foci are in bold and italic)*

2.4 Data Analysis

The consonants of the target stimuli were manually segmented by one trained phonetician in Praat (Boersma and Weenink, 2001) and checked by the other phonetician. The VOTs were labeled from the point of the stop release to the onset of the second formant of the preceded vowels. The Praat script ProsodyPro (Xu, 2013) was used for extracting VOT values. 112 tokens were suspected as incorrect production (i.e., produced in incorrect tones or had abnormal VOT) and confirmed by a native speaker. There were excluded from further analysis.

By plotting all the VOT values, we observed apparent inter-speaker variations. To investigate factors that significantly affect the VOTs, we fitted a basic linear mixed effect model to VOTs with subject as random effect by adopting the 'lmerTest Package' (Kuznetsova et al., 2017) in R (R Core Team, 2013). By adding stop types, discourse conditions and their interaction as fixed effects one after another, we improved the model. Next, we examined the contribution of stop types and discourse conditions, respectively, by fitting the linear mixed effect model again. The above model is sufficient to interpret the results.

Furthermore, we calculated the differences between different stops types in VOTs across four discourse conditions and compared them by plotting. All the figures were plotted by the 'ggplot2' package (Wickham, 2016) in R.

Fixed effect	Estimate	SE	*t*	*P*
(Intercept)	18.222	1.700	10.718	<0.001***
Stop type: unaspirated	-6.376	1.141	-5.586	<0.001***
Stop type: aspirated	20.169	1.159	43.289	<0.001***
Discourse condition: on-focus	-3.078	1.160	-2.653	<0.01**
Discourse condition: post-focus	-3.649	1.162	-3.140	<0.01**
Discourse condition: pre-focus	-3.515	1.155	-3.042	<0.01**

Signif. codes: ***p<0.001, **p<0.01, *p<0.05 (the same across the whole paper)

Table 3. *Linear mixed model of VOT (2192 observations) (Voiced stops and neutral focus as baseline)*

		voiced vs. unaspirated	voiced vs. aspirated	aspirated vs. unaspirated
Neutral focus condition		t = -5.461, p<0.001	t = 42.145, p<0.001	t = -47.86, p<0.001
Focus conditions	pre-focus	t = -3.276, p<0.01	t = 49.018, p<0.001	t = -52.90, p<0.001
	on-focus	t = -3.750, p<0.001	t = 45.341, p<0.001	t = -49.80, p<0.001
	post-focus	t = -2.971, p<0.01	t = 45.328, p<0.001	t = -48.75, p<0.001

Table 4. *t statistics and p-values in the linear mixed model of VOT in different discourse conditions*

3 Results

The VOTs extracted from the target stops were analyzed by a linear mixed effect model with subject as a random effect. The basic model was improved by adding Stop type (χ^2=4066.7, Df=2, p<0.001) and Discourse condition (χ^2=10.067, Df=3, p<0.05) as fixed effects. However, their interaction did not play a significant role (χ^2=8.9065, Df=6, p=0.1789), indicating that contrast of three stop types remained identical in different discourse conditions. The significant results are reported in Table 3. From Table 3, it can be inferred that the VOTs of unaspirated stops are significantly different from voiced stops. The difference between aspirated stops and voiced stops also reached significance. In terms of the discourse condition, the difference of neutral focus vs. on-focus, neutral focus vs. post-focus, and neutral focus vs. pre-focus reached significance.

We first analyzed the difference of stop types in four discourse conditions respectively. Results showed that the effect of stop type is significant across discourse conditions (Table 4). Pairwise significant differences were found among voiced, unaspirated, and aspirated stops when they were in neutral focus sentences or as pre-focused, on-focused, and post-focused elements of focused sentences.

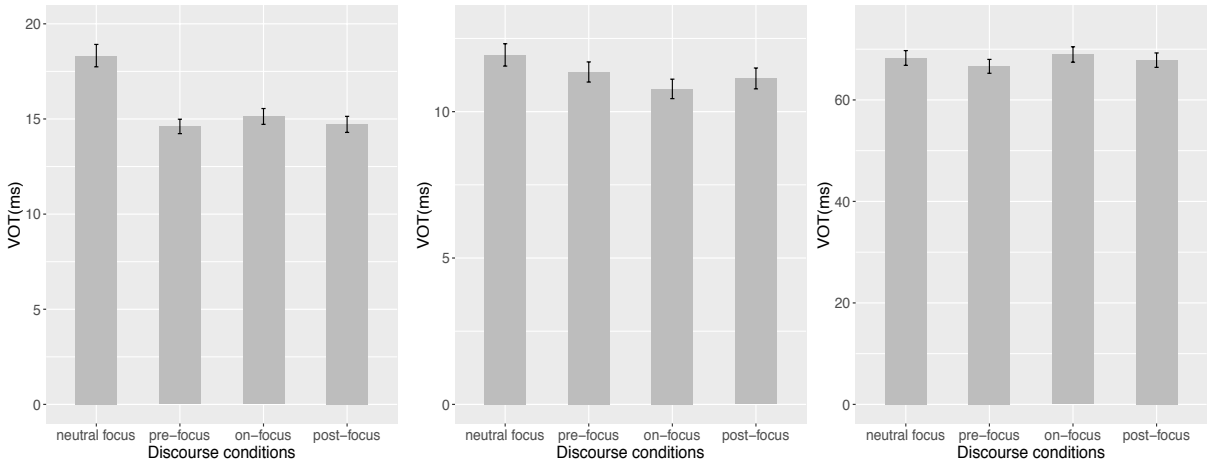
The results indicate that the differences among the stop types remains relatively stable in neutral focus condition and focus conditions. Three stop types were distinguishable from each other in VOT regardless of discourse conditions.

We then analyzed each stop type individually. The main effect of the discourse condition was found in the VOT of voiced stops, as analyzed below. The VOT of unaspirated stop also showed a significant difference between the neutral focus condition and the on-focus condition. No significant difference was found in the VOT of aspirated stop between neutral focus condition and any focus conditions.

When the voiced target syllable is in a pre-focus condition, its VOT is significantly shorter than in a neutral focus condition. The same significant differences were found in the on-focus vs. neutral focus and the post-focus vs. neutral focus condition (Table 5). The mean VOT of voiced stops in the on-focus condition decreases 17.54% compared to that in the neutral focus condition, while the mean VOT

Fixed effect	Estimate	SE	*t*	*p*
Intercept	18.2521	0.8791	20.763	<0.001 ***
pre-focus	-3.6721	0.5878	-6.247	<0.001 ***
on-focus	-3.2018	0.5902	-5.425	<0.001 ***
post-focus	-3.6383	0.5911	-6.155	<0.001 ***

Table 5. *Linear mixed model comparing VOT of voiced stop across discourse conditions (neutral focus condition as baseline)*



(a). *Voiced stops* (b). *Unaspirated stops* (c). *Aspirated stops*

Figure1. *Mean VOT of different stop types*

in pre-focus and post-focus conditions drops 20.12% and 19.93%, respectively, as shown in Figure 1(a). The figure indicates that when there is a focus before or after the target syllable, the VOT of the target voiced stop tends to be reduced more to signal the coming focus. When the focus is exactly the target syllable, the VOT also reduces, but to a lesser degree.

For unaspirated target stops, a significant difference was found between the VOT in the neutral focus condition and the on-focus condition (t= -2.450, p <0.05*), suggesting that speakers reduced the VOT of the target syllable when it was the focus of the sentence. A marginally significant difference was found between the VOT in the post-focus condition and the neutral focus condition (t= -1.754, p= 0.0798). Figure 1(b) shows the mean VOT of unaspirated stops. Similar to voiced stops, the mean VOT of unaspirated target stops is shorter in all the focus conditions, among which, the VOT in the on-focus condition drops the most. Speakers tended to shorten the VOT of the target syllables to indicate that a focus was addressed in the sentence.

Aspirated target stop demonstrates a different pattern [Figure 1(c)]. The on-focus condition has the longest mean VOT (68.96ms), longer than the neutral focus condition. The VOT of aspirated stops in the post-focus condition is slightly shorter than that in the neutral focus condition, while the VOT in the pre-focus condition remains the shortest (66.62ms). Although the differences do not reach significance, when the target syllable is the focus, a trend for subjects to lengthen their VOT is observed. Subjects tended to emphasize the on-focus target aspirated stop by extending the VOT. It is also possible that speakers tried to differentiate an aspirated stop from the other two stop types by enlarging their VOT difference in the on-focus condition.

For further comparison, we calculated the mean differences among the three stop types in neutral focus condition and focus conditions by a two-two subtraction. The results are shown in Figure 2. The differences between aspirated and voiced stops increase in all the focus conditions compared to the neutral focus condition. On the contrary, the difference between unaspirated and voiced stops decreases in all the focus conditions. The difference between aspirated and unaspirated stops enlarges in the on-focus condition, while it reduces in the pre-focus condition, in comparison with the neutral

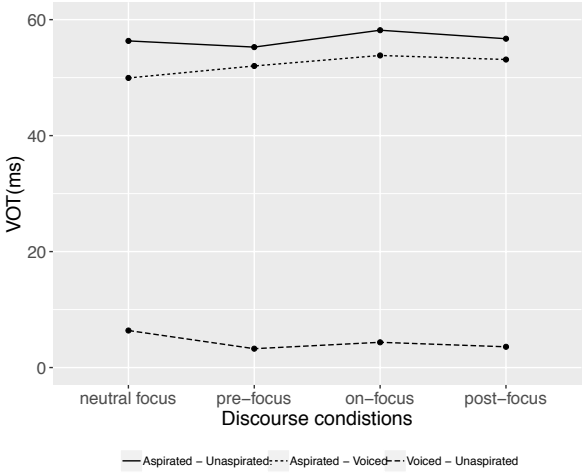


Figure 2. *Mean differences between stop types*

focus condition. The results suggest that, when a focus is indicated in the sentence, subjects are more likely to distinguish an aspirated target stop from unaspirated and voiced target stops. They reduced the distance of VOT between voiced and unaspirated target stops and increased the differences between aspirated and unaspirated and aspirated and voiced target stops. This tendency is more obvious when the target stops are the on-focused elements of the sentences.

The main findings of the current study are summarized as follows: 1) stop types show a significant effect on the VOT of target stops in neutral focus condition and three focus conditions; 2) in regards to discourse conditions, significant differences were found in the neutral focus vs. on-focus, the neutral focus vs. post-focus, the neutral focus vs. pre-focus condition of voiced stops and the neutral focus vs. on-focus condition of unaspirated stops; 3) compared to the neutral focus condition, the differences between VOT of aspirated and unaspirated and aspirated and voiced stops are enlarged in focus conditions, especially in the on-focus condition.

4 Discussion

The present study investigates the effects of stop types and prosodic focus on the VOT of Chongming Chinese. The results showed pairwise significant differences among VOTs of voiced, aspirated, and unaspirated stops in the neutral focus condition and the focus conditions. The VOT of voiced stops was significantly different between the neutral focus condition and the pre-focus/on-focus/post-focus conditions. For unaspirated stops, the difference between the VOT in the neutral focus condition and the on-focus condition also reached significance.

For Chongming Chinese, the VOT can distinguish the three-way contrast of stops. The difference of VOTs among stop types remained stable regardless of focus condition, which suggests that the manipulation of prosodic structure does not change the basic three-way distinction among stops in Chongming Chinese. The [+voiced] and the [+aspiration] features of stops are distinguishable from the measurement of VOT.

The VOT of voiced stops lies between the VOTs of aspirated stops and unaspirated stops in the current study, which may reveal its breathy nature, as indicated in Z. Chen (2014). It is commonly seen in Wu dialects that the voiced stops are pronounced in the manner of a weak voiceless onset proceeding a phonated breathy vowel (Ibid.). The VOTs of voiced stops were significantly longer than unaspirated stops in all the discourse conditions, indicating that VOT has the possibility of acting as a reliable acoustic cue for differentiating voiced stops and unaspirated stops in Chongming Chinese. It contradicts the general claim that in the Wu dialect, the voice feature and F0 shown in the following vowels are the major acoustic cues for distinguishing stops due to the similarity of VOTs between unaspirated and voiced stops (Ling and Liang, 2016; Z. Chen, 2014).

When the target syllables were on-focused, the VOTs of voiced and unaspirated stops were significantly shorter than VOTs in neutral focus condition, while the VOT of aspirated consonants is lengthened in the on-focus condition but did not reach significance. By doing so, the differences between unaspirated and aspirated stops and voiced and aspirated stops were enlarged. Our findings are consistent with the findings in other languages in that in the positions with prosodic emphasis, the phonetic components tend to be realized with the aim of enlarging phonological contrast (Chen, 2011; Cho and Keating, 2001; Choi, 2003). VOT is used for measuring both voicing and aspiration (Lisker and Abramson, 1964). For aspirated and unaspirated stops, the [-aspiration] feature is enhanced via the shortening of the VOT in the unaspirated stop to emphasize its difference from the aspirated stop. For voiced and aspirated stops, the [+voice] feature is strengthened in the voiced stop to reinforce its contrast with the aspirated stop. Therefore, the lexical contrast can be enhanced under focus. Moreover, the goal of spreading new information by providing focus in the utterance can be achieved.

Part of the results is not in line with the findings of Shanghai Chinese (Chen, 2011; Ling and Liang, 2016), in which aspirated stops and voiced stops increased the VOT significantly to signal the on-focus condition, while unaspirated stops remained stable. Because the prosodically conditioned lengthening effect (Cho and McQueen, 2005) is not apparent in the VOT of aspirated stops, Chongming Chinese speakers have to shorten the VOTs of unaspirated and voiced stops to compensate for maximizing the three-way stop contrast.

To find a possible explanation for the above inconsistent findings, we conducted a search and calculation of the *Chongming Fangyan Cidian* ('Dictionary of Chongming Dialect') (Zhang, 1993). We found that there is a total of 268 syllables with unaspirated stops, 174 syllables with voiced stops and 118 syllables with aspirated stops. All these syllables have independent lexical meanings in Chongming Chinese. The number of syllables with aspirated stops is smaller than the number of syllables with the other two stop types. It indicates that the frequency of appearance of voiced and unaspirated stops is much higher than aspirated stops in Chongming Chinese. Due to the lower frequency of occurrence of aspirated stops, speakers have less chance to practice manipulating their VOTs in different prosodic environments. They tend to rely more on adjusting the VOTs of the other two stop types to maintain the contrast, which may also explain why the increase of VOTs in on-focused aspirated stops was not significant.

The performance of voiced stops in focus conditions is unexpected. All the focus conditions demonstrated significant difference from neutral focus condition, suggesting that the voiced stops may play an extremely important role in maintaining stop contrast in Chongming Chinese.

As we discussed above, voiced stops in Chongming Chinese are presumably breathy stops. Breathy phonation refers to a situation when the vocal folds are both opening and vibrating (Davenport and Hannahs, 2013). Due to the escape of air, the energy for vibration is reduced and thus the [+voice] feature is not robust (Ibid.). When a focus is addressed, it is likely that Chongming Chinese speakers tried to compress the breathy nature of voiced stops by controlling the escape of air from the vocal folds and the VOTs were thus affected. By compressing the aspiration of breathy voiced stops, the [+voice] feature is strengthened and thus the voiced stops can distinguish themselves in the on-focus condition.

Another interesting issue revealed in our study is that the contrast between voiced and unaspirated stops was reduced in the on-focus condition, which seems to contradict the phonetic contract enhancement finding. It is likely that in Chongming Chinese, it is less important to draw the difference between voiced and unaspirated stops. We calculated all the minimal pairs for stops in the *Chongming Fangyan Cidian* ('Dictionary of Chongming Dialect') (Zhang, 1993). We found 317 minimal pairs for syllables with aspirated stops and unaspirated stops. Except for the difference in initial consonants (/p/ vs. /p^h/, /t/ vs. /t^h/, /k/ vs. /k^h/), their vowels and tones remain the same. In a similar manner, we found only 51 minimal pairs for syllables with voiced stops and unaspirated stops (/p/ vs. /b/, /t/ vs. /d/, /k/ vs. /g/). Fewer minimal pairs suggest the possibility of sacrificing the contrast between voiced and unaspirated stops and adding more efforts in distinguishing aspirated stops.

This study also tried to examine the target stops in pre-focus and post-focus positions. The post-focused influence was witnessed in voiced stops with a significant drop in the VOT from the neutral focus condition. Unaspirated stops also showed a marginally significant reduction of VOT in post-focus positions compared to the neutral focus condition. Previous studies have found that in a Verb Phrase (VP), when the initial verb was focused, the other arguments inside the VP (i.e., the oblique and thematic arguments) received prominence as well (Jun et al., 2006; Jun, 2011). The on-focused verb had the most robust emphasis (Ibid.). Referring back to the carrier sentence (1), the target syllable is the object of a VP, *ɦuo313/ɕia424 TARGET* ('say/write TARGET'). When the verb *ɦuo313/ɕia424* is on-focused, it is possible that the focus domain extends to the whole VP. As a post-focused component, as well as an argument of the VP, the target syllable may receive a certain degree of emphasis. Therefore, it is not surprising that in the post-focus condition, voiced and unaspirated stops still showed significant reduction to enlarge the three-way contrast in stops. This finding is in line with previous study of duration and intensity range change in post-focus condition in Chongming Chinese (Yang et al., 2019).

The VOTs were shortened in pre-focused words across all types of onset stops, among which, the reduction in voiced stops reached significance. The distinction between the voiced and unaspirated stops and the aspirated and unaspirated stops are reduced. Thus, it is hypothesized that speakers reduced the three-way stop contrast to differentiate the pre-focused elements from the focused elements and to signal the coming focus. It is also likely that speakers tried to save time and energy to produce the focused syllables. More studies should be

carried out to investigate the pre-focused items and test these hypotheses.

5 Conclusion

The present study examined the effect of prosodic focus on stops in Chongming Chinese. In the on-focus condition, the phonological contrast between voiced and aspirated stops and unaspirated and aspirated stops were enlarged to indicate their lexical contrast. Post-focus influence was found in the VOT of voiced and unaspirated stops, suggesting that the domain of the VP focus may contain not only the verb but other arguments. Pre-focus adjustments of VOT were also found, which is suspected as preparations for the following focus. Further investigation is needed. Our study verifies that the influence of prosodic focus demonstrates cross-linguistic differences. Aspirated stops remain relatively stable in different focus conditions, revealing their use in Chongming Chinese is in lower frequency. Voiced stops demonstrated the feature of breathy phonation, which may explain its significant manipulation in focus conditions.

As the goal of the current study is to demonstrate some acoustic cues for differentiating different focus conditions, further perception study should be carried out to examine the link between perception and production and testify whether the differences of VOT in different focus conditions can actually be used as acoustic cues in human perception. It is suggested from other studies that other acoustic cues such as F0, also played a supplementary role in differentiating phonological categories. Future study should also consider other acoustic cues and compare the results with the current study.

In addition, due to the language-specific feature in both the VOT and the realization of prosodic focus, more languages should be involved in study and contribute new findings. More attention should be paid on the pre-focused region of utterances and seek more convincing explanation.

Acknowledgments

This study was funded by a research grant from Faculty of Humanities, the Hong Kong Polytechnic University (grant number: 1-ZVHJ). We would like to thank all the informants and the three anonymous reviewers.

References

Alzaidi M. Swaileh, Yi Xu and Anqi Xu. 2019. Prosodic encoding of focus in Hijazi Arabic. Speech Communication, 106, 127-149.

Boersma Paul. 2002. Praat, a system for doing phonetics by computer. Glot international, 5.

Chen Si. 2014. A Phonetic and Phonological Investigation of the Tone System of Chongming Chinese. [electronic resource]. University of Florida. Retrieved from http://search.ebscohost.com/login.aspx?direct=true&AuthType=ip,uid&db=cat04364a&AN=ufl.033650748&site=eds-live

Chen Szu-wei, Bei Wang and Yi Xu. 2009. Closely related languages, different ways of realizing focus. In Tenth Annual Conference of the International Speech Communication Association.

Chen Zhongmin. 2014. On the relationship between tones and initials of the dialects in the Shanghai area. In proceedings of 4th International Symposium on Tonal Aspects of Languages 2014,116-119.

Chen Yiya. 2011. How does phonology guide phonetics in segment– f0 interaction? Journal of Phonetics, 39(4), 612-625.

Cho Taehong and Patricia A. Keating. 2001. Articulatory and acoustic studies on domain-initial strengthening in Korean. Journal of phonetics, 29(2), 155-190.

Cho Taehong and James M. McQueen. 2005. Prosodic influences on consonant production in Dutch: Effects of prosodic boundaries, phrasal accent and lexical stress. Journal of Phonetics, 33(2), 121-157.

Choi Hansook. 2003. Prosody induced acoustic variation in English stop consonants. In Proceedings of the 15th International Conference of Phonetic Sciences 2003, 261-264.

Davenport Mike and Stephen J. Hannahs. 2013. Introducing phonetics and phonology. Routledge.

De Jong Kenneth. 2004. Stress, lexical focus, and segmental focus in English: Patterns of variation in vowel duration. Journal of Phonetics, 32(4), 493-516.

Hwang H. Kyung. 2012. Asymmetries between production, perception and comprehension of focus types in Japanese. In Proceedings of the 6th International Conference on Speech Prosody 2012, 614-644.

Jun Sun-Ah. 2011. Prosodic markings of complex NP focus, syntax, and the pre-/post-focus string. In Proceedings of the 28th West Coast Conference on Formal Linguistics (pp. 214-230). Somerville, MA: Cascadilla Press.

Jun Sun-Ah, Hee-Sun Kim, Hyuck-Joon Lee and Jong-Bok Kim. 2006. An experimental study on the effect

of argument structure on VP focus. Korean Linguistics, 13(1), 89-113.
Kuznetsova Alexandra, Per B. Brockhoff and Rune Haubo Bojesen Christensen. 2017. “lmerTest Package: Tests in Linear Mixed Effects Models.” Journal of Statistical Software, 82(13), 1–26.
Lee Albert and Yi Xu. 2018. Conditional realisation of post-focus compression in Japanese. In Proceedings of the 9th International Conference on Speech Prosody 2018, 216-219.
Lee Yong-cheol. 2015. Prosodic Focus within and across Languages, (Doctoral dissertation). Available from ProQuest Dissertations and Theses database.
Ling Bijun and Jie Liang. 2016. The influence of syllable structure and prosodic strengthening on consonant production in Shanghai Chinese. In: 2016 10th International Symposium on Chinese Spoken Language Processing (ISCSLP), 1-5.
Lisker Leigh and Arthur S. Abramson. 1964. A Cross-Language Study of Voicing in Initial Stops: Acoustical Measurements. WORD, 20(3), 384-422.
Port Robert F. and Rosemarie Rotunno. 1979. Relation between voice-onset time and vowel duration. Journal of the Acoustical Society of America, 66(3), 654-662.
R Core Team. 2013. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.URL http://www.R-project.org/.
Schneider Walter, Amy Eschman and Anthony Zuccolotto. 2002. E-Prime user's guide. Pittsburgh: Psychology Software Tools.
Wickham Hadley. 2016. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York. ISBN 978-3-319-24277-4, http://ggplot2.org.
Xu Yi. 2011. Post-focus Compression: Cross-linguistic Distribution and Historical Origin. In Proceedings of the 7th International Conference on Phonetics Sciences 2018, 152-155.
Xu Yi. 2013. ProsodyPro — A Tool for Large-scale Systematic Prosody Analysis. In Tools and Resources for the Analysis of Speech Prosody (pp.7 – 10). Aix-en- Provence, France: Laboratoire Parole et Langage.
Xu Yi, Szu-Wei Chen and Bei Wang. 2012. Prosodic focus with and without post-focus compression: A typological divide within the same language family? The Linguistic Review, 29(1), 131-147.
Yang Yike, Si Chen and Kechun Li. 2018. Pitch realization of post-focus components in Chongming Chinese. The Journal of the Acoustical Society of America, 144(3): 1938-1938.
Yang Yike, Si Chen and Kechun Li. 2019. Effects of Focus on Duration and Intensity in Chongming Chinese. In Proceedings of ICPhS 2019, Melbourne.
Zhang Huiying. 1993 Chongming Fangyan Cidian [electronic resource]. Retrieved from http://img.chinamaxx.net.easyaccess2.lib.cuhk.edu.hk/n/abroad/hwbook/chinamaxx/

Web Page Segmentation for Non Visual Skimming

Judith Jeyafreeda Andrew, Stephane Ferrari, Fabrice Maurel, Gaël Dias and Emmanuel Giguet
Normandie Univ, UNICAEN, ENSICAEN, CNRS, GREYC
14000 Caen, France
Email: {judith-jeyafreeda.andrew, stephane.ferrari, fabrice.maurel, gael.dias,emmanuel.giguet}@unicaen.fr

Abstract

Web page segmentation aims to break a page into smaller blocks, in which contents with coherent semantics are kept together. Examples of tasks targeted by such a technique are advertisement detection or main content extraction. In this paper, we study different segmentation strategies for the task of *non visual skimming*. For that purpose, we consider web page segmentation as a clustering problem of visual elements, where (1) all visual elements must be clustered, (2) a fixed number of clusters must be discovered, and (3) the elements of a cluster should be visually connected. Therefore, we study three different algorithms that comply to these constraints: K-means, F-K-means, and Guided Expansion. Evaluation shows that Guided Expansion evidences statistically-relevant results in terms of compactness and separateness, and satisfies more logical constraints when compared to the other strategies.

1 Introduction

Skimming and scanning are two well-known reading processes, which are combined to access the document content as quickly and efficiently as possible. Scanning refers to the process of searching for a specific piece of information, and skimming is the action of passing through a document in a first glance to get an overview of its content. Skimming can easily be applied in a visual environment thanks to the visual, logical or textual document structure. Indeed, visual skimming relies on contrasted effects related to layout rendering and typographic styles. However, these effects are not available in a non visual environment. As such, reproducing the document content driven by its structure in a non visual setting is a much harder problem, but essential to be solved to improve web accessibility, for the visually impaired, for instance.

In this paper, we focus on the hypothesis that successful non visual skimming strategies can take advantage of a prior identification of the coarse-grained document structure. This specific task is known as Web Page Segmentation (WPS). WPS aims to break a page into zones that appear semantically coherent. A large number of approaches have been proposed to automate this process (Sanoja and Gançarski, 2014; Cai et al., 2003a; Zeleny et al., 2017). However, they deal with tasks that imply constraints far from ours. In our TAG THUNDER project[1], we consider that non visual skimming requires three characteristics to be filled.

First, the number of zones has to be fixed in order to foster the emergence of regularities in the output and to comply with the maximum number of concurrent oral stimuli a human-being can cognitively distinguish. Indeed, we assume that each semantically coherent zone can be summarized and simultaneously synthesized into spatialized concurrent speech acts. Within this context, (Guerreiro and Gonçalves, 2015; Manishina et al., 2016) have shown that the cognitive load can rise up to five different stimuli, thus limiting the number of zones resulting from the WPS process. The 5-zone WPS should also ease the association of a particular sound position to the logical function of the zone in a given

[1] https://tagthunder.greyc.fr/

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 423-431, Hakodate, Japan, September 13-15, 2019

web page. As a consequence, it may enable the advent of new non visual reading strategies. *Second*, each zone should be associated to a unique sound source spatially located in accordance with its position in the web page. Thus, each zone should be a single compact block made of contiguous web elements, and the zones should not overlap. *Third*, segmentation must be complete, which means that no web page element should remain outside a given zone, as the objective is to reveal the overall semantics of a document and not just parts of it, opposite to advertisement withdrawal for example.

In this paper, we study three different algorithms that comply to these constraints: the classical k-means (MacQueen, 1967), the F-K-means (a variant of K-means, which introduces the notion of force between elements instead of the euclidean distance), and the Guided Expansion algorithm (GE), which follows a propagation strategy including alignment constraints. A manual evaluation of the three algorithms is performed by three experts measuring two clustering indicators: compactness and separateness. However, human evaluation may be subject to bias as each expert evaluates the WPS process with his/her own subjectivity. As a consequence, we propose a quantitative evaluation that introduces different criteria of analysis.

The paper is structured as follows. Section 2 provides a brief overview of WPS and its evaluation policies. Section 3 introduces the three clustering algorithms. Sections 4 and 5 present the manual and automatic evaluations. Finally, 6 concludes the paper with a discussion and outlines future works.

2 Related Work

Web Page Segmentation. Efforts on WPS have focused on removing noisy content from web pages (Yi et al., 2003; Chen et al., 2003; Alassi and Alhajj, 2013; Barua et al., 2014). Later, (Yin and Lee, 2005) were the first to propose a structural viewpoint of web page segmentation, by developing a graph-based strategy to classify elements into categories. For that purpose, layout and Document Object Model (DOM) features were used, as well as some hand-crafted heuristics. Although this methodology shows an original research direction, it relies on a fixed structural semantics that does not correspond to the creativity on the Web. More recently, (Sanoja and Gançarski, 2014) proposed Block-O-Matic, a pipeline strategy, which combines content, geometric and logical structures. One of the main drawback of this approach is the fact that it heavily relies on the DOM, which can be prone to errors due to uncontrolled page creation (Zeleny et al., 2017). Moreover, the number of clusters is automatically determined and thus can greatly vary from page to page. Also, some elements can remain unclustered.

In order to overcome some of these limitations, visual-based strategies have been proposed, which mainly focus on the analysis of the visual features of the document contents as they are perceived by human readers. Notable works that follow this paradigm are VIPS (Cai et al., 2003a) and the Box Clustering Segmentation (BCS) algorithm (Zeleny et al., 2017). While VIPS still uses the DOM as a logical view of the document in combination with visual features, BCS exclusively relies on a flat visual representation of the document, that allows great adaptability to new web contents. In particular, BCS follows a sort of hierarchical agglomerative clustering algorithm that includes a threshold, which controls the gathering of visual elements into clusters. As a consequence, the number of coherent zones is automatically determined by the threshold and can vary, and some elements may remain unclustered, similarly to (Sanoja and Gançarski, 2014).

In this paper, we follow the same strategy as the BCS algorithm as we exclusively rely on visual elements to segment web pages, and thus rely on a flat structure. But, we propose three different clustering techniques that comply to the constraints imposed by the non visual skimming task: (1) segmentation into exactly 5 coherent zones, (2) completeness, where all visual elements belong to a given cluster and (3) connectivity of all the elements inside a cluster.

Evaluation. With respect to evaluation of WPS, two strategies have been predominantly proposed. On the one hand, qualitative evaluations can be performed, where human assessors are asked to validate the proposed segmentation against a human ground truth (Cai et al., 2003b).

On the other hand, studies propose quantitative evaluations relying on cluster correlation metrics.

Within this context, (Zeleny et al., 2017) compare BCS to VIPS using classical clustering evaluation metrics, the F-score and the Adjusted Rand Index. In particular, they create pairs of automatically detected areas and manually annotated areas, which share at least one rendered box. For each such pair, they calculate Precision and Recall. If there are any manually selected areas that do not share boxes with any automatically detected areas, the recall value for each of them is set to 0. The resulting F-score is calculated using average values of Precision and Recall for the entire web page. So, (Zeleny et al., 2017) use the techniques of a general clustering problem. However, WPS can not strictly be compared to a general clustering problem. For example, if just one visual element does not belong to its correct cluster, it may break the logical structure of the segmentation, but the quantitative metric will still remain high. Similarly, (Sanoja and Gançarski, 2015) create a ground truth database by segmenting web pages using the MoB tool Then, a block in the automatic segmentation is said to be correctly segmented if its geometry and location are equal to only one block in the ground truth database; thus proposing specifically-tuned metrics. But as they mostly rely on the DOM structure, they are limited to DOM-based methodologies.

3 Clustering Strategies

WPS for the specific task of non visual skimming can be defined as a clustering problem, where basic visual elements must be gathered into a K fixed number of clusters, where K is equal to 5. In particular, basic visual elements are retrieved from a web page after rendering on the user's browser. DOM elements are then enriched with calculated CSS features, and the basic visual elements correspond to the last block elements in each branch of the DOM tree[2]. In order to cluster the basic visual elements, we propose three different strategies: K-means, F-K-means, and Guided Expansion.

3.1 K-means

K-means (MacQueen, 1967) is a well-established algorithm, when the number of clusters must be fixed in advance. Within the context of WPS, some adaptations are required. In particular, the assignment phase is based on the shortest euclidean distance between two visual elements, noted $dist(.,.)$. For our task, the elements to cluster are not data points in an N-dimensional space, but blocks, i.e. rectangle shapes. In particular, we use a border-to border distance instead of a center-to-center distance. Indeed, as shown in Figure 1, a border-to-border distance is more appropriate in a visual context than a center-to-center distance. In our example, the center-to-center strategy selects the two visual bounding boxes positioned on the right while border-to-border strategy selects the two positioned on the left.

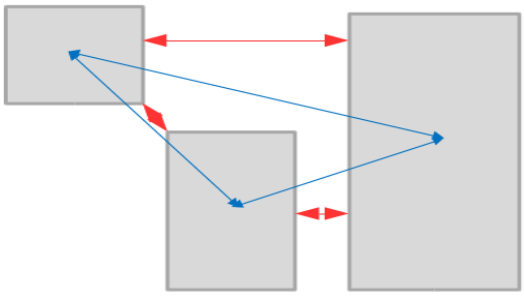


Figure 1: Blue lines showing center to center and red lines showing border to border distances.

Moreover, the classical K-means relies on the random selection of initial seeds. However, this strategy does not adapt to our approach because we need comparable algorithms. As a consequence, we propose to fix the 5 initial seeds following the diagonal reading strategy, i.e. if a diagonal is drawn on the web page from top-left to bottom-right, two seeds are positioned on each extremities, one in the center and the two other ones between the extremities and the center of the diagonal. The underlying idea is that within a skimming process, readers adopt a fast reading strategy, that focuses on particular areas of the web page. In this paper, we propose to test the diagonal strategy but other reading approaches exist (Fitzsimmons et al., 2014; Pernice et al., 2014), which study remains as future work. The K-means clustering process is detailed in Algorithm 1. An illustration of the K-means on a real web page is given in Figure 2.

[2]This is our unique use of the DOM structure.

Input: The set of basic visual elements; K
Output: K clusters
Initialization: Select K centroid elements;
while *true* **do**
  Assign each visual element to its closest centroid based on $dist(.,.)$;
  Compute K new centroids as the average virtual visual element of each cluster;
  if *centroids do not change* **then**
    break;
  end
end

Algorithm 1: K-means algorithm.

3.2 F-K-means

In the first proposal, the assignment phase is exclusively based on the geometric distance between visual objects. For this second algorithm, we propose a small variant, which takes into account the area covered by each visual basic element, the rationale being that visually bigger elements are more likely to "absorb" smaller elements than the contrary. So, if two visual elements are close to each other, their assignment function $force(b_1, b_2)$ will also depend on their differences of covered area as defined in equation 1, where a_{b1} (resp. a_{b2}) is the area of the visual element b_1 (resp. b_2) and $dist(.,.)$ is the shortest border-to-border euclidean distance between the basic elements.

$$force(b_1, b_2) = \frac{(a_{b1} * a_{b2})}{dist(b_1, b_2)} \quad (1)$$

So, the F-K-means algorithm follows the exact same procedures as algorithm 1, to the exception of the function used for the assignment step, which is the $force(.,.)$, i.e. the elements, which show the highest force to their centroids are selected. An illustration of the F-K-means on a real web page is given in Figure 3.

3.3 Guided Expansion

With the Guided Expansion (GE) algorithm, instead of assigning all visual elements to their closest centroid in a single step, only one visual element is assigned at a time to its centroid, controlled by a set of conditions that include the shortest euclidean distance between the borders of two elements, the alignment between elements, and their visual similarity. The GE is defined in algorithm 2.

In particular, visual similarity $vsim(.,.)$ between two elements b_1 and b_2 is computed as in equation 2 over their respective feature vectors $\overrightarrow{b_1}$ and $\overrightarrow{b_2}$ formed by the following CSS properties of each bounding box: font-color, font-weight, font-family and background-color.

$$vsim(\overrightarrow{b_1}, \overrightarrow{b_2}) = \sum_{i=1}^{|\overrightarrow{b_1}|} \mathbb{1}_{\overrightarrow{b_1^i} = \overrightarrow{b_2^i}} \quad (2)$$

It is important to notice that a cluster is a set of visual elements, except for the first step of the algorithm. So, when the distance and the visual similarity are computed between an element and its cluster candidate, this refers to the computation of each metric between the element and all the elements in the cluster. This situation is formalized in equations 3 and 4, where c_1 is the cluster candidate for b_1. An illustration of the GE algorithm on a real web page is given in Figure 4.

$$dist(b_1, c_1) = argmin_{b_i \in c_1} dist(b_1, b_i) \quad (3)$$

$$vsim(\overrightarrow{b_1}, c_1) = argmax_{b_i \in c_1} vsim(\overrightarrow{b_1}, \overrightarrow{b_i}) \quad (4)$$

4 Qualitative Evaluation

In this section, we propose to perform a qualitative evaluation, where 3 human experts are asked to evaluate two common indices in clustering, i.e. compactness and separateness (Acharya et al., 2014). Each expert must produce his/her own segmentation and evaluate both indicators on his ground truth. Compactness is defined at the cluster level and evaluates how many of the elements within a cluster belong to a same cluster in the (individual) ground truth. Separateness is defined at the web page level and evaluates how much the proposed segmentation guarantees the separability between clusters when compared to the expert ground truth segmentation. In this case, the expert must evaluate how much, on average, elements that should belong to the same cluster following the (individual) ground truth are separated in different clusters.

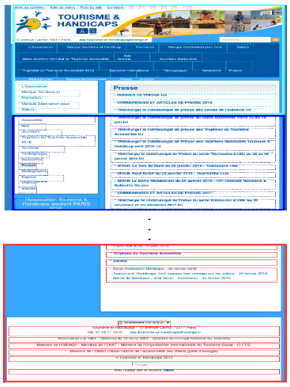


Figure 2: K-means

Figure 3: F-K-means

Figure 4: Guided Expan.

Figure 5: Manual Segm.

In particular, each expert must give a mark ranging from 0 (unacceptable), 1 (bad), 2 (passable), 3 (good) and 4 (perfect). Based on this protocol, the three algorithms presented in section 3 have been tested on a total of 53 web pages from 3 domains: Tourism (23 web pages), E-Commerce (12 web pages) and News (18 web pages), that are part of our TAG THUNDER project corpus[3]. To avoid bias, the experts are unaware of the algorithms strategies he/she is evaluating. Overall results are presented in table 1 and an example of the expert manual segmentation is illustrated in Figure 5.

It is clear that the GE algorithm shows the best figures both in terms of compactness and separateness for the 3 human experts. However, while compactness receives average values between passable and good, separateness receives much lower values, between passable and bad. This finding is transverse to all three algorithms, clearly evidencing that finding coherent zones that match human expectations is a hard task, while building internally semantically coherent zones is easier. Also, figures show differences between K-means and F-K-means. In particular, both algorithms show similar compactness, but the F-K-means evidences worst results for separateness. This result can easily be explained as the F-K-means tends to create unbalanced clusters, that are either very small or rather big. This is confirmed by the higher standard deviation in terms of compactness for F-K-means than for K-means, signifying that F-K-means tends to create very compact clusters (but small) and uncondensed big ones, thus penalizing separateness.

To statistically confirm these results, we computed a global segmentation score (GSS) taking into account both compactness and separateness (equation 5) and performed a Wilcoxon signed-rank test between all algorithms for each human expert. In equation 5, the evaluation scale refers to the scoring scale of separateness ($separat$) and compactness ($compact$), i.e. in our case 5 (0 to 4 grade). Results in table 2 show that GE evidences statistically superior results to both K-means and F-K-means, and that K-means provides statistically higher results than F-K-means, for all three experts in all tested situations, to exception for Expert 3 when comparing K-means and F-K-means.

$$GSS = \frac{(1 + separat) \times (1 + \overline{compact})}{|\text{evaluation scale}|^2} \quad (5)$$

5 Quantitative Evaluation

As seen from the manual evaluation, each expert evaluates the segmentation in a different way depending on his/her perception of coherency of the visual elements. In order to reduce human bias in evaluation, quantitative metrics should be used. However, as stated in section 2, clustering metrics are not adapted to our task. As a consequence, we propose to compute a set of metrics that characterize clustering results based on three different criteria: (1) number of broken logical constraints, (2) cluster balance and (3) cluster geometrical overlap.

[3]This dataset is freely available for research purposes.

Input: The set of basic visual elements; K
Output: K clusters
Initialization: Select K centroid elements (clusters) based on the reading strategy;
while *there are unclustered elements* **do**
  Select each closest element to every cluster using $dist(.,.)$;
  Order these elements by the minimum distance to their candidate cluster;
  Remove all elements that do not evidence the smallest distance for possible assignment;
  if *there are no ties* **then**
    Assign the closest element overall to its cluster;
  end
  else if *there are ties* **then**
    Check whether the elements are vertically or horizontally aligned with at least one element of their cluster;
    Order elements by alignment;
    if *there are no ties AND one aligned element* **then**
      Assign the aligned element to its cluster;
    end
    else if *there are ties OR no aligned element* **then**
      Order elements by the maximum visual similarity to their cluster;
      Remove all elements that do not evidence the highest visual similarity for possible assignment;
      if *there are no ties* **then**
        Assign the most visually similar element to its cluster;
      end
      else if *there are ties* **then**
        Assign all elements to their cluster;
      end
    end
  end
end

Algorithm 2: Guided expansion algorithm.

		Compactness		Separateness		GSS	
		Avg.	$\pm\sigma$	Avg.	$\pm\sigma$	Avg.	$\pm\sigma$
K-M	E1	2.42	1.16	1.15	0.64	0.30	0.12
	E2	1.90	0.87	1.20	0.60	0.26	0.11
	E3	3.10	0.74	0.70	0.80	0.29	0.15
F-K-M	E1	2.43	1.46	0.62	0.57	0.23	0.09
	E2	1.83	1.15	0.40	0.50	0.16	0.07
	E3	3.05	1.22	0.30	0.50	0.21	0.095
GE	E1	**2.89**	1.24	**1.62**	0.93	**0.42**	0.19
	E2	**2.41**	0.81	**1.90**	0.90	**0.41**	0.16
	E3	**3.40**	0.68	**1.50**	0.90	**0.44**	0.18

Table 1: Overall results for K-means (K-ME.), F-K-means (F-K-ME.) and Guided expansion (GE).

H_1	F-K-ME. $<$ K-ME.		F-K-ME. $<$ GE		K-ME. $<$ GE	
	z-score	S/NS	z-score	S/NS	z-score	S/NS
E_1	4.365	S	5.392	S	3.726	S
E_2	5.291	S	4.997	S	3.548	S
E_3	2.169	NS	4.304	S	3.021	S

Table 2: Wilcoxon signed-rank test for the GSS. S stands for significant statistical difference and NS for non significant. E_i is i-th expert. Tests are computed for $p < 0.05$.

Three criteria emerged from the manual evaluations conducted by all three experts. First, experts evaluated negatively when logical constraints were broken, i.e elements embodied by specific HTML tag sequences such as <li> items, <title> and the following paragraph <p>, <header>, <footer> or <nav> elements. So, each time one of these logical constraints is broken, this counts for one cut, and each web page is evaluated based on its overall number of cuts. The higher the number of cuts, the worst the clustering result must be evaluated. Overall results are given in table 3 (column 1). results show the superiority of the Guided Expansion algorithm over the other two algorithms in terms of number of cuts. In particular, it evidences a minimum average value of 1.47, while K-means shows a 2.12 score and F-K-means shows worst results with a score of 2.80. Moreover, the three algorithms can be sorted according to their ability to minimize the cut criterion with statistically significant values, i.e. GE is superior to K-means, which is in turn superior to F-K-means. This criterion seems all the more important that there seems to be a correlation between manual and automatic results. Indeed, as illustrated by figure 4 for GE and figure 5 for manual segmentation, similar behavior seems to stand. However, this situation does not stand for the other two algorithms, where for instance menu sec-

tions are cut as illustrated in figures 2 and 3.

Second, experts negatively evaluated strong imbalance between clusters, but also high balance between clusters. This can be motivated by the fact that a great deal of web pages contain a main (rather large) body section, while all other zones show similar sizes. Note that this issue is usually not taken into account by classical clustering metrics such as Adjusted Rand Index or F-score. As a consequence, this notion of balance is tested over three different properties of the clusters: surface area of the cluster, number of characters within the cluster, and number of visual elements within the cluster. In particular, the surface area of the cluster is calculated as the maximum rectangle that embodies all the visual elements contained in it. So, each web page receives an overall score that stands for the standard deviation between all clusters for each of the three balance criteria (i.e. surface, text and visual elements). Overall results are given in table 3 (columns 2-4).

Third, experts evaluated negatively when the zones were intertwined with each other, i.e. they penalized non rectangular clusters. To evaluate this phenomenon, we computed the number of overlaps between the outer rectangles of all clusters, i.e. the smallest rectangle including all the elements of each cluster. So, if two clusters overlap in terms of outer rectangle, this stands for the presence of a non rectangular zone. Overall results are given in table 3 (column 5).

Table 3 shows the results of the automatic evaluation for the three main criteria for a set of 150 web pages (47 tourist web pages, 58 e-Commerce web pages and 45 news web pages[4]) segmented using the three algorithms (K-means, F-K-means and Guided Expansion). In particular, each criterion receives the average value and the standard deviation $\pm\sigma$ for the set of 150 pages. Table 4 completes results of table 3 with statistical significance by including the Wilcoxon signed-rank test.

First, results show the superiority of the Guided Expansion algorithm over the other two algorithms in terms of number of cuts. In particular, it evidences a minimum average value of 1.47, while K-means shows a 2.12 score and F-K-means shows worst results with a score of 2.80. Moreover, the three algorithms can be sorted according to their ability to minimize the cut criterion with statistically significant values, i.e. GE is superior to K-means, which is in turn superior to F-K-means. This criterion seems all the more important that there seems to be a correlation between manual and automatic results. Indeed, as illustrated by figure 4 for GE and figure 5 for manual segmentation, similar behavior seems to stand. However, this situation does not stand for the other two algorithms, where for instance menu sections are cut as illustrated in figures 2 and 3.

Second, balance results show similar observations whether we compare surface area, text area or number of elements between clusters. In all cases, the F-K-means shows highest unbalance[5], while K-means shows the lowest unbalance. This situation can be observed in figures 2 and 3, where respectively, K-means tends to create evenly distributed zones and F-K-means usually discovers a large zone and a set of smaller clusters. Oppositely, the Guided Expansion algorithm evidences some tendency to unbalanced clustering, that seems to better approximate human segmentation as shown in figures 4 and 5, where human annotators may allow a disequilibrium between the main body of the web page and the satellite zones such as headers, footers or menus. Indeed, humans tend to prefer little unbalanced zones in order to both respect the task condition (i.e. non visual skimming) and maintain the structural and logical aspects of the web page. Note that with respect to statistical significance, we can conclude that F-K-means is clearly the algorithm that steadily produces more unbalanced results. while this hypothesis is not so strong between K-means and the GE algorithm.

Third, the "Exterior Rectangle" criterion, that aims to measure the number of non-rectangular shapes evidences similar results between all algorithms with around five overlaps per web page on average. Nevertheless, there is a clear statistical tendency of the F-K-means to produce less non-rectangular zones. This can be explained by the unbalance constraint. Indeed, as the F-K-means produces highly unbalanced clusters, i.e. usually a large big zone and a set of rather small clusters, it is unlikely that overlap between zones exist, and

[4] All part of our project corpus.

[5] This situation has already been evidenced in the qualitative evaluation.

	Nb. of Cuts Avg. $\pm\sigma$	Surface Area Avg. $\pm\sigma$	Text Area Avg. $\pm\sigma$	Nb. of Elements Avg. $\pm\sigma$	Exterior Rectangle Avg. $\pm\sigma$
K-means	2.12 $\pm$2.05	11.80 $\pm$6.46	11.40 $\pm$5.52	10.95 $\pm$8.01	5.21 $\pm$2.54
F-K-means	2.80 $\pm$2.76	21.14 $\pm$8.18	18.55 $\pm$7.74	22.79 $\pm$16.73	4.54 $\pm$2.20
GE	1.47 $\pm$1.85	17.34 $\pm$6.95	16.78 $\pm$6.37	19.67 $\pm$13.47	5.39 $\pm$2.22

Table 3: Automatic evaluation results for K-means, F-K-means and Guided Expansion (GE) for 150 web pages. Note that the column $\pm\sigma$ gives the standard deviation value over the 150 web pages.

H_1	F-Kmeans $>$ Kmeans		F-Kmeans $>$ GE		Kmeans $>$ GE	
	z score.	S/NS	z score	S/NS	z score	S/NS
Nb. of Cuts	3.64	S	7.08	S	4.85	S
Surface Area	10.29	S	5.65	S	9.12	NS
Text Area	9.59	S	2.36	S	8.83	NS
Nb. of Elements	9.96	S	2.53	S	9.54	NS
Exterior Rectangle	3.35	NS	3.60	NS	1.11	S

Table 4: Wilcoxon signed-rank test for the automatic evaluation for K-means, F-K-means and GE for 150 web pages. S stands for significant statistical difference and NS for non significant. Tests are computed for $p < 0.05$.

as a side-effect less non-rectangular zones are created. However, it is important to notice that the exterior rectangle criterion goes down to almost 0 for human annotators, who rarely proposed non-rectangular zones. As such, one might think that all algorithms are far from achieving human-like behavior. Although this is a strict reality from the figures, this difference against the manual evaluation observation may also indicate a lack of possible solutions by human annotators. Indeed, we think that acceptable segmentation can be proposed by some algorithms, although human annotators may not have thought about. For example, the top of figure 4 shows a non-rectangular red zone with an outer rectangle overlapping the yellow one, that might satisfy some logical coherence, as menus are gathered together. Although this situation has not been proposed by any of the three annotators, we agree that such a segmentation is clearly satisfactory. Based on a deeper manual analysis of these results, we found that the Guided Expansion algorithm seems to be best performing algorithm on this criterion by producing better non-rectangular zones. Nevertheless, further discussion should clearly be about the way to refine this criterion in order to distinguish between good and bad overlaps automatically.

6 Conclusions and Research Directions

In this paper, we presented Web Page Segmentation as a clustering problem driven by the task of non visual skimming. In particular, we tuned the well-known K-means algorithm and designed two other algorithms, namely the F-K-means and the Guided Expansion, all dedicated to our objective and respecting the task constraints of a fixed number of zones, completeness of the coverage, and connectivity of visual elements. In particular, we showed that human and automatic evaluations are complementary to rank the algorithms according to several parameters (the number of cuts of HTML elements, the number of overlaps between zones and the balance of created clusters), each parameter performing a specific complementary role for both compactness and separateness criteria. From both qualitative and quantitative evaluations, the Guided Extension algorithm seems to be the most efficient solution over all criteria. The superiority of the GE algorithm is probably due to the introduction of the alignment constraint. Indeed, the alignment constraint is more difficult to encode in a K-means family algorithm as alignment is a local feature. Still, some clear limitations exist. The clustering process is highly sensitive to the initial seeds positions. By following a diagonal reading strategy, we noted that most algorithms evidence an horizontal segmentation, i.e. vertical cluster are difficult to identify. Another related issue concerns the F-K-means. If some seed is associated to a small element, this cluster will hardly expand as the $force(.,.)$ metric tends to benefit larger visual elements, thus clearly disadvantaging this algorithm compared to the other ones. As such, immediate future work must deal with finding optimal reading strategies for all algorithms.

References

Sudipta Acharya, Sriparna Saha, José G. Moreno, and Gaël Dias. 2014. Multi-objective search results clustering. In *25th International Conference on Computational Linguistics (COLING)*, pages 99–108.

Derar Alassi and Reda Alhajj. 2013. Effectiveness of template detection on noise reduction and websites summarization. *Information Sciences*, 219:41–72.

Jayendra Barua, Dhaval Patel, and Ankur Kumar Agrawal. 2014. Removing noise content from online news articles. In *20th International Conference on Management of Data (SIGMOD)*, pages 113–116.

Deng Cai, Shipeng Yu, Ji-Rong Wen, and Wei-Ying Ma. 2003a. Extracting content structure for web pages based on visual representation. In *5th Asia-Pacific Web Conference on Web Technologies and Applications (ApWeb)*, pages 406–417.

Deng Cai, Shipeng Yu, Ji-Rong Wen, and Wei-Ying Ma. 2003b. Vips: a vision-based page segmentation algorithm. Technical Report MSR-TR-2003-79, Microsoft, November.

Yu Chen, Wei-Ying Ma, and Hong-Jiang Zhang. 2003. Detecting web page structure for adaptive viewing on small form factor devices. In *12th International Conference on World Wide Web (WWW)*, pages 225–233.

Gemma Fitzsimmons, Mark J. Weal, and Denis Drieghe. 2014. Skim reading: an adaptive strategy for reading on the web. In *ACM Web Science Conference (WebSci)*, pages 211–219.

João Guerreiro and Daniel Gonçalves. 2015. Faster text-to-speeches: Enhancing blind people's information scanning with faster concurrent speech. In *17th International ACM SIGACCESS Conference on Computers & Accessibility (ASSETS)*, pages 3–11.

James MacQueen. 1967. Some methods for classification and analysis of multivariate observations. In *15th Berkeley Symposium on Mathematical Statistics and Probability (BSMSP)*, volume 1, pages 281–297.

Elena Manishina, Jean-Marc Lecarpentier, Fabrice Maurel, Stéphane Ferrari, and Busson Maxence. 2016. Tag Thunder : Towards Non-Visual Web Page Skimming. In *18th International ACM SIGACCESS Conference on Computers and Accessibility (ASSETS)*.

K. Pernice, K. Whitenton, and J. Nielsen. 2014. *How People Read on the Web: The Eyetracking Evidence*. Nielsen Norman Group.

Andrés Sanoja and Stéphane Gançarski. 2014. Block-o-matic: A web page segmentation framework. In *International Conference on Multimedia Computing and Systems (ICMCS)*, pages 595–600.

Andrés Sanoja and Stéphane Gançarski. 2015. Web page segmentation evaluation. In *30th Annual ACM Symposium on Applied Computing (SAC)*, pages 753–760.

Lan Yi, Bing Liu, and Xiaoli Li. 2003. Eliminating noisy information in web pages for data mining. In *9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 296–305.

Xinyi Yin and Wee Sun Lee. 2005. Understanding the function of web elements for mobile content delivery using random walk models. In *14th International Conference on World Wide Web (WWW)*, pages 1150–1151.

Jan Zeleny, Radek Burget, and Jaroslav Zendulka. 2017. Box clustering segmentation: A new method for vision-based web page preprocessing. *Information Processing & Management*, 53(3):735–750.

Automatic Speech Act Classification of Korean Dialogue based on the Hierarchical Structure of Speech Act Categories

Youngeun Koo
Dept. of German Linguistics & Literature, Sungkyunkwan University /
25-2, Sungkyunkwan-Ro, Jongno-Gu, Seoul, Korea
sarah8835@skku.edu

Jiyoun Kim
Dept. of German Linguistics & Literature, Sungkyunkwan University /
25-2, Sungkyunkwan-Ro, Jongno-Gu, Seoul, Korea
kite92@skku.edu

Munpyo Hong*
Dept. of German Linguistics & Literature, Sungkyunkwan University /
25-2, Sungkyunkwan-Ro, Jongno-Gu, Seoul, Korea
skkhmp@skku.edu

Abstract

Speech act is an utterance intention and should be understood correctly for a successful communication. However, in some cases, analyzing speech act of an utterance is not simple. This is because we can roughly grasp 'representative' speech act relatively easily through 'utterance-internal' features, but 'concrete' speech act varies depending on 'utterance-external' features. Therefore, this paper proposes a hierarchical structure of speech acts and a two-step classification method of speech acts, for a better understanding of the human conversation and the improvement of automatic speech act classification. The experiment, using Korean tutorial dialogues and telephone calls, showed 83% for the 1st step and 84% for the 2nd step, while using a flat structure showed 71% of accuracy.

1 Introduction

Speech act (SA) is an intention of an utterance. Austin (1962) argues that SA is "a functional unit in communication". For successful communication and correctly understanding the intention of an utterance, SA is very important. Understanding correct SA is crucial not only in a real-life, but also in the field of 'Intelligent tutoring systems(ITS)' as well as in various dialogue systems such as Apple's 'Siri' assistant and Amazon's 'Echo' speaker. Along with the increasing demand for such systems which can interact with human naturally, the system is expected to improve its performance through implementing a better SA classification method into it. However, it is often not easy to analyze it clearly. Sometimes, it is hard to describe SA because it varies according to in which situation the utterance is made. Besides, some utterances are difficult to be defined as one specific SA.

In this study, we examine various factors that disrupt SA analysis and claim the necessity of a hierarchical structure of SA categories. We propose the hierarchical structure of SA by comparatively analyzing two different corpora: Korean tutorial dialog and Korean telephone call.

This paper is organized as follows. In section 2 we look through related works of speech act theory. Section 3 shows some difficult cases of speech act classification and proposes the hierarchical structure of speech act categories to solve this problem. Then, in section 4, we present how we set the experiment of an automatic speech act classification to verify our methodology and then discuss the result of the experiment. Finally, we conclude this paper with future works.

* Corresponding author

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 432-441, Hakodate, Japan, September 13-15, 2019

2 Related Works

In pragmatics, the study of the use of language, speech act theory is one of the most important topics. Speech act theory was established by J. Austin and J. Searle. Austin (1962) insisted that a language of itself is an action and introduced the concept of a 'performative' sentence. Then, the works of Searle further developed the speech act theory. According to Austin (1962), depending on the 'force' that affects an utterance, there are three actions: locutionary act, illocutionary act, and perlocutionary act. Normally, we call illocutionary act as speech act.

Discovering and classifying SA categories are of great importance in that it allows us to understand our language use and real-life interaction. Nevertheless, to classify SA is a tough task, because an utterance can have diverse intentions depending on a situation.

Category	Explanation
Representatives	to commit the speaker to something's being the case, to the truth of the expressed proposition
Directives	attempts by the speaker to get the hearer to do something
Commissives	commit the speaker to some future course of action
Expressives	express the psychological state specified in the sincerity condition about a state of affairs specified in the propositional content
Declaratives	brings about the correspondence between the propositional content and reality

Table 1. Speech Act Categories of Searle (1976)

Table 1 is the SA categories of Searle (1976). Later, this had a great influence on further works. Fraser (1974), Kats (1977), Bach and Harnish (1979) and Leech (1983) attempted to improve categories of Searle (1976). In European countries, researchers concentrated on 'sub-classifying' SAs of Searle (1976) (Lee, 2015).[1]

Among SAs of Searle (1976), Hindelang (1978, 1981) concentrated on directives: demanding and questioning directive. Hindelang (1978) subdivided demanding directive into 18 SAs with criteria, such as 'obligation of counterpart to carry out the request' and 'relationship between the speaker and the counterpart'. Similarly, Hindelang (1981) sub-classified questioning directive into 10 SAs. Next, Rolf (1983) subdivided representatives into 36 SAs with two criteria: 'existence of its preceding speech act' and 'speaker's attitude toward the information'. Furthermore, Graffe (1990) sub-classified commissives into 'sp1-preferred type', 'sp2-preferred type' and 'complex type', depending on who has an interest in realizing the commissive. Lastly, Marten-Cleef (1991) dealt with sub-classifying expressives with respect to 'speaker and counterpart's attitude and judgment on the uttered situation'.

These works enabled not only better understanding of SAs of Searle (1976), but also suggesting various SAs that can appear in our real life. Yet, an empirical investigation on whether each SA actually appears in real communication is not completed.

Speech act theory is also studied in the field of computer science. Lampert et al. (2006) and Qadir et al. (2011) adopted 5 SA categories of Searle (1976). Kim (2006) and Buckley et al. (2008) utilized the DAMSL(Dialog Act Markup in Several Layers) (Core and Allen, 1997) tag-set. Some studies like Lee et al. (1997) and Bayat et al. (2016) proposed their own SA categories.

3 Multi-level Speech Act Categories

SA categories proposed in previous researches are mostly in a flat structure. In other words, all categories are on the same level. In fact, some studies show an approach to deviate from a flat structure. Core and Allen (1997) suggested each utterance having multiple SA labels in the DAMSL tag-set. They analyzed SA of an utterance in 3 layers: forward communicative function, backward communicative function and utterance feature.[2]

[1] Kohl and Kranz (1992) explain why sub-classifying 'the global speech act type' of Searle's taxonomy is necessary. First, some speech acts need to be divided in much detail. When we focus on the speech act types of Searle (1976), the types are distinguishable each other. But, in fact, in many cases, the boundary between the types is ambiguous. Second, 'global speech act type' is insufficient to explain the link between the utterance intention and its uttered expression. Third, through sub-classifying 'global speech act type', we can ascertain whether the overall taxonomy is well-founded and plausible.

[2] Some studies describe the top label of DAMSL in 4 dimensions including 'communicative status', 'the

However, they focus on the 'role of an utterance' in a dialogue flow, rather than the 'intention of an utterance'.

Kang et al. (2013) also took an approach to use hierarchical structure in the SA classification. 'Question type', 'response type' and 'other type' are suggested as 3 SA types of the first layer. By structuralizing SAs into the hierarchical structure for SA classification, the accuracy reached 85% in hotel, airline, tour reservation corpus and 91% in schedule management corpus. But the corpora of Kang et al. (2013) are mostly composed of 'question-answer' pair. Thus, it remained yet as a limitation, that the hierarchical structure is highly restricted to the 'question-answer' paired domain corpus.

This paper proposes the multi-level hierarchical structure of SA for a better understanding of human conversation and the improvement of automatic SA classification. In the following section, we look at some difficult cases of SA analysis and discuss the reasons why the hierarchical structure of SA is necessary.

3.1 Importance of the Hierarchical Structure of Speech Act Categories

In some cases, it is difficult to classify an utterance into a specific SA. First, each person can understand SA differently. To be specific, not all people understand the utterance intention 'same', as 'one' specific SA.

(1) A: He is our new teacher!
B: It's not him.
(*'disagree', 'inform'*)[3]

In example (1), speaker B's utterance can be understood as either disagreeing with the speaker A's assertion or informing new information to speaker A. It depends on how people read and perceive this utterance. Even if people perceive the utterance similarly, not all people would denote its SA with the same SA category. People can denote (1B) as 'inform', 'assert', 'disagree', 'dispute', 'react' or 'response'. To solve this, these similar SAs should be grouped into the same type.

Second, we understand SAs differently depending on a communication situation. Examples (2) ~ (5) explain this in more detail.

(2) A: This experiment is due tomorrow!
(*'command', 'request'*)
B1: Yes, Mrs. Jones.
B2: Okay, no need to worry.

In example (2), SA of speaker A's utterance differs depending on the relationship between two speakers. If speaker A is at a higher position or can impose a sanction against, or if speaker B is bound to perform what speaker A demands, the utterance of speaker A is a 'command' (Hindelang, 1978). However, if not, the utterance would be a 'request'.

(3) A: Can you pass me that?
(*'question', 'request'*)

In example (3), SA of speaker A varies according to the situation where the utterance is made. If it happens during a doctor's appointment, it is a 'question'. Instead, if it happens in daily life and speaker A is pointing something close to another speaker, it is likely to be a 'request'.

(4) A: You know I hate messing up the house, don't you?
(*'criticism', 'warning'*)
B1: I'm sorry, mom.
B2: Yes, I'll keep in mind.

In example (4), speaker A's intention differs by to whom the utterance is made. If it is toward a kid who messed up the house, speaker A intends to 'criticize' the kid. However, if speaker A utters toward the other kid who did not mess up the house, speaker A intends to 'warn' this kid.

(5) A: He gave a Christmas present to his boss again this year!
(*'compliment', 'criticism'*)

In example (5), the utterance of the speaker A can be interpreted differently depending on the speaker A's attitude. If the speaker A is favorable to something/someone, which he/she is talking about, his/her utterance is a 'compliment'. Instead,

information level', 'forward-looking function', and 'backward-looking function' (Fisel, 2007). For this paper, whether the number of dimensions in top label is 3 or 4 does not make much difference. Here, we focus on the fact that DAMSL attempted multiple-labeled annotation. Therefore, this paper follows the description in Core and Allen (1997).

[3] Speech acts are marked in italic font.

if the speaker A is hostile, the utterance is a 'criticism'.

Lastly, SA analysis is often difficult since SAs proposed so far inevitably overlap somehow each other. This is because it is almost impossible to establish SA categories fully complimentarily.

For these several reasons, we insist that SAs should be understood and classified automatically based on the hierarchical structure of SAs, rather than on the flat structure.

3.2 Proposing the Hierarchical Structure of Speech Act Categories

The proposed hierarchical structure is designed through empirical analysis of two different domains of conversation corpora: Korean tutorial dialogues and Korean telephone calls, both built by the National Institute of the Korean Language (NIKL). They are comprised of two separate one-to-one conversations. The tutorial dialogue corpus consists of 1,833 utterances between a teacher and a student. Most of the utterances in this tutorial dialogue were collected in a math class. The telephone call corpus consists of 2,005 utterances between a graduate student and an undergraduate student. Table 2 shows further information about corpora.

Speech Act	Utterance		Speech Act	Utterance	
	Tut	Tel		Tut	Tel
Accept	3	1	Exclamation	56	153
Acknowledge	221	254	Greeting	0	4
Agree	45	44	Guess	42	65
Answer	288	185	Induce	110	28
Apologize	2	1	Inform	360	496
Ask-answer	338	269	Praise	1	8
Ask-confirm	44	28	Reject	3	7
Assert	161	293	Request	33	18
Avoid	13	9	Suggest	28	23
Command	14	2	Thank	0	4
Correct	10	9	Will	11	8
Criticism	32	36	Wish	2	11
Disagree	16	49			

Table 2. Number of Speech Acts in each Corpus (Tut: tutorial dialogue, Tel: telephone call)

This paper aims at organizing utterance intentions overall, rather than sub-classifying only one particular SA. Also, instead of designing completely new SA categories, we utilize SA categories proposed in Koo (2018) to build the hierarchical structure of SAs.

At first, we inspected which SAs are a 'representative speech act' that represents SAs with similar features. A representative SA can be understood easily without a complex analysis of the utterance and the conversation. We did not adopt 5 deductively derived SAs of Searle (1976) as our representative SA.[4] Instead, we investigated the corpora to determine the representative SAs.

We analyzed the frequency of each SA to judge whether it is a domain-independent representative SA. This enabled us to presume the status or position of each SA. The one, which is biased in one domain or does not appear often in both domains, cannot be considered as a representative category. In this case, we searched for other similar SAs and put them together in the same upper class, the representative SA.

For example, in the tutorial dialogues, there are many 'avoid' utterances compared to the telephone calls. Considering that 'avoid' is one of the negative reactions to the counterpart's demand, 'avoid' shares many features with 'reject' or 'disagree'. Also, there are many utterances that a teacher 'induces' a student to respond in the tutorial dialogue. Since, the speaker intends to get information through inducing, 'induce' can be combined into an upper SA category with 'ask-answer' and 'ask-confirm'.

On the other hand, 'assert' appears relatively a lot in the telephone call corpus. This is quite easy to infer, because telephone calls mostly occur when one of the speakers has something to the other to talk about. Since 'assert' and 'inform' are similar, in that they both deliver information to others, we can combine them together.

As a result, we propose a two-levelled hierarchical structure of SAs as Figure 1. [5] Representative SAs, the upper level SAs, are marked with all letters capitalized. Concrete SAs, the lower level SAs, are marked with the capitalized first letter.

[4] Pöring and Schmitz (1999) brought out a hierarchical structure of speech act by borrowing 5 speech act types of Searle (1976) and categorizing them into 3 types of speech act in the first class: representatives as 'information-searching', directives and commissives as 'obligative', and expressives and declaratives as 'constitutive'.

[5] Indeed, there might be some speech acts, omitted in this hierarchical structure. However, this is left as a further study. In this paper, we attempt to explore possibility of improved automatic speech act classification by using a hierarchical structure of speech act categories rather than a flat structure.

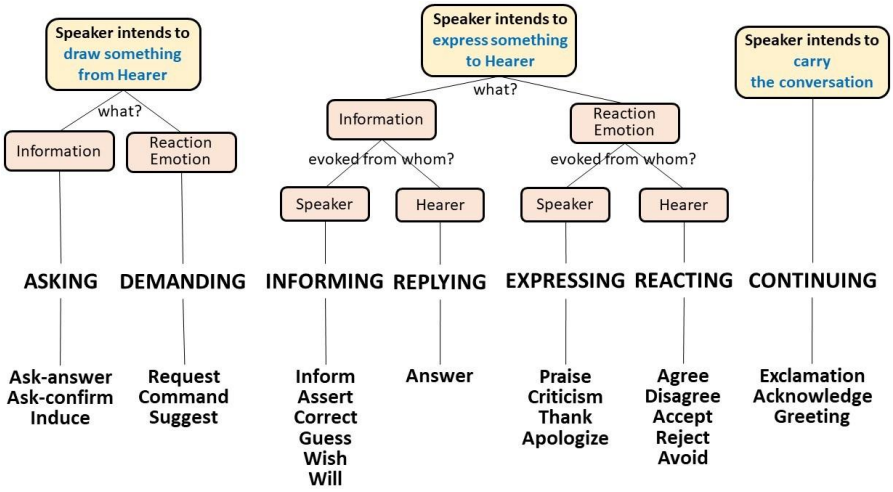


Figure 1. Proposed Hierarchical Tree Structure of SAs

There are a few important issues in the proposed hierarchical structure. This structure was made not only by an inductive approach, by empirically analyzing corpus, but also by a deductive approach, by theoretically analyzing our utterance intentions in terms of linguistic interactions. When making a hierarchical structure, it is important to understand an utterance as a part of an interaction between two or more speakers (Kang, 2004; Franke, 1990). Also, 'what' is involved in the interaction is important. So we considered it as one of the criteria in structuring SAs.[7]

Representative Speech Act	Concrete Speech Act
ASKING	Ask-answer, Ask-confirm, Induce
DEMANDING	Request, Command, Suggest
INFORMING	Inform, Assert, Correct, Guess, Wish, Will
REPLYING	Answer
EXPRESSING	Praise, Criticism, Thank, Apologize
REACTING	Agree, Disagree, Accept, Reject, Avoid
CONTINUING	Exclamation, Acknowledge, Greeting

Table 3. Proposed Hierarchical Structure of SAs

Once representative SA categories are determined, concrete SA categories can be mapped into their corresponding representative SA category. Table 3 shows how each SA of Koo et al. (2018) is linked to 7 representative SAs. Most of these subgroups are self-explanatory, except for the following cases.

First, 'Wish' and 'Will' are a concrete SA in the 'INFORMING' representative SA. Whereas, Rolf (1983) considered them as the expressives. Compared to other SAs, 'Will' and 'Wish' are the multi-faceted category. This causes researchers to analyze them differently. For this study, we focused on their essential intent, conveying some information, rather than their slightly different nuance, 'desiring to achieve' and 'hoping to achieve'.

Second, we mapped 'Answer' to the 'REPLYING' representative SA. In many studies, 'Answer' is regarded as 'INFORMING'. Nevertheless, 'REPLYING' and 'INFORMING' are different, especially with respect to which SA they pairs with. Therefore, we mapped 'Answer' to 'REPLYING', since we propose SA categories not only for understanding the human conversation, but also for automatically classifying SAs.

Third, 'Acknowledge', 'Exclamation' and 'Greeting' fall into the 'CONTINUING' SA. This paper emphasizes the role of a SA in a linguistic interaction. In this respect, we consider 'Acknowledge', 'Exclamation' as a neutral and an emotional reception signal, and 'Greeting' as a socially conventional expression in a conversation.

[7] Bunt (2013) classified 'general-purpose functions' into 'information-transfer functions' and 'action-discussion functions'.

4 Experiments

We conducted an experiment to verify our classification method and the feasibility of the proposed hierarchical structure. We also compared the accuracy of the automatic SA classification between using the flat structure and the hierarchical structure.

We utilized WEKA version 3.8 for machine learning and 'Support Vector Machine (SVM)' as a machine learning algorithm. As an evaluation measure, we employed '10-fold cross validation'. We used two corpora mentioned earlier, Korean tutorial dialogues and Korean telephone calls, as a training corpus. Each corpus consists of 1,833 utterances and 2,005 utterances, respectively.

The experiment used unigram, bigram, which are extracted from each utterance, and linguistic features proposed in Koo et al. (2018) for the automatic classification: 9 sentence features and 4 context features.

Feature Type	Feature Name
Sentence feature	sent_type, tense, sub_person, negation, interrogative, verb_num, sent_length, first two words, last two words
Context feature	prev SA, prev SA_oppo, SA pair, turn chng

Table 4. Linguistic Features for Speech Act Classification

First, we conducted experiments on two corpora together to verify the methodology of this paper. We compared the accuracy of SA classification with the flat structure and with the hierarchical structure. Table 5 shows the result of the experiment.

	Tutorial dialogue + Telephone call	
	1st level	2nd level
Baseline (flat)	71.04	
Proposed (hierarchical)	83.44	84.47

Table 5. Accuracy of the Experiment on Combined Corpus (%)

The accuracy of the '1st level' of the hierarchical structure indicates the accuracy of classifying an utterance into a representative SA. Similarly, the accuracy of the '2nd level' indicates the accuracy of classifying an utterance of a specific representative SA into a concrete SA. For now, we designed two levels of classification separately, aiming for a preliminary examination on our hierarchical structure.

As a result, the accuracy of the 1st level was 83.44% and the accuracy of the 2nd level was 84.47%. Since we classify SAs in two steps, the first step is very important. If the first step is incorrectly analyzed, then the next step is bound to fail.

To evaluate our method, we conducted an additional experiment in a deep learning approach. We used 'Convolutional Neural Networks (CNN)' (Kim, 2014) to classify utterances into 7 representative SAs by setting the model as following: filter windows of 2, 3, 4, batch size of 32, epoch of 50 and learning rate of 0.01. Also, same with the evaluation measure of the earlier experiment using SVM, randomly selected 10% of the training data is used as the test data.

SVM	CNN
83.44	81.75

Table 6. Accuracy of the Experiment on the 1st Level (%)

The accuracy of the CNN model for classifying utterances into the representative SA is 81.75%. Of course, this result will improve, if we elaborate the model. Nonetheless, even with this preliminary experiment, it is still enough to figure out the feasibility of the proposed representative SAs. Moreover, through comparing the accuracy of the feature-based machine learning approach and the deep learning approach, we could conclude that linguistic features of Koo et al. (2018) perform nearly as good as a deep learning model.

In Table 5, the performance of the 2nd level is comparatively lower than the 1st level, though an utterance is classified into 7 classes on the 1st level and 4 classes in average on the 2nd level. This happens to be attributed to the features for the SA classification. Specifically, the features that we used are not suitable or sufficient to classify a concrete SA of an utterance. Following examples are the incorrectly classified utterances.

Num	Utterance	Predicted SA	Actual SA
1 (Tel)	A: 냉면 좀 사줘봐. Buy me some cold noodles. (nayng-myen com sa-cwe-pwa.)	Suggest	Request
2 (Tut)	A: 숙제니까 알아서 적어. It's the homework, so write it down by yourself. (swuk-cey-ni-kka al-a-se cek-e.)	Request	Command
3 (Tel)	A: 니 좀 유종의 미 좀 바꿔요! You should change that 'end well'! (ni com yu-cong-uy mi com pa-kkwe-yo!)	Suggest	Command

Table 7. Incorrectly Classified Utterances of 'DEMANDING' type

In example (1), the Korean expression '~(해)봐(~(hay)pwa)' carries a meaning of suggesting something to someone for his/her sake. However, as this expression is made in the utterance, which is in the conversation between the two close friends and uttered toward the speaker who has more ability to perform something, it is more likely to be a 'Request' rather than a 'Suggest'. Linguistic features derived from the utterance are usually enough to understand the intention of an utterance quite roughly. Nevertheless, they are not enough to understand the intention of an utterance in precise. In other words, 'utterance-internal' features are apt for classifying representative SA, but not for concrete SA. To catch precisely the utterance intention, we need more detailed, 'utterance-external' features.

Similar to example (1), examples in (2) and (3) also require detailed information about the participants of the conversation, in order to be correctly classified. Example (2) is classified automatically as a 'Request'. However, when there is detailed information about two speakers, a teacher and a student, it can give us a clue to classify the SA correctly as a 'Command'.

In addition, considering only the information given from the utterance, the utterance that ends with Korean honorific expression '~요(~yo)', like example (3), could be a 'Suggest'. However, when the information is given that this conversation is between two close friends, the example (3) might be classified correctly.

Num	Utterance	Predicted SA	Actual SA
4 (Tel)	A: 그래서 피곤한가 보지. Guess that's why (he's) tired. (ku-lay-se phi-kon-han-ka po-ci.)	Inform	Guess
5 (Tel)	A: 니는 비디오 언제 산 거야? When did you buy videos? (ni-nun pi-ti-o en-cey san-ke-ya?)	Ask-answer	Ask-answer
	B: 중학교 때는 없었던 거 같은데? I guess it did not exist during middle school. (cwung-hak-kyo ttay-nun eps-ess-ten ke kath-un-tey)	Guess	Inform

Table 8. Incorrectly Classified Utterances of 'INFORMING' type

Example (4) and (5B) show the difficulty of distinguishing 'Inform' from 'Guess'. Whether the speaker has knowledge or information about the subject of the utterance or not, determines the concrete SA. Especially, the 'utterance-external' features could be employed usefully, when the 'utterance-internal' features, such as Korean expression '~거 같다(ke kath-ta)' that indicates an uncertain attitude, in example (5B), might affect a SA classification wrongly.

Next, we conducted an additional experiment on two corpora separately. This experiment is to test whether the performance of a SA classification based on the hierarchical structure can be affected by the domain of the training corpus or not. Table 9 shows the results of the experiment.

	Tutorial dialogue		Telephone call	
	1st level	2nd level	1st level	2nd level
Baseline (flat)	70.03		71.40	
Proposed (hierarchical)	82.11	83.26	86.25	87.18

Table 9. Accuracy of Experiments on each Corpus (%)

As to the accuracy of the baseline, the two corpora show a similar result. However, when we look at the accuracy of the proposed method, the results are different. Telephone call corpus has a higher accuracy on both levels, compared to a tutorial dialogue. This seems due to the complexity of an utterance in each corpus. In the tutorial dialogue, a considerable amount of utterances are not directly related to the conversation between a teacher and a student (Koo, 2018). For example, such utterances are the readings of a textbook by the speaker. Likewise, utterances in the tutorial dialogue are relatively complex, which makes it hard to train input sentences. Whereas, telephone calls are simpler and relatively restricted types of conversation patterns appear in telephone calls.

On top of that, the accuracy of the 2nd level of the tutorial dialogue is low, compared to that of the telephone call. This is presumed to be caused by the lack of data. In fact, for an accurate experiment, fair and balanced amount data for each representative and concrete SA is required. However, as Table 2 shows, the tutorial dialogue particularly lacks data.

5 Conclusion

In this paper, we proposed the hierarchical structure of SA categories by comparatively analyzing the corpus of two different domains: Korean tutorial dialogues and Korean telephone calls. With these corpora, we conducted an experiment of a SA classification using the hierarchical structure. On the 1st level where an utterance is classified into a representative SA, the accuracy of the classification was 83%. On the 2nd level where an utterance of a specific representative SA is mapped into a concrete SA, the accuracy showed 84%. We also discussed the results of the experiment with some examples of incorrectly classified utterances.

Through the experiment, we can infer that the hierarchical structure is adequate for an automatic SA classification. However, a more elaborated analysis of SA categories is necessary. In future works, we plan to verify the SA categories of this study in more detail. The linguistic motivation of the proposed SAs must be investigated further.

In addition, as mentioned earlier in this paper, to classify SA on a concrete level, more features are needed. 'Utterance-external' features like the information about participants of the conversation is necessary for a more sophisticated method for a SA analysis. Not only that, various conversational analysis based features are presumed to be useful.

Above all, for the completeness of this methodology, we plan to connect two steps of the SA classification and identify the performance of the hierarchical structure for automatic SA classification more accurately.

Acknowledgments

This work was partly supported by Institute of Information & communications Technology Planning & Evaluation(IITP) grant funded by the Korea government(MSIT) (R7119-16-1001, Core technology development of the real-time simultaneous speech translation based on knowledge enhancement) and National Research Foundation of Korea (NRF) grant funded by the Korean government(MOE) (2018H-1A2A-1062645, Global Ph.D. Fellowship Program).

References

Andrew Lampert, Dale Robert, and Paris Cécile. 2006 Classifying speech acts using verbal response modes. In Proceedings of the Australasian Language Technology Workshop 2006. 34-41.

Ashequl Qadir, Ellen Riloff. 2011. Classifying sentences as speech acts in message board posts. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing(EMNLP), Association for Computational Linguistics, 748-758.

Berken Bayat, Christopher Krauss, Agathe Merceron, Stefan Arbanowski. 2016. Supervised Speech Act Classifica-tion of Messages in German Online Discussions. In: Proceedings of the International Florida Artificial Intelligence Research Society Conference(FLAIRS Conference), 204-209.

Bruce Fraser. 1974. An examination of the performative analysis. Papers in Linguistics, 7, 1-40.

Changwoo Kang. 2004. Possibility of Subclassification of Speech Act Types and Its Problem. Journal of Germanic Linguistics, 9(0):195-215. (in Korean)

Eckard Rolf. 1983. Sprachliche Informationshandlungen. Göppingen. (in German)

Geoffrey N. Leech. 2016. Principles of pragmatics. New York: Longman Inc.

Götz Hindelang. 1978. Auffordern. Die Untertypen des Aufforderns und ihre sprachlichen Realisierungs-formen. Göppingen. (in German)

Götz Hindelang. 1981. Zur Klassifikation der Frage-handlungen. In: Hindelang, Götz/Zillig, Werner (Hrsg.): Sprache: Verstehen und Handeln. Bd. 2. Tübingen, 215-226. (in German)

Harry Bunt. 2005. A Framework for Dialogue Act Specification. In: Paper presented at the 4th Joint ISO-SIGSEM Workshop on the Representation of Multimodal Semantic Information, Tilburg.

Hyeyong Lee. 2015. Korean Expressive Speech Acts. Youkrack, Seoul. (in Korean)

Hyunjung Lee and Jungyun Seo. 1997. Analysis of Speech-acts for Korean Dialog Sentences. Human and Language Technology, 24(2 Ⅱ):259-262. (in Korean)

Jerrold J. Katz. 1977. Propositional Structure and Illocutionary Force: a study of the contribution of sentence meaning to speech acts. Hassocks: Harvester Press.

John Austin. 1962. How to Do Things with Words. The William James Lectures Delivered at Harvard University in 1955, Clarendon Press.

John Searle. 1976. A classification of illocutionary acts. Language in society, 5(1), 1-23.

Jürgen Graffe. 1990. SICH FESTLEGEN UND VER-PFLICHTEN: Die Untertypen kommissiver Sprechakte und ihre sprachlichen Realisierungs-formen. Münster/New York. (in German)

Kent Bach and Robert M. Harnish. 1979. Linguistic Communication and Speech Acts. Cambridge: The Mit Press.

Mark Buckley, Magdalena Wolska. 2008. A classification of dialogue actions in tutorial dialogue. In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1, 73-80.

Mark Fisel. 2007. Machine learning techniques in dialogue act recognition. Estonian Papers in Applied Linguistics, 3, 117-134.

Mark G. Core and James Allen. 1997. Coding Dialogs with the DAMSL Annotation Scheme, Association for Advancement of Artificial Intelligence(AAAI), 56:28-35.

Mathias Kohl und Bettina Kranz. 1992. Untermuster globaler Typen illokutionärer Akte. Zur Untergliederung von Sprechaktklassen und ihrer Beschreibung. In: König, Peter-Paul und Helmut Wiegers (Hrsg.): Sprechakttheorie. Münster: LIT. S. 1-44. (in German)

Min-Jeong Kim, Kyoung-Soo Han, Jae-Hyun Park, Young-In Song and Hae-Chang Rim. 2006. Dialogue Act Classification for Non-Task-Oriented Korean Dialogues, Human and Cognitive Language Technology, 2006(10):246-253. (in Korean)

Ralf Pörings and Ulrich Schmitz. 1999. Sprache und Sprachwissenschaft: eine kognitiv orientierte Einführung. Gunter Narr Verlag. (in German)

Sangwoo Kang Youngjoong Ko and Jungyun Seo. 2013. Hierarchical Speech-act Classification for Discourse Analysis, Pattern Recognition Letters, 34(11):1119-1124.

Susanne Marten-Cleef. (1991). GEFÜHLE AUS-DRÜCKEN. Die expressiven Sprechakte. Göppingen. (in German)

Walther Kindt. 2009. Pragmatik: die handlungs-theoretische Begründung der Linguistik. In H. M. Müller (Ed.), Arbeitsbuch Linguistik. UTB: Vol. 2169. Paderborn: Schöningh. 289-305. (in German)

Wilhelm Franke. 1990. Elementare Dialogstrukturen. Darstellung, Analyse, Diskussion. Tübingen. (in German)

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), 1746–1751.

Youngeun Koo. 2018. A Study on Speech Act and Automatic Speech Act Classification of German Dialog (Master's Thesis). Sungkyunkwan University, Seoul. (in Korean)

Youngeun Koo, Jiyoun Kim, Munpyo Hong, and Youngkil Kim. 2018. A Linguistic Study of Speech Act and Automatic Speech Act Classification for Korean Tutorial Dialog. Journal of KIISE, 45(8), 807-815. (in Korean)

Investigation of Mandarin Clickbait Headlines: A Case Study of *Biàn Zhèyàng*

Chi-Ling Lee
Master's Program in Teaching Chinese as a Second Language
National Chengchi University
No.64, Sec.2, ZhiNan Road
Taipei City 11605, Taiwan
106161007@nccu.edu.tw

Siaw-Fong Chung
Associate Professor
Department of English
National Chengchi University
No.64, Sec.2, ZhiNan Road
Taipei City 11605, Taiwan
sfchung@nccu.edu.tw

Hui-Wen Liu
Professor
College of Communication
National Chengchi University
No.64, Sec.2, ZhiNan Road
Taipei City 11605, Taiwan
huiwen@nccu.edu.tw

Abstract

With the progress of Internet technology, the ways how news are composed and spread have changed. Nowadays, the click rate of the online news has been used in evaluating the popularity of media platform. Therefore, headlines are now presented with the intent to attract from more 'clicks' from the readers. Given this phenomenon, this research aims to investigate the linguistic techniques employed in Mandarin headlines. The clickbait manifestations of topics, word choices, and presentation patterns will be presented in the result. The results suggested that topics related to appearance and celebrities were often the target of using clickbait headlines. Furthermore, forward-referent serves as one of the techniques to lure the readers to click on the news. Writers' presuppositions were often hidden using positive or negative words in the headlines. Such presuppositions often affect the emotions of the readers upon reading the news. Finally, two- and three- phrase/sentence headlines with various patterns were found. Our study provided a detailed analysis of semantic and pragmatic uses of clickbait headlines, an aspect not often seen in past linguistic analysis of Mandarin news.

Keywords: *clickbait*, news *headlines*, *biàn zhèyàng, forward-referent, presupposition*

1 Introduction

With the widespread of Internet technology, the ways how news were composed and spread have changed. This change is having serious effects on news media business. Chen et al.(2015:1) reported that "the online media economy is largely based around monetization of "views" through advertising revenue" and because of this, "clicks and page views translate directly to a dollar amount". This phenomenon leads to issues such as disguises of news information in headlines by use of non-specific referents, questioning techniques, etc., (Blom and Hansen, 2015; Chen et al, 2015) most of which could be seen as 'clickbait'.

In the Oxford Online Dictionary[1], a 'clickbait' is defined as "Internet content whose main purpose is to encourage users to follow a link to a web page, esp. where that web page is considered to be of low quality or value." With the purpose to attract a larger number of clicks, editors (or sub-editors) and writers employed various techniques in the news headlines to attract a number of readers, sometimes regardless of the decrease of news quality (Chen et al, 2015; Liliana et al, 2015).

In Mandarin online news platforms, it is not hard to find clickbait in news headlines. For example, sentence (1) presupposed that the audience knew about the 325 billion dollars' news, although this could also be a bait because people who did not know it would want to know more about it by clicking it.

(1) 被 川普 3250 億 美元 新聞 嚇 醒
bèi chuānpǔ 3,250 yì měiyuán xīnwén xià xǐng
BEI Trump 325 billionsUSD news scarewake
下 週一 台股 聚焦 **這 件 事**
xià zhōuyī táigǔ jùjiāo zhè jiàn shì
next Monday Taiwan stock focus this CL thing

[1] URL: https://www-oed-com.autorpa.lib.nccu.edu.tw/view/Entry/37263110?redirectedFrom=clickbait#eid (Retrieve date: 2019.05.22)

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 442-451, Hakodate, Japan, September 13-15, 2019

'Due to the shock of the 325 billion dollars news from Trump, this thing should be noticed by Taiwan stock market next Monday'

Second, the sentence did not give any hint about 'this thing'. Readers' curiosity was provoked by the covert of a definite and specific referent, although this referent is non-referring to the readers (Givon, 1978; 1995). This is somehow violating the rules of definiteness and markedness. When we have an object that is definite and specific such as *She wanted to buy this house* versus one that is indefinite and non-specific (*She wanted to buy house*) (cf. Givon, 1995: 61), the former should be a referring-object, while the latter is a non-referring one. The news headline is the first thing that appears in a news article. It is very unlikely that information in it could be referring unless it has been mentioned previously. The use of 'this thing should be noticed' entailed that the writer knew about this thing but did not make it clear to the readers on purpose. It is this manipulation of curiosity that helps increase the number of clicks. Sentence (2) shows a similar kind.

(2) **10種** 食物不 能 放 冰箱
shí zhǒng shíwù bù néng fàng bīngxiāng
10 CL food not can put refrigerator
'Ten kinds of food can't be placed in the fridge'

Sentence (2) provided numeral information -- '10 things' but the number did not provide sufficient information for the readers to know more about the ten things. This kind of technique often appears in book titles. The difference is that this headline is in interrogative form, indicating the writer's unconfirmed source of information. An interrogative headline is also used as a way to avoid commitment so that any reports that turn out to be challenged or incorrect will not be penalized by the news administrative body.

This paper intends to investigate a different kind of clickbait -- the type that appeared fairly recently-- *biàn zhèyàng* 'change to become as such' or 'become like this'. It is the manipulation of one's curiosity by indicating 'a change in something' that deserves their attention. *Biàn* is an 'inchoative' or 'become' verb (cf. Jakendoff, 1990; McCawley, 1971) that mean 'change', whereas *zhèyàng* means 'this or so' or 'as such'. *Zhèyàng* does not clearly describe any detail about the content of news. According to Biq (2007:130), a study of *nà*, the opposite of *zhe4* 'this', *nà* has a 'detached' meaning that shows "约似 (approximation)" and "模糊的认定 (vague identification)". *Zhè*, which should be less detached than *nà* was however used to mean roughly a similar meaning with *na4yang4* in the news headlines, but with more definiteness.

(3) 30年 後 世界 **變 這樣**
30 nián hòu shìjiè biàn zhèyàng
30 year after world become this
'30 years later the world will become like this'

However definite, because *zhèyàng* is vague in content, the news headline still becomes opaque and mysterious. Blom and Hansen (2015) have also discussed the Danish headlines with *sådan* and *så*, both meaning 'like so' or 'this is how'. The paper suggested that these forward-reference words make information gaps between headlines and the origin articles for purpose to trap readers into clicking into the articles.

Since this phenomenon seems to have become a globalized issue, it is essential to look closely at our news headlines. The main purpose of this study is to gain understanding of clickbait headlines in Mandarin and their distributions and types. The following research questions will be answered:

1. What topic of news often employ *biàn zhèyàng* as clickbait to attract readers' attention?
2. What language features are used in Mandarin clickbait headlines?
3. What pragmatic functions are manipulated in *biàn zhèyàng* clickbaits?

Based on the above questions, this study will provide an analysis of *biàn zhèyàng* in Taiwan news headlines.

2 Literature Review

The topic of the news articles is one of the leading factors influencing how many clicks the news will get. Journalism has distinguished different types of hard and soft news. Hard news is considered more society-beneficial, while soft news is often related with tabloidization and 'infotainment' (Reinemann et al. 2012; Lin, 2016; Otto, Glogger, and Boukes, 2017; Skovgarrd,

2014). In recent years, online news has become softer and softer. Unlike the traditional tabloid, which was said to include sex, scandal, tragedy, paranormal, supernatural phenomena, outrageous behavior, how-to-tips on self-improvement, household tasks, information about celebrities, physical deformities or freakish physical accomplishments (Schaffer and Deborah, 1995), Reinemann et al. (2012) discovered that soft news was more episodic and less politically relevant. Moreover, soft news frequently reported personal and private events, although the news topics could be about politics (Jebril, Albaek, and De Vreese, 2013).

For the soft news, the number of clicks was not only affected by the news topics, but also the effect of word choice in the headlines. Word choices of news headlines would catalyze readers' curiosity or suspense. Reinemann et al. (2012) mentioned that commercialized news was now compiled in personal and emotional style. Plenty of personal elements and emotional words can be found in the text to stimulate growth of clicking rate (Bednarek and Caple, 2017; Bednarek, 2006). Yet, scholars such as Bas and Grabe (2015) found that emotion-provoking news may increase clicking rate but did nothing good to close the knowledge gap.

In addition, employing anaphora in headline is another technique to making larger click number. Blom and Hansen (2015) found that Danish headlines frequently used forward-references. Eight manifestations of forward-references in Danish headlines were found, including demonstrative pronouns, personal pronouns, adverbs, definite articles, ellipsis of obligatory arguments, imperatives with implicit deictic reference, interrogatives, and general nouns with deictic reference. These kinds of words were used as triggers of presuppositions about the original articles which made readers fascinated about the content of the news. The research "confirmed that forward-referring headlines are primarily used as clickbait luring the readers into clicking on and reading the full article thus making the news site more attractive for advertisers" (Blom and Hansen, 2015: 98-99).

As mentioned, words used in the headlines also carry certain presuppositions. A presupposition is proposition that the speaker believed but not uttered. Bonyadi and Samual (2013) analyzed headlines in New York Times and Tehran Times and found that existential presupposition and lexical presupposition were constantly found. They suggested that existential presuppositions brought negative attributes in headlines, such as 'stolen election' or 'belated truth', while lexical presupposition revealed strong positive or negative inclination. With the powerful positive or negative connotation words, "the writers tried to presuppose their intended propositions" (Bonyadi and Samual, 2013: 4).

On top of the above, organizing stories in different order of plots also instigate different readers' reactions. Knobloch et al. (2004) testified readers' reflection from various narrative patterns, including linear, reversal, and invert types. The linear pattern represented the stories in chronological order. The reversal patterns presented the stories in reverse order. The invert type of writing was profoundly used and was considered professional in news writing. This type put most important event and outcome in the initial part of the story, remaining less essential elements at the end. The empirical evidences of their research suggested that the linear type of writing evoked more suspense than other two types but higher curiosity in reaction was found in the reversal narratives.

To sum up, with the intent to stimulate click rates, selecting appealing topics, choosing words with presupposition, and using different ways to organize the headlines were some of the common techniques used to increase the click number of news. These methods were seen as measures of clickbait. From analyzing the headline of German news, Kuiken et al. (2017) offered the features of clickbait. Clickbait features included questions, negative words, quotes, signal words, pronouns, and numbers. They testified that using features of the clickbait in headlines significantly increased the number of clicks.

However, the above studies were mainly analyzing western language headlines. In our discussion, we need to take account of the features of Mandarin headlines. Hsieh and Li (2011) suggested that because of the limitation of newspaper space and the immediateness of news, Mandarin headlines usually only contain two short sentences, and 12.7 words on average. The reduction of the sentences and words leads to the simplification of the complete information and the

fragment of the whole meaning. Analyzing the content of two short sentences in headlines on TV, they discovered that these sentences showed five main patterns -- [EVENT + OUTCOME], [EVENT + EVENT], [EVENT + QUESTION], [COMMENT +EVENT], and [EVENT+ REASON]. The expression of [EVENT + OUTCOME] is a cause-effect relationship, which is the core of the news. Therefore, the frequency of this structure is the highest of all the headlines in the research. The [EVENT + EVENT] pattern provides details of the news, whereas [EVENT + QUESTION] shows uncertainty of the event, which also fulfils the neutral and justified role of the media. The [COMMENT + EVENT] pattern adds subjective opinion onto the headlines. The [EVENT+ REASON] pattern shows that the media understand the event well. These patterns of the headlines reveal what kind of perspective the editors would like to give the audience. Hsieh and Li (2011) mainly worked on the headlines on TV. The analysis of the Mandarin clickbait headline patterns will be demonstrated in session 4.3.

In addition, in order to simplify headlines, some methods were often adopted. For example, function words, classifiers, and second verbs were either removed or added on purpose. Using pivotal sentences and ungrammatical sentences was also part of the strategies to make headlines simplified. Furthermore, to increase readers' curiosity, it is also common to use connotation-rich vocabulary, first name of celebrities, well-known nicknames, pseudo-quotes, and literary poetic devices in Mandarin headlines. In what follows, we present the results of our study.

3 Methodology

Headlines containing the keyword *biàn zhèyàng* in Mandarin were collected. Such news concealed the outcome of what have changed. By not revealing detail of the information, readers' curiosity could be stimulated and thus increased the click rate. Through gathering and investigating headlines with *biàn zhèyàng*, this paper attempts to figure out how Mandarin clickbait headlines entice the readers to click on the headlines. Google News was employed to retrieve the headlines due to its tremendous database. The headlines mostly came from main online news platforms in Taiwan; for instance, *China Times*, *Ettoday*, and *United Daily News*. The data was collected in every five days from April to May in 2019. There were 57 tokens news articles altogether. The news types, keywords, and patterns of the collected data were analyzed in order to find out the techniques of Mandarin clickbait headlines.

4 Analysis

This research observed three aspects of headlines, namely the topics of news, keywords in the headlines, and the patterns of headlines.

4.1 Topics of News

All of the news articles were analyzed in terms of their topics. Based on the study Schaffer et al. (1995) had accomplished, the collected data were categorized into eight topic types, namely 'appearance', 'celebrity', 'outrageous', 'new policy', 'scandal', 'sex', 'spatiotemporal', and 'how-to-tips'. The results are given in Table 1.

Topics of News	Freq.	%
Appearance	22	35.48
Celebrity Privacy	21	33.87
New Policy	8	12.91
Outrageous	4	6.45
How-to-Tips	3	4.84
Sex	2	3.23
Spatiotemporal	1	1.61
Others	1	1.61
Total	**62**	**100**

Table 1. Topics of Headlines

A majority of the headlines of *biàn zhèyàng* was the topic of appearance. This topic contained outlook changing of people's faces and people's figures. In addition, there are more than one third of appearance headlines contain well-known names (8 out of 22). With the topic of appearance, the writer drew the viewers' attention by making them curious about how a famous someone's outlook had become, or how had he or she changed. Sentence (4) and (5) are instances of 'Appearance'

(4) 宋 慧喬 化 煙燻 妝 **變 這樣**
sòng huìqiáo huà yānxūn zhuāng ***biàn zhè yang***
Song Hyekyo put smoky makeup become like so
網 嚇 喊 不 適合 她
wǎng xià hǎn bú shìhé tā

Netizens scared call not suit her
'Hyekyo Song put on smoky makeup and became like this; terrified netizens called out and said this did not suit her'

(5)56 歲 關之琳 發福 **變 這樣**
56 suì guān zhīlín fāfú biàn zhèyàng
56 year-old Rosamund Kwan fat become like so
'56 years old Rosamund Kwan became fat like this.'

Like sentence (4) which provided the hint as to how disfavored Song would look, sentence (5) also stated the recent gain of body weight by a well-known beautiful actress. These news reported the negative change of the actresses but sentence (4) had a better connotation than sentence (5), for the headline told specifically that it was the makeup that did not suit the actress (not the actress' problem). It is a constant trick of the media to report weight gain or weight loss of the celebrities, which appeared to be a matter of seriousness to the celebrities.

Topic of celebrity privacy was the second highest among the headlines. These headlines often contained the full name of well-known people and were concerned about changes of their personal relationships, or their unknown privacy. For headlines that pointed to appearance per se were grouped under 'Appearance'. Sentence (6) is an example of celebrity privacy.

(6) 許志安 痛哭 懺悔
xǔzhì'ān tòngkū cànhuǐ
Andy Hui cry-hard repent
鄭 秀文 臉書 封面
zhèng xiùwén liǎngshū fēngmiàn
Sammi Cheng Facebook page
竟 **變 這樣**
jìng ***biàn zhèyàng***
surprisingly become this.
'Andy Hui blubbered and repented; Sammi Cheng Sau Man's Facebook page became like this.'

The boxed lexical items indicate the theme (Facebook background) rather than the reason in which something had changed. It is slightly different from the examples in (4) and (5).The 'Policy' type as in (7) below conveys changes after a new policy had been implemented.

(7) 莫迪 經濟學 實施 五年後
mòdí jīngyìxué shíshī wǔ nián hòu
Modi economics implement five year after
印度 經濟 **變 這樣**
yìndù jīngjì ***biàn zhèyàng***
India economy become so
'After the implementation of Modi's economy policy for five years; the economy of India became like this.'

The boxed lexical items indicate the theme (economy respectively) in which something had changed like the boxed item in sentence (7). Examples of 'Outrageous' headlines contained events that evoked negative emotions, such that in (8) below.

(8) 夏語心 遭鎖門 討 10 倍 車資
Xià Yǔxīn zāo zuǒmén tǎo 10 bèi chēzi
Xia yuxin got locked request 10 times taxi fare
曾 國城 聽完 結局 **變 這樣！**
Zēng Guóchéng tīngwán jiéjú biàn zhèyàng
Zeng listened outcome became like this.
'Xia got locked and asked for ten times taxi fare; after Zeng listened and the outcome became like this.'

This kind of headline first told what the 'unbelievable' act was. The first part often triggered anger or negative emotion from the readers; the second part then lured the readers to click on the news. In many cases, *biàn zhèyàng* was used to introduce how people in the events had become. It is often an exaggeration of the real, often tiny outcome. Bednarek (2008: 67) listed six "core evaluative parameters" for evaluation of the press, namely 'comprehensibility', 'emotivity', 'expectedness', 'importance', 'possibility/ necessity', and 'reliability'. Clickbait news fall short of many of these but may be high in 'emotivity'.

On the other hand, 'How-to-Tips' type of headlines used *biàn zhèyàng* to show the readers what had happened after using the tips that the headline had mentioned (9). Example (10) shows *biàn zhèyàng* was used to mark the unknown future ('Spatiotemporal').

(9) 夫妻 例行 開 週會 生活 **變 這樣**
fūqī lìxíng kāi zhōuhuì shēnghuó ***biàn zhèyàng***
spouse hold week meeting life become like so
'Spouse meet every week; marriage life becomes like this.'

(10) 30 年 後的世界 **變 這樣**：
*30 nián hòu de shìjiè **biàn zhèyàng***
30 year after de world become like so
你的鞋、你的車
nǐ de xié 、nǐ de chē
your shoes your car
你的內衣比你還 聰明
nǐ de nèiyī bǐ nǐ hái cōngmíng
'The world will become like this after 30 years; your shoes, your cars, and your underwear will be smarter than you.'

One 'Others' example was found because the news reported an event that someone did crazy things and was caught by police.

In the next section, the keywords in the headlines will be analyzed.

4.2 Forward-Referent and Presupposition

Forward-referents such as 'like so' or 'this is how' leave information gaps between headlines and the origin articles. They are often employed in headlines to affect the number of clicks. Presuppositions in headline also have an impact on how the audience would interpret the story. This section provides our analysis of keywords.

All the headlines with *biàn zhèyàng* utilized the forward-referent *zhèyàng*. However, most of the referents of *biàn zhèyàng* in the headlines were also forward-reference words such as pronouns, or non-previously introduced nick names, etc. The number of forward-reference words used in as the subjects of *biàn zhèyàng* was calculated. Table 2 shows the counts and percentages of each label.

Forward-Referents	Freq.	%
General Noun (結局 'results', MV)	24	38.7
Full Name (邱品叡)	17	27.42
Modifier+ general noun (正妹老婆 'pretty.girl-wife')	15	24.19
Nick Name (小馮 'little Feng', 阿帕契姐 'Sister AH-64 Apache')	2	3.23
Pronoun (他 'he')	2	3.23
Zero anaphora	2	3.23
Total	**62**	**100**

Table 2. Forward-Referents in Headlines

From Table 2, 38.7% referents of *biàn zhèyàng* were general nouns, and another 24.19% were of 'modifier+general noun'. There was no clue as to know what the specific thing of the 'general noun' was. Therefore, it encourages the readers to read the whole article.

The full names were usually the name of well-known persons. The persons themselves were more attractive than what really happened to them. A total of 27.42% used the person's full names, while only less than 10% used pronouns and zero anaphora. For pronouns and zero anaphora, they were used for a reason. In (11), a pronoun is used to conceal the core information which refers to the main love rival of the female star. This technique might lead to higher click rate. For (12), a zero anaphora was missing because it was referring to oneself (the speaker).

(11) 周 慧敏 頭號 情敵 是 她
zhōu Huìmǐn tóuhào qíngdí shì tā
Vivian Chow main love rival is her.
51 歲 單身 港姐 **變 這樣**
*51 suì dānshēn gǎngjiě **biàn zhèyàng***
51 year-old single Ms. Hong Kong become this.
'Vivian Chow's main love rival is her. 51 year-old Ms. Hong Kong became like this. '

(12) 受過 國民 黨 教育
shòuguò guómín dǎng jiàoyù
received KMT education
館長 嘆：Ø 怎麼 **變 這樣**
*Guǎnzhǎng tàn Ø zěnme **biàn zhèyàng***
Curator sigh : Ø why become this
'Educated by KMT, the curator sighed: why Ø became like this'

As for presupposition-identifying, Mandarin clickbait headlines often used exaggerating, sensational words to attract viewers' attention. Some of these words reveal the author's presupposition. Here is one of the examples to show this phenomenon.

(13) 昔日 龍 女郎 爆 肥 80 公斤
xírì lóng nǚláng bào féi 80 kōngjīn
past dragon lady exploding fat 80 kilogram
靠 這 招 甩 肉 **變 這樣**
*kào zhè zhāo shuǎi ròu **biàn zhèyàng***
depend this method dump meat become like so
'The dragon like lady rapidly gained weight for

80 kilograms; Ø lost weight by this tip and became like this.'

The writer used the adverb *bào féi* 'exploding fatness' to exaggerate how fast the lady had gained weight. In addition, *lóng nǚláng* 'dragon lady' was used to describe the woman's figure in the past, which means the woman used to have a dragon like body. With these exaggerating and sensational words, readers were enticed with curiosity and lured to click on the news even if they did not know who this woman was. A comparison of a better past to a worse present is what the readers would like to read about. In addition, this headline also utilized the forward-referents. Both *lóng nǚláng* 'dragon lady' and *zhè zhāo* 'this method' hid the important elements of the story.

Words with presuppositions increased the possibility of clicking. Strong negative and positive words were used to impose the presuppositions on the audiences, regardless of whether the audience agreed or not. In our collected data, there are plenty verbs which inherited extreme emotions. For instants, '對不起 *duìbùqǐ* (sorry)', '害 *hài* (harm)', '走鐘 *zǒuzhōn* (become deformed)', '腫 *zhǒng* (swell)', '遭 *zāo* (to undergo)', '打趴 *dǎpā* (defeat)', '嘆 *tàn* (sign)', '淚崩 *lèibēng* (uncontrollably cry)', '驚恐 *jīngkǒng* (terrify)', '痛哭 *tòngkū* (bitterly cry)', '懺悔 *chànhuǐ* (repent)', 驚 *jīng* (stun)', '暴怒 *bàonù* (fury)', '爆肥 *bàoféi* (exploding fatness)', etc.

4.3 Patterns of Online News Headline

As mentioned, Mandarin headlines usually include two phrases or sentences (Hsieh and Li, 2011). From the total 62 headlines, 50 headlines (80.6%) were composed with two phrases or sentences and the remaining 12 headlines (19.4%) contained three phrases/sentences. In this part, we followed the analysis structure of Hsieh and Li (2011).

Based on the annotation system of Hsieh and Li (2011), we annotated phrases/sentences in the headlines as EVENT, REASON, OUTCOME, and COMMENT. An EVENT is the main issue that the news conveyed. Our target phrase *biàn zhèyàng* was also marked. A REASON explains why *biàn* (change) had occurred. An OUTCOME marks the part which shows the consequences of the change. A COMMENT represents the judgment or reactions from the author, writers, editors, and others. A COMMENT is usually a response from the netizens.

(14) 掃墓 卻 見 祖墳 **變** **這樣**
sǎomù *què jiàn zǔfén* ***biàn*** ***zhèyàng***,
sweep tomb but see grave change like this
他 po 圖 網 全 怒[COMMENT]
tā *pō* *tú* *wǎng* *quán* *nù*
he postpicture netizens all angry
'He went to clean the tomb but saw the tomb became like this; he posted the picture and the netizens become angry'

The patterns of the two headlines are given in Table 3 below.

Pattern	Freq.	%
EVENT +*biàn zhèyàng*	22	44
Biàn zhèyàng +COMMENT	11	22
REASON + *biàn zhèyàng*	9	18
OUTCOME +*biàn zhèyàng*	2	4
biàn zhèyàng +EVENT	2	4
Biàn zhèyàng +OUTCOME	1	2
REASON and *biàn zhèyàng* + COMMENT	1	2
REASON and *biàn zhèyàng* + OUTCOME	1	2
REASON + *biàn zhèyàng* and COMMENT	1	2
Total	**50**	**100**

Table 3. Patterns of Headlines

(15) 捕獲 許 純美 落跑 情郎 [EVENT]
bǔhuò Xǔ Cúnměi luòpǎo qíngláng
catch Xu Cunmei runaway boyfriend
邱 品叡 過 十年 **變** **這樣**
qiū pǐnruì guò shínián ***biàn zhèyàng***
Qiu Pinrui through ten years become this
'Catchinging runaway Xu's boyfriend, Qiu becomes like this after ten years.

(16) 不 徵 空 店 稅 [REASON]
bù zhēng kōng diàn shuì

Not levy empty store tax
Sway 預言 東區 接 下來 **變 這樣**
Sway *yùyán dōngqū jiē xiàlái* ***biàn zhèyàng***
Sway predict east-area following become this
'Not levying empty-store tax, Sway predicted that the eastern area would become like this.'

Sentence (15) demonstrates a pattern of [EVENT +*biàn zhèyàng*]. The first short sentence presents a scandal which happened about ten years ago. The next *biàn zhèyàng* sentence conveys how the man has become after 10 years. Example (16) is a [REASON + *biàn zhèyàng*] sentence. The first short sentence shows the reason of changing; the second short sentence is the prediction of the changing outcome, but the outcome was concealed by '*biàn zhèyàng*', which enticing the readers to read the full articles.

(17) 土酷 風 打趴 眾 女星 [OUTCOME]
tǔkù fēng dǎpā zhòng nǚxīng
local-cool style defeat all female stsr
最 醜 女團 翻身 **變 這樣**
zuì chou nǚtuán fānshēn ***biàn zhèyàng***
ugliest female team stand-up become this
' Local and cool style defeated all the female stars. The ugliest girl group stands up and becomes like this.'

(18) 女星 生 2 娃 **[**REASON**]**
nǚxīng shēng 2 wá
female star borned 2 babies
變 這樣
biàn zhèyàng
become this
鏡子 反射 身材 太 真實
jìngzi fǎnshè shēnchái tài zhēnshí
Mirror reflect figure too real
網 看 傻！[COMMENT]
wǎng kàn shǎ!
netizens watch stunned
'A star gave birth to two babies and becomes like this. The reflection in the mirror tells the truth; netizen were all shocked.'

The Pattern of sentence (17) is [OUTCOME +*biàn zhèyàng*]. The first short sentence gives the information about the outcome of the changing. The pattern of (18) is [REASON and *biàn zhèyàng* + COMMENT]. The first short sentence contains the reason of changing and *biàn zhèyàng;* the second short sentence is the reflection of the *biàn zhèyàng* from netizens.

Reinemann et al. (2012) stated that soft news is prone to include the journalist's personal impressions, interpretations, and opinions. In this research, the preponderance of the data included comments. These comments provided more emotional and sensational elements. Twenty-two headlines (40%) were found to consist of comments in all the 62 headlines. Among these, five of them had two comments, shown in (19).

(19) 感恩 公投[COMMENT]
gǎn'ēn gōngtóu
appreciate referendum
同志 教育 改良 版
tóngzhì jiàoyù gǎiliángbǎn
homosexual education improved version
變 這樣
Biàn zhèyàng
become so
網： 教育 部長 超 帥[COMMENT]
Wǎng: jiàoyù bùzhǎng chāo shuài
Netizens education minister super cool
'Thanks to the referendum, the improved version of homosexual education policy becomes like this; Netizens: the Minister of Education is cool.'

The pattern of sentence (19) is [COMMENT+ *biàn zhèyàng*+ COMMENT]. The first part expressed the author's personal opinion of the new policy version of the homosexual education. The second sentence used a forward-referent to evoke the viewers' curiosity. The last comment showed the response from the netizens, which presented the netizens' feeling about the minister.

5 Conclusion

Our analysis demonstrated that the clickbait headlines are used more in appearance and celebrity topics. As a kind of Forward-Referent technique, the use of general nouns is the most common way to lure audience into reading the whole articles. In addition, negative and exaggerated words are employed in the headlines to provoke the curiosity of the readers. Finally, writing comments about the news is a kind of method to increase the click rate.

This study analyzed news headlines in Mandarin that contained a clickbait keyword *biàn zhèyàng*. Clickbait news have increased abundantly in

recent years and readers are lured into reading them without notice. A linguistic analysis of how clickbait headlines manipulate readers' response was needed. With the commercialization of media industry, newspapers seek larger and larger click number for survival purposes. Clickbait headlines used various techniques often unknown by readers.

Our research observed that Mandarin headlines often fell into certain news topic, with certain word choice, and employed particular presentation patterns. Using forward-referents and general nouns become common although they normally violate the old-before-new discourse order. Forward-referents reveal limited information about the content. Furthermore, news writers hid their presuppositions behind negative or positive words and by doing so, readers' response could be manipulated. This study provided an analyzing schema as to how one should treat this kind of clickbait news with more caution.

Acknowledgments

This research was supported by MOST project 106-2410-H-004-109-MY2 and NCCU Grant 108H112-01.

References

Bas, O., and Grabe, M. E. 2015. *Emotion-provoking Personalization of News: Informing Citizens and Closing the Knowledge Gap? Communication Research*, 42(2):159-185.

Bednarek, Monika and Helen Caple. 2017. *The Discourse of News Values: How News Organisations Create Newsworthiness.* Oxford/New York: Oxford University Press.

Bednarek, Monika. 2006. *Evaluation in Media Discourse: Analysis of a Newspaper Corpus.* London/New York: Continuum.

Bednarek, Monika. 2008. *Semantic Preference and Semantic Prosody Re-examined. Corpus Linguistics and Linguistic Theory*, 4(2):119-139.

Biq, Yung-O. 2007. *Lexicalization and Pharsalization of na Collocates in Spoken Taiwan Mandarin. Contemporary Linguistics*, 9(2):128-138.

Blom, J. N., and Hansen, K. R. 2015. *Click Bait: Forward- reference as Lure in Online News Headlines. Journal of Pragmatics, 76*:87-100.

Bonyadi, A., and Samuel, M. 2013. *Headlines in Newspaper Editorials: A Contrastive Study. SAGE Open, 3*(2):1-10.

Chen, Y., Conroy, N., and Rubin, V. 2015. *Misleading Online Content: Recognizing Clickbait as 'False News'. Proceedings of the 2015 ACM on Workshop on Multimodal Deception Detection:* 15-19.

Chen, Y., Conroy, N., and Rubin, V. 2015. *News in an Online World: The Need for an 'Automatic Crap Detector'. Proceedings of the Association for Information Science and Technology, 52*(1): 1-4.

Givón, Talmy. 1978. *Definiteness and Referentiality. Universals of Human Language, ed. by Joseph Greenberg*, Cambridge, MA: MIT Press.

Givón, Talmy. 1995. *Functionalism and Grammar. Amsterdam: John Benjamins.*

Hsieh，Chia-Ling and Li， Jia-Hao 2011. *A Study on Taiwanese Television News Headlines and their Pedagogical Implications. Journal of Chinese Language Teaching,8(3)*:79-144.

Jackendoff, Ray S.1990. *Semantic Structures*. Cambridge: MIT Press.

Jebril, N., Albaek, E., and De Vreese, C. H. 2013. *Infortainment, Cynicism and Democracy: The Effects of Privatization vs Personalization in the News. European Journal of Communication, 28*(2):105-121.

Knobloch, S., Patzig, G., Mende, A., and Hastall, M. 2004. *Affective News: Effects of Discourse Structure in Narratives on Suspense, Curiosity, and Enjoyment While Reading News and Novels. Communication Research,* 31(3): 259-287.

Kuiken, J., Schuth, A., Spitters, M., and Marx, M. 2017. *Effective Headlines of Newspaper Articles in a Digital Environment. Digital Journalism*, 5(10): 1300-1314.

Liliana Alves, Nuno Antunes, Olga Agrici, Carlos M. R. Sousa, and Celia M. Q. Ramos. 2016. *Click Bait: You Won't Believe What Happens Next. Fronteiras: Journal of Social, Technological and Environmental Science*, 5(2):196-213.

Lin, S.-P.. 2016. Readers under Cultural Forces of Competitive News Sources: A Study on Taiwanese Local Newspaper Audience. *Journal of Communication Research and Practice*, 6(2): 173-207.

McCawley, James D. 1971. *Prelexical syntax. Report of the 22nd Annual Roundtable Meeting on Linguistics and Language Studies,* Washington, D.C.: Georgetown University Press.

Otto, L. Glogger, I. and Boukes, M. 2017. *The Softening of Journalistic Political Communication: A Comprehensive Framework Model of Sensationalism, Soft News, Infotainment, and Tabloidization. Communication History*, 27(2): 136-155.

Reinemann, C., Stanyer, J., Scherr, S., Legnante, G., Esser, F., Strömbäck, J., and De Vreese, C. 2012. *Hard and Soft News: A Review of Concepts,*

Operationalizations and Key findings. Journalism, 13 (2): 221-239.

Schaffer, D. 1995. *SHOCKING SECRETS REVEALED! The Language of Tabloid Headlines. ETC: A Review of General Semantics*, 52(1): 27-46.

Skovgarrd, M. 2014. *A Tabloid Mind? Professional Values and Organizational Pressures as Explanations of Tabloid Journalism. Media, Culture and Society,* 36(2): 200-218.

On the "Easy" Task of Evaluating Chinese Irony Detection

An-Ran Li
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
an-ran.li@connect.polyu.hk

Emmanuele Chersoni
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
emmanuelechersoni@gmail.com

Rong Xiang
Department of computing, The Hong Kong Polytechnic University
csrxiang@comp.polyu.edu.hk

Chu-Ren Huang
Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University
churen.huang@polyu.edu.hk

Qin Lu
Department of computing, The Hong Kong Polytechnic University
qin.lu@polyu.edu.hk

Abstract

In this paper, we present a discussion on the problem in the evaluation of irony detection in Mandarin Chinese, especially due to the difficulties of finding an exhaustive definition and to the current lack of a gold standard for computational models. We describe some preliminary results of our experiments on an irony detection system for Chinese, and analyze examples of irony or other related phenomena that turned out to be challenging for NLP classifiers.

1 Introduction

In recent years, irony became a hot topic which draws the attention of both cognitive linguists and computational linguists. As a special kind of rhetorical device, its most striking feature is the incongruity between its literal meaning and contextual meaning. This feature means that the processing procedure of irony should be more complex than other expressions for both humans and machines. And since ironic expressions are highly context-dependent while analyzing contextual information is not easy for computational systems, the automatic detection of irony is a hard task. However, if we cannot effectively detect it, entire sentences will be understood in a totally different way, affecting the performance in many NLP tasks.

Generally speaking, ironies are often defined as the expressions whose literal meanings are incongruous with their contextual meaning. However, according to our observation, the use of the word "incongruous" is inadequate. Some other kinds of expressions can also show incongruities between their literal and contextual meanings:

- Exaggeration or Hyperbole:

(1) I was fired and caught a cold in the same day. I must be the most luckless people in the world!!!

The speaker should know that there must be some more luckless people than him/her in the world. He/she does not mean to state the fact that he/she is the most luckless people, but just

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 452-460, Hakodate, Japan, September 13-15, 2019

want to express his/her strong emotion by the exaggeration device.

- Metaphor:

(2) Mark Twain's work is a mirror of America.

The speaker does not use simile by using the words "as" or "like" in this sentence. Literally speaking, the meaning of the sentence is that "Mark Twain's work is a mirror". However, it is obvious that literature works cannot be a real mirror. The contextual meaning of the sentence is that Mark Twain's work shows the actualities of America.

Another rhetorical device which needs to be emphasized in this discussion is the pun. Ironies and puns at least have the following similarities:

- They both have some incongruities in several linguistic levels;
- In communication, not all the listeners can understand their meaning;
- Speakers may make pun / ironies unintentionally;
- There are bad ironies/puns like "icy jokes". That is, the expression is so unfunny / non-ironic / non-punny that it is kind of funny / ironic / punny.

However, they are not the same concept, neither. Puns just ask for double entendre. If an expression can express more than one meaning, it can be a pun.

For example:

(3) 人类失去联想，世界将会怎样？
ren2 lei4 shi1 qu4 lian2 xiang3, shi4 jie4 jiang1 hui4 zen3 yang4?
What will happen to the world if human lost their imagination?

This is a famous advertising slogan of Lenovo (a technology company) in China. Their Chinese name "联想 (lian2 xiang3)" means "imagination" in Chinese, so here the word "联想 (lian2 xiang3)" is a pun. It can not only refer to imagination, but also to Lenovo. However, since "imagination" and "Lenovo" are not contrast with each other, it just a pun instead of an irony.

(4) The dear leader played the "trump" card and played it very well.

In this sentence, the word "trump" refers to the Joker card in poker game, its basic meaning. It can also refer to the advantage that makes people more likely to succeed. It can even be a pun since the word "trump" can also refer to President Trump. However, this sentence is not necessarily an irony. Only when we can get further context, which can prove that the speaker is an opponent of President Trump (or at least, not the supporter of him), the sentence can be interpreted as ironic.

Compare with puns, although most of the ironies also have double meanings, there are some restrictions. If the two meanings of an expression are just incongruous instead of contradict with each other, it can't be an irony. Besides that, having double meanings even isn't an essential condition to ironies. For expressions like counterfactuals (such as "太阳从西边出来 (tai4 yang cong2 xi1 bian1 chu1 lai, The sun rise in the west.)") and satiations (such as "你相信吗？你相信吗？ (ni3 xiang1 xin4 ma? ni3 xiang1 xin4 ma?, Do you believe? Do you believe?)"), all the words are in their original meanings. However, the listeners still feel they are ironic.

From examples above, we know that a suitable definition of irony should not only show its linguistic features but also can effectively differentiate them from other linguistic phenomena. Both "incongruity" and "opposite" are not sufficient or necessary features, although it can involve both of them. Huang (2019) proposed that "reversal" is the critical nature of irony. All the features including "incongruity" and "opposite" can be seen as tools to achieve this "reversal".

The concept of "reversal" comes from "reversal theory" (Apter, 1982) in social psychology field. It is an important theory for personality changes as well as for change of belief / knowledge, in the context of studies on persuasion. This concept can include not only the reversed meanings, but also the situations in which the literal meaning is left unchanged and the reversal concerns only its semantic or pragmatic effect.

In this paper, we accept this definition and view irony as the expression which makes people experience a reversal during the understanding process.

In our experiments, we tried to use classifiers and word embedding methods to classify different types of ironic expressions, on the basis of the available resources. Starting from the definition basing on the concept of "reversal", we hope to find an efficient method to build reliable dataset which can be used to train and test irony detection models. Then we also apply our dataset to some machine learning models and report our preliminary results.

2 Related Work

A lot of efforts have recently been devoted to irony detection in natural language processing. The model by Buschmeier et al. (2014) give more than ten features including Imbalance (the overall polarity of words is different from the polarity of the whole sentence), Hyperbole (a sequence of three positive or negative words in a row), Quotes (up to two consecutive adjectives or nouns in quotation marks have a positive or negative polarity), Pos/Neg Punctuation (a span of up to four words contains at least one positive (negative) but no negative (positive) word and ends with at least two exclamation marks or a sequence of a question mark and an exclamation mark (Carvalho et al., 2009), Punctuation, Interjection (such as "wow" and "huh"), Laughter and Emoticon. Basing on these features, they use five different classifiers to analyze an Amazon review corpus which has 437 ironic reviews as well as 817 non-ironic reviews. The best F1-score they reported (74.4) is reached by combining star-rating with bag-of-words and specific features, and then using Logistic Regression to classify the corpus. A lot of later researches use features that are similar to their set.

Comparing with them, Reyes and Rosso (2014) used more lexical features. They divided the features they used into three layers. The signatures layer includes the features like Pointedness (refers to the contents that reflect a sharp distinction in the information. E.g. punctuation, emoticons and capitalized words), Counter-factuality (words hint an opposition. E.g. nevertheless, nonetheless and yet) and Temporal compression (words represent an abrupt change. E.g. *suddenly*, *now* and *abruptly*). The emotional scenarios layer includes the features like Activation (means the degree of emotion), Imagery (whether the words are easy to form a mental picture) and Pleasantness (the degree of pleasure of the words). The unexpectedness layer includes the features like Temporal Imbalance and Contextual Imbalance (whether there is an opposition of polarity or attitude in the time line / context). Their basic idea is much closer to our definition of irony since the three triggers of reversal (contingent events, frustration and satiation, see Apter, 1982) are all included in their features. They claimed that negation should be a useful grammatical category to detect ironies and also report the difficulties in the automatic detection task.

Sarcasm is a rhetorical device which share many important features with irony. The main difference between irony and sarcasm is whether the speakers intend to hurt someone by their words. Similar to irony, sarcasm experience a reversal in both meaning and sentiment, so sarcasm detection models can also give us some inspiration on irony detection. Ghosh et al. (2015) used a word embedding method to detect sarcasm and introduce a useful platform Amazon Mechanical Turk (MTurk). This platform can rephrase the sarcastic utterances to their intended meanings (e.g. "I love going to the dentist" can be rephrased as "I hate going to the dentist" or "I don't like going to the dentist"). They use this platform to rephrase 1,000 sarcastic Tweets and get 5,000 sarcastic – intended message pairs (each sarcastic message has five intended candidates). Meanwhile, they use co-training algorithm and statistical machine translation alignment method to extract 80 semantically opposite paraphrases. By extracting the context vectors with word embedding, they got the contextual features of each sarcastic utterance as well as its intended pairs. By using the distributional approach *w2vsg* with the Kernel classifier, they achieved the highest F1-score of 97.5% in their study.

Joshi et al. (2017) summarized the main approaches on sarcasm detection tasks. Rule-based approaches focus on the rules which rely on indicators of sarcasm. Feature Sets approaches usually use bag-of-words as features. Learning algorithms mainly rely on different kinds of SVM models. And deep learning-based approaches can give us further insights when the datasets are big enough. They claimed that pattern discovery was the early trend in sarcasm detection, while the use of context will be the new trend of the task.

Author-specific context, conventional context and topical context will play more and more important roles in future research, which can be also be true for the irony detection task.

However, researches on Chinese irony detection are still limited. Only one Chinese irony corpus has been built by Tang and Chen (Tang and Chen, 2014). Basing on the NTU Sentiment Dictionary, they extracted 304,754 messages from Plurk. These messages are the texts with negative emoticons and positive words. They believed they are potential candidates of ironic expressions. Then they retrieved all the messages which contain the pattern "degree adverb + positive adjective" from the candidates then manually reviewed whether they are ironic expressions.

During the review task, if annotators found some new irony patterns, they would also retrieve all the messages which contain this new pattern and then manually check them. Finally, they found 1,005 ironic messages from all the candidates and divided them into five groups according to the patterns they have. These patterns are: degree adverbs + positive adjective, positive adjective with high intensity, positive noun with high intensity, "很好" (hen3 hao3, very good) and "可以再…一点" (ke3 yi3 zai4…yi4 dian3, It's okay to be worse).

They used these patterns to extract ironic expressions from Yahoo corpus and obtained 36 ironic texts from it. Their work is, without a doubt, a meaningful try but the patterns that they found are too short. A lot of dynamic and relatively abstract ironic expressions are not included in their corpus and whether just use one pattern (degree adverbs + positive adjective) at the very beginning of the bootstrapping procedure is adequate for this task is worth discussing.

Besides that, Deng, Jia and Chen (2015) construct a feature system for Weibo irony identification task. The system contains six features:

- the basic emotion feature of the words in the sentences: be recorded by unigram
- homophonic words: such as "河蟹 (he2 xie4, river crab)" and "和谐 (he2 xie2, harmony)"
- sequential punctuations: more than three
- length of the text: Weibo texts are divided into short, middle and long. They believed that the length of the text will affect the quantity of sentiment information.
- verb passivization: abnormal collocation of the structure "被 + verb" like "我被就业了 (wo3 bei4 jiu4 ye4 le, I am gotten a job)
- incongruities between emotions in and out of the quotation marks: whether the emotion words in the quotation marks is positive while the emotion words out of the quotation is negative or vice versa.

Basing on this system, they reported the highest precision rate and F-score from the Logistic Regression Model (Precision rate: 78.31%, F1-score: 71.13%) and the highest recall rate from Decision Tree Model (71.86%).

From current studies we can see that now we lack of Chinese irony resources. It is no doubt a big problem. On the one hand, we do not have a suitable corpus for both machines and researchers to extract features and find patterns. Only hundreds of examples cannot effectively help us to summarize the rules. Moreover, they usually do not cover enough types of ironic expressions. On the other hand, since both the theoretical and applied researches on Chinese ironies are limited, we do not have an adequate corpus as well as a standard to evaluate the quality of Chinese irony detection. However, constructing such a corpus completely by annotators is a hugely difficult task since ironic expressions account for a very small percentage in most corpora (usually less than 1%). In other words, ironic expressions are just like needle in the haystack. Therefore, it is meaningful to find a method which can filter ironic expressions automatically and precisely.

3 Classification Experiments

3.1 Data Collection

For our study, first we need to build a provisional dataset for the classifiers. The dataset need to include enough ironic expressions as well as non-ironic expressions that share some features with them, as representatives of the negative class. The Taiwanese irony corpus built by Tang and Chen (2014) is a suitable resource to form the ironic part. According to what they reported in their paper, this corpus has 1,005 ironic messages

collected from Plurk and five ironic patterns can be extracted from them. Therefore, its scale is big enough for the classifiers to detect features. And since it is manually-checked, the expressions are reliable and typical.

The non-ironic expressions, for the moment, are of two different types. The first type is sentences extracted from microblogs. Among those 1,005 ironic messages in Taiwanese corpus, 993 of them have both "positive sentiment" tag and "ironic" tag. It means that they are not only ironic but also have positive lexicons. We use them as patterns. Meanwhile, we screen out all the positive sentences from two sentiment-labeled microblog corpora as candidates. After that, we calculate all the cosine similarities between each patterns and candidates by using bag-of-word vectors. We filter out all the candidates whose cosine similarities are less than 0.5 and choose the best five candidate matches of each pattern. We finally extract 2,241 candidates from microblogs corpus. All of these sentences share high similarities of words with the ironic expressions in Taiwanese corpus. However, they are non-ironic since the sentiment tag of each whole sentence is still positive.

The second type is puns. As we mentioned, puns share a lot of similarities with irony. It can even confuse humans in some cases so we wonder whether they are also confounding factors to computers. Introducing puns in our dataset can broaden our range from a theoretical point of view, no matter what the classification results show us.

If the classifiers can correctly classify most of (or even all of) the ironic puns as ironies and filter out the non-ironic ones, it is no doubt an exciting result to show that the detection method is strong enough to filter out irony-like expressions from real ironies. If the classifiers classify all the puns as ironies, at least it shows that the filter can identify double entendre from other expressions. Finding out the features that different kinds of double entendre have in common and work out why these features are effective enough to differentiate double entendre from other expressions should be a new and meaningful topic to do research on. Even if the result shows that ironic and non-ironic puns are classified randomly, it can also be a treasurable resource for error analysis and future researches. There should be some rules inside the wrong results. Why some of the non-ironic puns can confound the classifiers while others cannot? It is also a worthwhile topic.

For this part of our dataset, first we extract 906 candidates online. Then we manually checked these candidates to find typical puns. The second meaning of these puns can be easily recognized and the two meanings are different enough to differentiate from each other. We finally selected 176 puns from 906 candidates. Besides that, we also add 30 *xiehouyu* to the dataset. *Xiehouyu* is a kind of Chinese idiom that usually has two parts. The first part of it is descriptive while the second part carries its double meanings. Since *xiehouyu* are conventional expressions, no matter whether the second parts are shown in the discourse, they are typical and popular template for puns.

Finally, the three parts we mentioned above construct our classification dataset. They are:

Positive examples:

- 993 ironic expressions with positive lexicons (from Taiwanese corpus, Tang and Chen (2014))

Negative examples:

- 2,241 non-ironic expressions which have high similarities with those ironic expressions (from sentiment-labeled microblog corpora, Zhou et al. (2018) and Wang et al. (2016))
- 206 puns: 176 complete sentences with typical puns (randomly extract from different websites) and 30 popular xiehouyu.

Therefore, now we have 3440 sentences in the database. We automatically mix them and divide them into training set (3,097 sentences) and test set (343 sentences).

3.2 Classification Task and Results

In this section we use two widely-used classifiers (Support Vector Machine (SVM) and Logistic Regression (LR)) to train and classify our data. The Support Vector Machine takes as input a sentence feature vector, a representation that is

computed as the mean vector of the embedding of the words in the sentence. The results of SVM are as follows:

Kind of Sentences	Number of Sentences	Precision	Recall	F1-score
Negative Examples	244	0.98	0.97	0.97
Positive Examples	99	0.92	0.95	0.94
Average / Total	343	0.96	0.96	0.96

Table 1: Results of Support Vector Machine (SVM)

Similar to SVM, the Logistic Regression (LR) also uses a form of binary model to analyze the input sentences. Here the results of LR are as follows:

Kind of Sentences	Number of Sentences	Precision	Recall	F1-score
Negative Examples	244	0.98	0.98	0.98
Positive Examples	99	0.96	0.96	0.96
Average / Total	343	0.98	0.98	0.98

Table 2: Results of Logistic Regression (LR)

As we can see, both of the classifiers show good results. It is no doubt excited but it seems that the classification of irony sentences is too easy for computational models. Therefore, we still wonder that whether the limitation of the data affect the result. The reason why we have this consideration is that the ironic sentences in the Taiwanese corpus just have five typical ironic constructions but in actual discourses there should be more. We want to confirm whether the diversity of our data lower the difficulty of the classification task.

In future, we plan to collect more varieties of ironic expressions to extent our database. However, not matter whether the results will change significantly, our method should be a good standard to evaluate Chinese irony detection tasks.

3.3 Error Analysis

According to the results of LR model, these sentences are not correctly classified.

Ironic sentences which are classified as non-ironic:

(5) 以為訂的是晚上七點半回新竹的票，七點在自動售票機取到票才發現是晚上八點。特地再到網路訂位取票的窗口問可不可以換到七點半，結果票務人員發現我訂的是晚上八點從新竹到台北的票。我真的可以再天兵一點。。。:'-(

(6) 很好，團購了一個可以拿來澆花的、可能會摔破的、大陸製自行車水壺:-(

(7) 我真是太幸運啦！今天朋友抽籤，竟然抽到裡面最老的一位奶奶，８４歲。我真的是開心到不知道要說什麼．．．就是傻很大～而且他又喜歡話中代日語，所以我想．．這次專題真的很有挑戰性～:-(

(8) 今天真是太太太幸了,在台九上，大卡迎面而，不知哪一拳大的石，1秒2秒的，恰恰好直落在我的上方!碰的巨我－跳:-好人平安。

Non-ironic sentences which are classified as ironic:

(9) 回到纯爱的美好记忆。”””””欢迎爱光临”””””。”

(10) 飞蛾扑火-自取灭亡

(11) 上完课啦 终于告别早起的苦逼日子来全家吃个盒饭补充一下元气先下午接着上bec3 fighting!!!

(12) 为公司年会特意准备的谢谢！是不是很有：欧巴桑的feel“‘o(∩)o

We are not sure why (5) and (6) are not correctly classified since it seems that they have enough features for classifiers to make correct decisions (Example (5) have ironic features “ 真的 (zhen1 de, really)” and “ 可以再…一点 (ke2 yi3 zai4…yi4 dian3, it can be more…)” with negative emoticon “:'-(”. Example (6) has an ironic

feature as well as positive adjective " 很好 (hen2 hao3, very well)" with negative emoticon ":-(".). However, for example (7) and (8), even humans may also confuse whether the speakers are ironic. It is because that the event they described can either be considered as lucky or unlucky. The judgements just rely on the parties view the events from which angle. However, the speakers here use a lot of positive markers with relatively same quantity of negative markers. These markers will give too many vectors to the sentences and finally confuse the classifier. For non-ironic sentences, example (9) and (12) just have positive markers, we guess maybe the occurrences of sequential punctuations and seldom-used emoticon are marked as ironic features. Example (11) may be affected by the co-occurrence of "苦逼 (ku3 bi1, bitter)" and "fighting". According to this analysis, how to give different ironic constructions a reasonable weight in the classification task should be a meaningful topic.

Meanwhile, although we are not sure about the reason why example (10) are classified as irony, it is more than excited to see all the puns are correctly classified except it. It shows us although both irony and pun have double meanings, there must be some features which are strong enough to differentiate them from each other. It also supports our hypothesis that the critical nature of irony it reversal instead of incongruous or opposite in meaning since puns also have the latter. Finding out the features which can differentiate ironies from puns can be a valuable theoretical contribution.

4 Towards New Datasets for Chinese Irony Detection

As what we mentioned in 3.2, we wonder whether the scale of the database will affect our results. In order to richer the database, we need more ironic sentences as positive examples. These sentences must be as various as possible so that we can include enough actual instances for the classifiers to extract features. In the first step, we use the strategy which is similar to the Taiwanese corpus. We'll use some ironic constructions as key words to find candidates then manually check them. During the checking task, we may find some new ironic constructions. We'll use these new constructions to find more candidates as well as more ironic sentences. It just like makes a snowball. Using more constructions as key words in this step should be good for us to get more candidates. According to our definition, ironies should either express or facilitate reversals at different linguistic levels, so ironic detection should be more effective and accurate if we model it as a reversal detection instead of an incongruity detection task. For now, we've found at least seven kinds of ironic reversal:

1. Rhetorical Reversal: In Chinese, rhetoric questions can be formed in different ways such as:

   a) adding tag question with the verb 是 (shi4, to be) followed by a question particle 吗 (ma).

   b) adding emphasis with wh-words on manner/degree. For example:

(13) noun phrase+有这么+ verb phrase +的吗

Noun phrase + you3 zhe4 me + verb phrase + de ma

Is there anything can be done like this?

(14) 你以为你是谁？

ni3 yi3 wei2 ni3 shi4 shui2?

Who do you think you are?

   c) repetition of a normal question: It is a kind of satiation in reversal theory. Repeating an expression (no matter it is a question or not) again and again will makes listeners to question whether it is in its original meaning. It indicates stronger ironic intention.

2. Imperative sentences as dares: It is a kind of threaten to stop listeners from doing what they dare to do. Speakers use imperative but actually it is a prohibition. For example:

(15) noun phrase + 再 + verb phrase + (一+ quantifier) + (试试)

noun phrase + zai4 + verb phrase + (yi2 + quantifier) + (shi4 shi)

(Somebody) can (try to) do it (once more)

3. Evaluative reversal: This kind of reversal usually include some special lexical markers such as "亏 (kui1, fortunately)", which is marked in 现代汉语词典 (Xian Dai Han Yu Ci

Dian, 2016) to express irony/sarcasm.

4. Opposite pairs: This kind of expressions show ironic meaning by directly using contrastive linguistic pairs. (Ding, 2018)

5. Counterfactual constructions: These constructions reverse the factuality of a statement. It can be marked with adverbs such as "要不是(yao4 bu2 shi4, but for)", or formulaic counterfactual expressions such as "太阳从西边出来(tai4 yang cong2 xi1 bian1 chu1 lai, The sun rise in the west.)". (Jiang, 2019)

6. Reversal of sentiment: This happens when positive emotion words are used to express negative emotion, and vice versa.

7. Satiation: As what we mentioned, if speakers repeat an expression several times, listeners will question whether it is in its original meaning. Similarly, if speakers overuse certain polarity words (such as hyperbole), the listeners will also experience a reversal. If there are more than one assertive words or high degree adverbs in one sentence, it is highly possible to be an ironic expression.

Each kind of reversal can separate out more than one ironic constructions. Only using constructions from first four kinds we can easily extract 2,363 candidates from a single microblog corpus. Since most of the constructions are highly formalized and easy to retrieve, we are confident of finding more ironic constructions as well as positive examples by this method.

Meanwhile, in order to manually check the candidates in a standard way, similar to what Pragglejaz Group (2007) and Gerard J. Steen et al. (2010) did on metaphors, we construct an Irony Identification Procedure (IIP) to help annotators to make judgements. In short, the procedure should be as follows:

1. Read the entire sentence as well as the context (if available) to sketch an overall understanding of the meaning.

2. Determine the contextual meaning of core constructions of the text. These core constructions include idioms, adjective phrases, rhetorical devices, clauses which are linked by conjunctions and some other constructions which can express the attitudes of the speakers. Annotators should pay special attention to sentiments, evaluations and logic relations which are shown by these constructions in the given context.

3. Determine the literal meaning of each core construction. When finding literal meanings, researchers should neither consider about the construction meanings emerge after the combination of the components nor refer to any context. Literal meanings have to be: direct (can be understood without any context), formal (can be found in dictionaries) and common (frequently-used but do not use any rhetorical devices).

4. Compare the contextual meaning and the literal meaning of the construction to see whether the contextual one is the reversal of the literal one. Researchers should notice that the evaluation criterion is whether there is a reversal in the expression instead of just "incongruous". For example, if the literal sentiment of the construction is joy while the contextual sentiment of the construction is grossness or even wrath, it can be a reversal. If sentiment just changes from joy to excitement or from grossness to wrath, it is an "incongruity".

5. If the contextual meaning of a construction experiences a reversal from its literal meaning, mark it as an "ironic construction". If it hasn't been included in current ironic construction set, add it to the set and further use it to retrieve new candidates.

6. Basing on core constructions, judge whether the whole text experience a reversal. If so, chose it as a positive example.

As what Joshi et al., 2017 claimed in their paper, pattern discovery was the early trend of sarcasm detection and researchers will rely more on context information in the future. Therefore, we will also try to take context features into consideration. Features like logic confusion and topical context will be new topic we concern about.

References

Micheal J Apter. 1982. The Experience of Motivation: The Theory of Psychological Reversals. Academic Press.

Micheal J Apter. 1984. Chinese Irony Corpus Construction and Ironic Structure Analysis. In *Journal of Research in Personality*, vol. 18(3): 265-288.

Konstantin Buschmeier, Philipp Cimiano, and Roman Klinger. 2014. An Impact Analysis of Features in a Classification Approach to Irony Detection in Product Reviews. In *Proceedings of the EMNLP Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*.

Paula Carvalho, Luis Sarmento, Mario Silva, and Eugenio De Oliveira. 2009. Clues for Detecting Irony in User-Generated Contents: oh...!! It's So Easy ;-). In *Proceedings of the International CIKM workshop on Topic-Sentiment Analysis for Mass Opinion*.

Zhao Deng, Xiu-Yi Jia, Jia-Jun Chen 邓钊, 贾修一, 陈家骏. 2015. Research on Chinese irony detection in microblog 面向微博的中文反语识别研究. In *Computer Engineering Science* 计算机工程与科学, vol. 37(12):2312-2317.

Jing Ding. 2018. A lexical semantic study of Chinese opposites. Springer Singapore, Singapore.

Debanjan Ghosh, Weiwei Guo, and Smaranda Muresan. 2015. Sarcastic or Not: Word Embeddings to Predict the Literal or Sarcastic Meaning of Words. In *Proceedings of EMNLP*.

Pragglejaz Group. 2007. Chinese and Counterfactual Reasoning. In *MIP: A Method for Identifying Metaphorically Used Words in Discourse. Metaphor and Symbol*, Vol. 22 (1): 1-39.

Chu-Ren Huang. 2019. Double Meaning and Reversal: Toward an empirical linguistic account of irony. In 2*019 Joint Conference of Linguistic Societies in Korea The 26th Joint Workshop on Linguistics and Language Processing* (JWLLP-26), Seoul, Korea.

Yan Jiang. 2019. Chinese and Counterfactual Reasoning. In Chu-Ren, H., Barbara, M., and Zhuo, J.-S. (eds.):*The Routledge Handbook of Chinese Applied Linguistics*. Routledge, Abingdon.

Aditya Joshi, Pushpak Bhattacharyya, and Mark J Carman. 2017. Automatic Sarcasm Detection: A Survey. In *ACM Computing Surveys*, vol. 50(5).

Shu-Xiang Lv, Sheng-Shu Ding (eds.) 吕叔湘, 丁声树 编撰. 2016. Modern Chinese Dictionary 现代汉语词典. The Commercial Press 商务印书馆, Beijing, China.

Antonio Reyes and Paolo Rosso. 2014. On the Difficulty of Automatically Detecting Irony: Beyond a Simple Case of Negation. In *Knowledge and Information Systems*, vol. 40(3): 595–614. Springer.

Gerard J. Steen, Aletta G. Dorst, J. Berenike Herrmann, Anna A. Kaal, Tina Krennmayr and Trijntje Pasma. 2010. A Method for Linguistic Metaphor Identification: From MIP to MIPVU. John Benjamins Publishing Company, Amsterdam/Philadelphia.

Yi-Jie Tang and Hsin-Hsi Chen. 2014. Chinese Irony Corpus Construction and Ironic Structure Analysis. In *Proceedings of COLING*.

Zhongqing Wang, Yue Zhang, Sophia Yat Mei Lee, Shoushan Li and Guodong Zhou. 2016. A bilingual attention network for code-switched emotion prediction. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 1624-1634.

Frank Z Xing and Yang Xu. 2015. A Logistic Regression Model of Irony Detection in Chinese Internet Texts. In *Research in Computing Science*, no. 90: 239–249.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu and Bing Liu. 2018. Emotional chatting machine: Emotional conversation generation with internal and external memory. In *Proceedings of Thirty-Second AAAI Conference on Artificial Intelligence*.

Towards Better Ad Experience: Click Prediction Leveraging Sequential Networks Derived Specifically From User Search Behaviors

Shengzhe Li [†], **Tomoko Izumi** [§], **Yu Kuratake** [§], **Jiali Yao** [§], **Jerry Turner** [§], **Daisuke Kawahara** [†] and **Sadao Kurohashi** [†]
[†]Graduate School of Informatics, Kyoto University, Sakyo-ku, Kyoto 606-8501, Japan
[§]Microsoft Development Co., Ltd., Minato-ku, Tokyo 108-0075, Japan
{lisz, dk, kuro}@nlp.ist.kyoto-u.ac.jp
{tomitsui, yu.kuratake, jiayao, jeromet}@microsoft.com

Abstract

We propose a sequential modeling approach to improve click prediction for search engine advertising. Unlike previous studies leveraging advertisement content and their relevance-to-query information, we employ only users' search behavioral features such as users' query texts and actual click records of both organic search results and advertisements. By leveraging long short-term memory (LSTM) networks, we successfully modeled users' sequential search behaviors and fully utilized them in click predictions. Through experiments conducted with large-scale search log data obtained from an actual commercial search engine, we demonstrated that our method combining users' current and previous search behaviors reaches better prediction performance than baseline methods.

1 Introduction

Search advertisement (ad) is the fundamental source of revenue for internet search services. It plays a major role in digital advertising, which is estimated to account for 43.5% of the whole advertising market.[1,2] In order to gain more ad clicks, a search engine may simply show more ads to the user. Although this may help to increase revenue in a short time, it could hurt long term revenue as this increases users' ad blindness (Hohnhold et al., 2015), meaning they learn to simply ignore ads. Also, showing many unsatisfactory ads may eventually result in losing users, letting them switch to another search engine. For example, when a user searches about "Amazon CEO", the user's intent is obviously to look for information. But a search engine may understand as a search for item "CEO" at Amazon. This kind of errors may dissatisfy users, because mismatched ads occupy the best position in the page, where is supposed to be the answer of the information lookup. In order to satisfy both users and advertisers while not deteriorating the search service's revenue opportunities, it is essential to show ads at the right time when the user's search intent matches with the ad. This will lead the user to conduct a desired action, such as purchasing a product, at the advertiser's site. Therefore, when a search arrives, it is necessary to determine whether it is appropriate to display ads. In other words, if ads are displayed, we predict whether user would click at them. If the probability of click is high, we induce that it is suitable to present the ads. Otherwise, it would be better not to display ads. Therefore, the problem is converted to click prediction, and we determine the appropriateness of the ad display according to the click prediction results.

Click prediction is a widely used technology to improve the ad-related user experience by increasing click through rate (CTR). Click prediction anticipates the probability of ads to be clicked by leveraging various information such as ad contents, ad position, relevance scores between ads and queries, and detailed user's intent signals such as dwell time of each ad click.

[1]https://www.emarketer.com/content/emarketer-total-media-ad-spending-worldwide-will-rise-7-4-in-2018

[2]https://www.statista.com/outlook/216/100/digital-advertising/worldwide

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 461-470, Hakodate, Japan, September 13-15, 2019

However, if search ads are provided through a 3rd party ad platform, the search service may suffer poor ad performance. Because ads are often served through a simple API request between the search service and the ad platform, it is difficult to communicate complex signals such as user's search histories and relate those signals with ad features such as ad contents to maximize ad performance. Due to the limitations on available features, the ad performance of the search engine could not be optimized for the user and it may simply result in showing lots of ads (ad over-triggering) and thus hurt user experience. As of today, there are many internet search services that provide ads through a 3rd party ad platform, and so this is not a trivial issue.

In order to overcome this constraint and provide a better ad experience, we propose a click prediction approach leveraging sequential networks derived specifically from user search behavioral signals available for the search service. Our method utilizes long short-term memory (LSTM) networks to capture the user's one-shot search intent and overall personal preference over ads, and leverages this information to estimate the probability of ad clicks. In this study, we focus solely on user behavioral features, and thus the prediction model is designed without any ad-related information. Our experiments that used a large-scale search log validated the effectiveness of our proposed method.

The remaining of this paper is organized as follows. Section 2 summarizes previous studies on click prediction, Section 3 introduces our methodology of sequential ad click prediction model using users' search behavioral features. Section 4 details our experiments and analysis. Section 5 concludes the paper.

2 Related Work

Studies on ad click prediction have a long history. Using logistic regression with statistical features is the most common method in ad click prediction (Richardson et al., 2007; McMahan et al., 2013; He et al., 2014) because of its small computation complexity but relatively good performance. In recent years, factorization machines (FMs) (Juan et al., 2016; Chen et al., 2009; Ta, 2015), gradient boosting decision trees (GBDTs) (Trofimov et al., 2012), conditional random fields (CRF) (Xiong et al., 2012), and deep neural networks (DNNs) (Zhang et al., 2016) have also been utilized for the ad click prediction task within a single ad impression, and have achieved impressive results. Ling et al. (2017) make an ensemble of these models and apply the ensemble model to a real world search engine.

Recently, sequential ad click prediction based on user behavior has become a hot topic. For example, recurrent neural networks (RNNs) (Mikolov et al., 2010) are used to model users' click and behavioral sequences (Zhang et al., 2014; Liu et al., 2017a) for its good ability at capturing sequential information. In some other related tasks like query suggestion, RNN-based approaches also show their superiority (Chen et al., 2018; Sordoni et al., 2015). Recent studies also use long short-term memory (LSTM) networks (Hochreiter and Schmidhuber, 1997) to model query sequences thanks to its better ability on handling long sequences than RNNs (Deng et al., 2018; Zhu et al., 2017). Moreover, it is observed in Zhang et al. (2014) that the longer the time span between different two searches is, the less impact the former search brings to the latter search. Therefore, hierarchical architectures are proposed to model the difference impact by query level, session level and user level (Sordoni et al., 2015; Chen et al., 2018). All these pieces of evidence indicate that in addition to the information obtained from the current query, short-term and long-term historical features also play an important role in predicting user's ad click behaviors.

In regards to features that reflect users' behaviors, many studies indicate a variety of solutions. First, for the utilization of query texts, employing statistical language models (Salton and Buckley, 1988; Salton et al., 1975; Murdock et al., 2007; Raghavan and Iyer, 2008; Shaparenko et al., 2009; Liu et al., 2015) is common for ad prediction tasks. With the progress of neural networks, deeper utilization of semantics in texts have appeared. CNNs are used to capture semantic information in texts to conduct ad click prediction (Edizel et al., 2017), while RNNs are employed to encode query texts for next query prediction (Sordoni et al., 2015). Zhai et al. (2016) use RNN/LSTM networks to extract intents behind the query texts. In fact, in other areas such as machine translation, sequence-to-sequence model-

ing with an RNN or LSTM text encoder has become a de facto standard (Bahdanau et al., 2014; Neubig, 2017). Besides, in other tasks, we also find evidence of the availability of text encoding. Smith et al. (2018) employ RNNs to encode event texts, which are quite similar to query texts. These references are solid for us to try encoding query texts with sequential neural models. In the meantime, time interval is proved to be an important indicator on user intent (Liu et al., 2017b; Zhu et al., 2017). These related studies provide a solid base for our feature construction.

3 Methodology

First, we define three terminologies that represent how we partition a search log. An **impression** refers to the point when a search/ad result is shown to a user given a query. In this study, we only use ad impressions, which are logged when ads are shown to users. A **session** is a higher level unit which consists of a sequence of impressions within a short period of time (e.g., 30 minutes (Boldi et al., 2008)) by a specific user. Sessions expire due to various reasons such as task completion, timeout and unexpected cutoff. A **user**'s search history consists of a sequence of sessions by the same user.

As stated in Section 1, our model utilizes users' sequential behavioral information to enhance ad click predictions. To make the sequence better imply the user's preference on ads, we need to make this span as long as possible (Zhang et al., 2014). In regards to sequential modeling, we treat every single query as an estimation unit, and the whole history of the user as a sequence, and employ LSTM networks to process the sequence. In each query, we employ a series of features to represent user's search behaviors such as query texts and click histories on both sponsored and organic search results.

3.1 Observation and Principles

Let us consider the flow of a query to observe what information will render a user's search behavior. First, having a specific intent in her mind, the user issues a query from some entrance to the search engine, which we call **entry point**. Given the **query text**, the search service returns both organic and sponsored search results. The user may give **clicks** to both kinds of results. After a **time interval**, the user may search for another query with a different intent. This is a basic search cycle of a single query. Without any ad content information, it is obvious that one single query is not sufficient to predict ad click probabilities. In order to overcome this constraint, we want to learn the user's behavioral features not only from the current query but also from all the previous queries by this user. Especially, queries within the same session could indicate the transition of the user intent, and all the previous queries may indicate the user's personal preference to ads. Therefore, we propose a sequential framework to better capture the user's sequential behavioral information. For each query, we propose a set of behavioral features to represent the user behavior.

3.2 Framework

As shown in Figure 1, we employ an LSTM network to adapt the inter-impression task. Each state of LSTM in Figure 1 is corresponding with an ad impression returned for one query. Note that the queries corresponding with these ad impressions are not necessarily continuous, because queries that do not trigger ads are merged into features. We use the hidden state of cell t in the LSTM as the distributed representation of the t-th query, and use a multi-layer perception (MLP) to decode the hidden state to determine the probability of the ads being clicked. Note that the probability of a click refers to the likelihood that at least one of the ads presented in the result page is clicked. The probability of no click represents that no ads in the page are clicked. Unlike the hierarchical methods proposed in Sordoni et al. (2015) and Chen et al. (2018), we expand all sessions in the user level with the following two main reasons. First, the separation of sessions does not necessarily mean the gap of intent, manual boundaries in the session level may introduce noise. Second, according to the results in Zhang et al. (2014), the prediction for longer histories works better. Therefore, we connect all the sessions, that is, all the ad impressions in a row as shown in Figure 1.

In our work, we use a weighted softmax cross en-

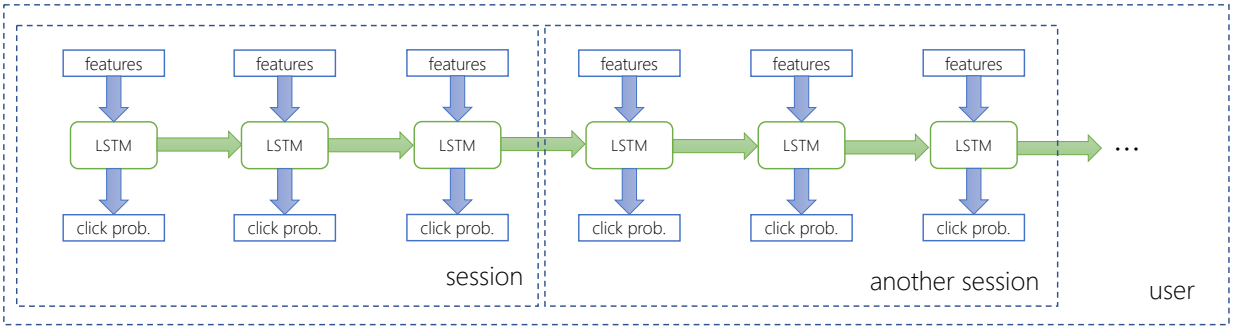


Figure 1: The architecture of our model. We expand all ad impressions of a user in a row, and apply LSTMs to the sequence. In each state, one-shot features are used as the input of the LSTM, and the hidden state is used to predict the ad click probability through an MLP.

tropy as our loss function,

$$L = -\frac{1}{N}\sum_{N}\left[w_0 y_{\text{true}} \log p_0 + w_1(1-y_{\text{true}})\log p_1)\right], \tag{1}$$

where N is the number of training instances, w_1 and w_0 are the weights of clicked and not clicked ad impressions respectively. p_1, p_0 are the predicted probabilities of ad being clicked respectively, and $y_{\text{true}} \in \{0, 1\}$ is the true label indicating whether the ads are clicked.

3.3 Behavioral Feature Construction

As stated above, we define each ad impression as a minimum unit. To reflect user behavioral features as much as possible, we select a series of features that are listed below.

1. **Entry point**. We use signals that indicate where the user issues the query, such as "from the top page of serach service" and "from a browser's address bar".

2. **Query text**. Query text contains most of the information about user's intent for this query. Assume that each query is composed of several words $\{w_1, w_2, \cdots, w_k\}$. Then we obtain their embeddings $Q = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \cdots, \boldsymbol{e}_k\}$. Through an encoder, the query is compressed into a D_h-dimensional embedding, as

$$\boldsymbol{h} = \text{Enc}(\boldsymbol{e}_1\boldsymbol{e}_2\cdots\boldsymbol{e}_k),\ \ \text{Enc} : \mathbb{R}^{k\times D_e} \to \mathbb{R}^{D_h}, \tag{2}$$

where D_e is the dimensionality of word embedding vectors. Since RNN and LSTM networks have shown incredible ability to capture sequential and semantic information in texts, as Smith et al. (2018) have done, we employ bidirectional LSTMs to encode each query as shown in Figure 2, and use the concatenation of the final states of the forward and backward cells as the encoding of the query text as $\boldsymbol{h}_Q = \text{concat}(\overrightarrow{\boldsymbol{h}_k}, \overleftarrow{\boldsymbol{h}_1})$. We compare the results of handling the query texts with mean-pooling, character-level CNNs, bidirectional RNNs, bidirectional LSTMs and CNN-BiLSTMs, and find that bidirectional LSTM is the optimal method.

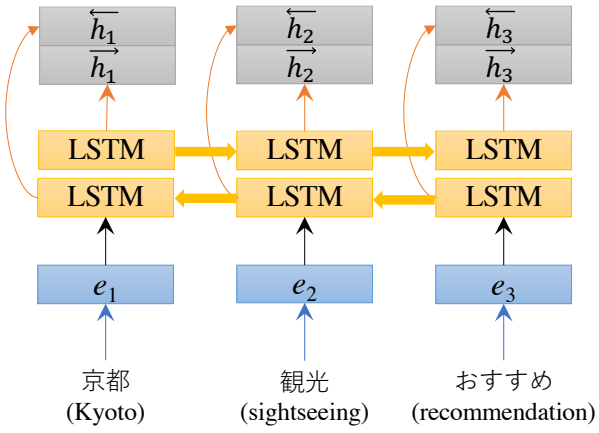


Figure 2: A BiLSTM query encoder to utilize query texts.

3. **Clicks to organic search results**. We count clicks that are made to organic search results between two ad impressions and incorporate this information into our model. We believe

users' clicks to organic search results are also a good indicator to capture the user's search intent. For example, getting more clicks to organic search results is more likely to indicate that the user is looking for pure informational contents, where ads are not useful.

4. **Ad click history**. We straightforwardly use the actual ad click history of the user to predict a personal preference for clicking ads.

5. **Time interval**. We use the time interval between two adjacent ad impressions. We believe that this time interval reflects behavioral features of the user (Liu et al., 2017b; Zhu et al., 2017).

6. **Authentication**. We use the boolean flag that indicates whether the user logged in.

7. **Day of week**. According to our statistics, we observed the likelihood of ads being clicked differs depending of the day of the week, which may be due to the nature of search intent difference between weekends and weekdays. Hence, we believe that it is useful for representing user behavioral features.

As stated in Section 1, our approach is under the limitation that the ad-related information is not available for a search engine employing a 3rd part ad platform. Therefore, no ads related features such as ad contents and relevance scores between ads and queries are used in this study, which are utilized in Zhang et al. (2014).

4 Experiments

We conduct two experiments to validate our proposed approach over a real-world search log dataset. First, we validate the model architecture by comparing it with existing baselines. Then, we explore the impact and relative importance of each feature used in the model. At last, we present two real cases to give an intuitive analysis on the model performance.

4.1 Dataset

We sampled a one-week search log that consists of 18 million impressions from one of the commercial search engines in Japan. From the log, we created a training set with 5.87M ad impressions derived from 508K users, and a validation set with 835K ad impressions from 72.3K users. Because we observed different ad click tendencies between weekdays and weekends, we created two types of test sets, one created from Wednesday's data which consists of 142K ad impressions from 15.4K users, and the other created from Saturday's data which consists of 136K ad impressions from 14.5K users. The Wednesday and the Saturday are in the following week of the week from which we retrieve the training data.

4.2 Evaluation Metrics

We use **AUC** and **F3** scores to evaluate a model from different viewpoints. The AUC score indicates the classification performance of the model, while the F3 score indicates the balance of prediction accuracy and business impact. It is worth mentioning that instead of the F1 score, which indicates the same weights for precision and recall, we set $\beta = 3$ to value recall more. The reason is that recall directly relates to revenue earned by the search engine provider; if a user is highly likely to click ads but the model erroneously predicts that the user would not click ads and does not show ads, the search service would miss its revenue opportunities. We believe that considering the impact to the real business, recall is far more sensitive in this study. Besides, we present a reference measurement, **reduction rate**, to directly measure how much a model is capable to reduce ad displays. A higher reduction rate means more ads removed, but at the same time it increases the risk of losing true clicks. Therefore, we hope to control the reduction rate within an acceptable range instead of making it as high as possible.

4.3 Sequential Framework Effectiveness

In order to prove the effectiveness of our framework, we compare our model with a non-sequential deep neural network method (DNN), as well as logistic regression (LR), which is the most commonly used simple classifier. We use the two test sets to evaluate the performances of these models. The results are listed in Table 1.

The results on the AUC and F3 scores indicate significant superiority of our model to the DNN and LR baselines. Comparing the results of our method and DNN, our utilization of a sequential model like

Table 1: Our model versus two baselines. Our approach is equipped with sequential architecture more than the DNN baseline, and the LR baseline represents with the common industrial practice which is simple and fast. Both ours and DNN are deep neural models, while LR is not.

Model	AUC score (%)			F3 score (%)			Reduction rate (%)		
	Dev.	Test (weekday)	Test (weekend)	Dev.	Test (weekday)	Test (weekend)	Dev.	Test (weekday)	Test (weekend)
Ours	**86.18**	**82.22**	**83.00**	**72.02**	**64.58**	**67.09**	67.63	67.59	65.24
DNN	84.52	80.28	80.97	70.10	63.00	65.73	67.54	65.30	63.19
LR	79.40	76.54	77.01	62.29	56.84	57.98	68.62	73.90	72.91

Table 2: The impacts of each feature tested over the validation set. We remove each feature while keeping other features inboard to see how much the scores deteriorate.

Model	AUC score (%)	F3 score (%)	Reduction rate
Our model	86.18	72.02	67.63
w/o entry point	85.66	69.62	71.84
w/o query text	84.11	70.52	65.19
w/o #click organic	**74.61**	**58.20**	63.59
w/o ad click history	84.40	68.26	68.87
w/o time interval	85.84	71.19	68.89
w/o authentication	86.14	69.11	73.13
w/o day of week	86.10	69.46	72.65

LSTMs indeed improves the performance of ad click prediction. Meanwhile, the improvement of DNN over LR indicates that deeper utilization of information makes sense in this task. Moreover, the comparison between our approach and LR indicates that our method would bring huge improvement over the industrial practice.

Moreover, the results on the two test sets indicate significant differences. This implies the fact that users have higher interest on ads on weekends than on weekdays. This conclusion is in accordance with our intuition that the user search intent is more likely to be informational on weekdays while it is more likely to be transactional, such as shopping, on weekends.

4.4 Feature Impact

In order to discover how much impact each feature brings to the model performance, we remove each feature while keeping other features inboard and observe how much the scores deteriorate on the validation set. A larger deterioration indicates a higher importance of the feature. The results are listed in Table 2.

The user behavioral features include the first five: entry point, query text, organic click count, actual ad click history and time interval, as described in Section 3.3. From the results, it is obvious that the count of organic result clicks from the previous ad impression brings a huge improvement to the model. Without this feature, the AUC score drastically deteriorates by 11.6%, and the F3 score drops by a shocking 13.8%. We explain the reason why it is so powerful as follows. (1) Clicks on organic search results reflect the actual intent of the user's query. From our observation, more clicks to organic results indicate stronger intent to look for information. (2) Only this feature contains sequential information in non-ad impressions. This information would be discarded for non-sequential ad click prediction tasks as no prediction task would be conducted if there is no ad. We also verify that the utilization of query texts is critical, as it could directly reflect the user's behavior and intent. When we stop using query texts, the AUC score drops by over 2%. The ad click history also proves its power with both AUC and F3 results. The drops of 1.78% on AUC and 3.76% on F3 suggest that it is important and useful historical information.

User	Ans.	Pred.	Query	Interval	Auth	Entry	ORC	Ad Click History
User X	0	0	ハヤシライス トマト缶 市販ルー (hashed beef tomato can source)	20	0	8	1	0, 0, 0
	0	0	**子供旅行 おすすめ (good place for traveling with kids)**	18	0	2	1	0, 0, 0, 0
	1	1	**格安航空券 国内 (domestic cheap flight ticket)**	**12**	0	2	**0**	0, 0, 0, 0, 0

Figure 3: A successful case.

User	Ans.	Pred.	Query	Interval	Auth	Entry	ORC	Ad Click History
User A	0	1	パワーポイント (PowerPoint)	<begin>	1	2	0	None
	0	1	パワーポイント 使い方 (how to use PowerPoint)	1	1	8	0	0
	0	0	パワーポイント 結合 (PowerPoint merge)	1	1	8	1	0, 0
User B	1	1	パワーポイント (PowerPoint)	<begin>	1	2	0	None
	0	1	パワーポイント (PowerPoint)	5	1	2	0	0
	0	0	パワーポイント 2018 (PowerPoint 2018)	2	1	8	1	0, 0

Figure 4: A failed case.

Supplementary features including authentication and day of week are also proved to contribute to accurate prediction. Although the differences on AUC are quite small, they have shown considerable impacts on F3. Our interpretation is that they mainly influence the trade-off between precision and recall. In this case, they help the model achieve higher precision, and thus generate differences on F3.

4.5 Case Analysis

We present two actual instances in Figures 3 and 4 to intuitively explain the advantages and shortcomings of our proposed method. Each case contains several continuous ad impressions, which are featurized with query texts, time intervals (larger number means longer interval), authentication (boolean), entry points (category), organic result clicks (ORC; count) and actual ad click histories (boolean sequence).

A successful case is shown in Figure 3, where the second query "good place for travelling with kids" does not yield an ad click while the third query "domestic cheap flight ticket" yields an ad click. In this case, there exists a transition of search intent between the second and the third queries: the second "good place for travelling with kids" being informational while the third "domestic cheap flight ticket" being transactional. To our interpretation, not only the query texts, but also the moderate time interval and the record of zero organic result click contribute to the success of predicting the intent transition, winning over the zero ad click histories.

Meanwhile, a failed case is shown in Figure 4. Two users with exactly the same features searching for the same word "PowerPoint" acted differently. According to their follow-ups, User A has an informational intent, while User B has a transactional one. However, in this case, our proposed model is unable to distinguish the informational intent, as no history is given. This case indicates that our model is highly context-dependent. In the same manner as common models, it cannot handle the randomness of

queries well at the beginning of a user's search log.

5 Conclusion

In this paper, we presented an ad prediction method relying solely on user's sequential search behavioral information. Given the constraint that ad-related information is not available, we only used user's search behavioral features. Through the experiments using real data, we proved that our approach reaches better prediction performance than the baselines, and verified the effectiveness of our sequential framework and behavioral feature construction.

In the future, we will apply our model to an actual commercial search engine to validate the effectiveness of our proposed method for better ad experience. Furthermore, we would like to combine our proposed model with the existing ad click prediction models that leverage ad-related information. Because our proposed method and the existing ad click prediction method are complementary, we believe that our proposed method will contribute to further improving the existing click prediction tasks.

Recently, transfer learning methods have shown a strong ability for improving various NLP tasks. BERT (Devlin et al., 2018) based models refreshed almost all the records of open NLP tasks in one night. Therefore, we believe that it is worth trying to handle the query texts with BERT to further improve the performance of the model. Moreover, we plan to conduct visualization on user's behavior in order to better observe the transition of user intent among queries.

Acknowledgement

This research was supported by Rinna Co-operative Research Project, Microsoft Japan Co., Ltd.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

Paolo Boldi, Francesco Bonchi, Carlos Castillo, Debora Donato, Aristides Gionis, and Sebastiano Vigna. The query-flow graph: model and applications. In *Proceedings of the 17th ACM conference on Information and knowledge management*, pages 609–618. ACM, 2008.

Wanyu Chen, Fei Cai, Honghui Chen, and Maarten de Rijke. Attention-based hierarchical neural query suggestion. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, pages 1093–1096, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5657-2. doi: 10.1145/3209978.3210079. URL `http://doi.acm.org/10.1145/3209978.3210079`.

Ye Chen, Michael Kapralov, John Canny, and Dmitry Y Pavlov. Factor modeling for advertisement targeting. In *Advances in Neural Information Processing Systems*, pages 324–332, 2009.

Weiwei Deng, Xiaoliang Ling, Yang Qi, Tunzi Tan, Eren Manavoglu, and Qi Zhang. Ad click prediction in sequence with long short-term memory networks: an externality-aware model. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, pages 1065–1068. ACM, 2018.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Bora Edizel, Amin Mantrach, and Xiao Bai. Deep character-level click-through rate prediction for sponsored search. *arXiv preprint arXiv:1707.02158*, 2017.

Xinran He, Junfeng Pan, Ou Jin, Tianbing Xu, Bo Liu, Tao Xu, Yanxin Shi, Antoine Atallah, Ralf Herbrich, Stuart Bowers, et al. Practical lessons from predicting clicks on ads at facebook. In *Proceedings of the Eighth International Workshop on Data Mining for Online Advertising*, pages 1–9. ACM, 2014.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8): 1735–1780, 1997.

Henning Hohnhold, Deirdre O'Brien, and Diane Tang. Focusing on the long-term: It's good for users and business. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1849–1858. ACM, 2015.

Yuchin Juan, Yong Zhuang, Wei-Sheng Chin, and Chih-Jen Lin. Field-aware factorization machines for ctr prediction. In *Proceedings of the 10th ACM Conference on Recommender Systems*, pages 43–50. ACM, 2016.

Xiaoliang Ling, Weiwei Deng, Chen Gu, Hucheng Zhou, Cui Li, and Feng Sun. Model ensemble for click prediction in bing search ads. In *Proceedings of the 26th International Conference on World Wide Web Companion*, pages 689–698. International World Wide Web Conferences Steering Committee, 2017.

Pengqi Liu, Javad Azimi, and Ruofei Zhang. Contextual query intent extraction for paid search selection. In *Proceedings of the 24th International Conference on World Wide Web*, pages 71–72. ACM, 2015.

Qiang Liu, Shu Wu, and Liang Wang. Multi-behavioral sequential prediction with recurrent log-bilinear model. *IEEE Transactions on Knowledge and Data Engineering*, 29(6):1254–1267, 2017a.

Yiqun Liu, Xiaohui Xie, Chao Wang, Jian-Yun Nie, Min Zhang, and Shaoping Ma. Time-aware click model. *ACM Transactions on Information Systems (TOIS)*, 35(3):16, 2017b.

H Brendan McMahan, Gary Holt, David Sculley, Michael Young, Dietmar Ebner, Julian Grady, Lan Nie, Todd Phillips, Eugene Davydov, Daniel Golovin, et al. Ad click prediction: a view from the trenches. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 1222–1230. ACM, 2013.

Tomáš Mikolov, Martin Karafiát, Lukáš Burget, Jan Černockỳ, and Sanjeev Khudanpur. Recurrent neural network based language model. In *Eleventh Annual Conference of the International Speech Communication Association*, 2010.

Vanessa Murdock, Massimiliano Ciaramita, and Vassilis Plachouras. A noisy-channel approach to contextual advertising. In *Proceedings of the 1st international workshop on Data mining and audience intelligence for advertising*, pages 21–27. ACM, 2007.

Graham Neubig. Neural machine translation and sequence-to-sequence models: A tutorial. *arXiv preprint arXiv:1703.01619*, 2017.

Hema Raghavan and Rukmini Iyer. Evaluating vector-space and probabilistic models for query to ad matching. In *SIGIR Workshop on Information Retrieval in Advertising (IRA)*, 2008.

Matthew Richardson, Ewa Dominowska, and Robert Ragno. Predicting clicks: estimating the click-through rate for new ads. In *Proceedings of the 16th international conference on World Wide Web*, pages 521–530. ACM, 2007.

Gerard Salton and Christopher Buckley. Term-weighting approaches in automatic text retrieval. *Information processing & management*, 24(5): 513–523, 1988.

Gerard Salton, Anita Wong, and Chung-Shu Yang. A vector space model for automatic indexing. *Communications of the ACM*, 18(11):613–620, 1975.

Benyah Shaparenko, Özgür Çetin, and Rukmini Iyer. Data-driven text features for sponsored search click prediction. In *Proceedings of the Third International Workshop on Data Mining and Audience Intelligence for Advertising*, pages 46–54. ACM, 2009.

Noah A Smith, Yejin Choi, Maarten Sap, Hannah Rashkin, and Emily Allaway. Event2mind: Commonsense inference on events, intents, and reactions. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 463–473, 2018.

Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 553–562. ACM, 2015.

Anh-Phuong Ta. Factorization machines with follow-the-regularized-leader for ctr prediction in display advertising. In *Big Data (Big Data), 2015 IEEE International Conference on*, pages 2889–2891. IEEE, 2015.

Ilya Trofimov, Anna Kornetova, and Valery Topinskiy. Using boosted trees for click-through rate

prediction for sponsored search. In *Proceedings of the Sixth International Workshop on Data Mining for Online Advertising and Internet Economy*, page 2. ACM, 2012.

Chenyan Xiong, Taifeng Wang, Wenkui Ding, Yidong Shen, and Tie-Yan Liu. Relational click prediction for sponsored search. In *Proceedings of the fifth ACM international conference on Web search and data mining*, pages 493–502. ACM, 2012.

Shuangfei Zhai, Keng-hao Chang, Ruofei Zhang, and Zhongfei Mark Zhang. Deepintent: Learning attentions for online advertising with recurrent neural networks. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1295–1304. ACM, 2016.

Weinan Zhang, Tianming Du, and Jun Wang. Deep learning over multi-field categorical data. In *European conference on information retrieval*, pages 45–57. Springer, 2016.

Yuyu Zhang, Hanjun Dai, Chang Xu, Jun Feng, Taifeng Wang, Jiang Bian, Bin Wang, and Tie-Yan Liu. Sequential click prediction for sponsored search with recurrent neural networks. In *Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

Yu Zhu, Hao Li, Yikang Liao, Beidou Wang, Ziyu Guan, Haifeng Liu, and Deng Cai. What to do next: Modeling user behaviors by time-lstm. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, pages 3602–3608, 2017.

Cantonese turn-initial minimal particles: annotation of discourse-interactional functions in dialog corpora

Andreas Liesenfeld
The Hong Kong Polytechnic University
Hung Hom, Kowloon, Hong Kong
lies0002@ntu.edu.sg

Abstract

This interactional linguistic study is concerned with the annotation of discourse-interactional functions of turn-initial particles in Cantonese conversation. These particles (or intersections) are commonly transcribed as *ngo* (哦), *ng* (嗯), *aa* (啊), *aak* (呃) and can format a range of functions both as turn-initial utterances or as stand-alone turns. Based on the analysis of 20 hours of naturally-occurring video corpus data, the study identifies five discourse-interactional functions that the most 'minimal' (i.e. shortest and mostly monophthongic) of these utterances can format: continuers, positive response tokens, change-of-state tokens, turn management tokens and repair initiators. I then show that three dimensions have to be taken into account to annotate those functions: sequential position, pitch contour and phonetic production format. In contrast to existing annotation taxonomies that directly map production format to function, I argue that discourse-interactional functions of these particles can only be annotated with reasonable accuracy if at least these three structural dimensions are taken into account. I conclude with discussing the relation between sequential position, sound and pitch format for each function.

1 Introduction

Turn-initial particles are short utterances such as *oh*, *huh* or *mmhm* in English that appear in turn-initial position and that can stand alone as turns. These particles have important functions in the joint construction of conversation and can constitute various actions depending on their sequential environment and production format. This study examines turn-initial particles in naturally-occurring Cantonese conversation, explicates some of their discourse-interactional functions, and examines the relation of their interactional uses to some aspects of their phonetic and prosodic production formats.

The utterances under study are a range of particles that occur in turn-initial or turn-constructional unit (TCU)-initial position and that are commonly described as particles or interjections (嘆詞) and transcribed using Chinese characters such as *ngo* (哦), *aa* (啊), *ng* (嗯), *aak* (呃). Specifically, the study focuses on the most 'minimal' turn-initial particles, those that are formatted using monophthongic and nasal utterances. It is not concerned with other 'larger' particles such as *ei* (誒), *ai yo* (唉唷) or *ji aa* (咦呀).

Figure 1 illustrates that such particles are not only a common phenomenon in naturally-occurring conversation, but also that these minimal utterances can format a range of different discourse-interactional functions depending on their sequential position and specific production formats. One of the speakers (P2) here produces three such particles in only a couple of seconds of talk that each format a different discourse-interactional function, a change-of-state token (Line 04), a continuer (Line 06) and a repair initiator (Line 08).

The accurate annotation of these particles is an integral part of any larger dialog act taxonomy that aims to further process speech act formation or model speaker intent. Notably, in Figure 1, each of the minimal utterances features a different produc-

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 471-479, Hakodate, Japan, September 13-15, 2019

Figure 1: Corpus excerpt showing three different discourse-interactional functions

Data excerpt (MYCANCOR 022 (04:53-05:06)) from a conversation between two participants, P1 and P2. Previous to the beginning of this excerpt, the topic of P1's partner was brought up and P2 inquired how P1 met her partner.

01 P2 form four ge3 si4 hau6 sik1(.) keoi5 tung4 nei5 tung4 hok6
form four嘅時候識佢同你同學
form four PAR time know 3SG with 2SG classmate
Got to know ((your partner)) in form four? Was he studying with you?

02 P1 m4 jat1 joeng6(.)ngo5 ngo5 dei2 hai6 jan1 wai4 tung4 jat1 go3 lou5 si1 maa3
唔一樣我我哋係因爲同一個老師嘛
NEG same 1SG 1PL be because same one CL teacher PAR
It wasn't like this, I, we were, because of the same teacher.

03 P1 go3 daa2 ngok6 tyun4 [>go3 lou5 si1<
個打樂團 個老師
CL band CL teacher
That band... teacher.

04⇒P2 [↑ngo::
哦
INT
Oh.

05 P1 jin4 zi1 hau6 ngo5 dei2 hok6 haau6 heoi3 keoi5 dei2 daai6 hok6 gaau1 lau4
然之後 我哋學校去佢哋大學交流
then 1PL school go 3PL university exchange
And then our school had an exchange with their university.

06⇒P2 ↓aa1.
啊
INT
Okay.

07 P1 zi1 hau6 hai6 go2 dou6 zau6 sik1 zo2 keoi5 lo3
之後係嗰度就識咗佢咯
then be there just know MOR 3SG PAR
Then I got to know him over there.

08⇒P2 ↑aa1 go2 jat1 baai3 ze1
啊[嗰一拜嗌
INT that one week PAR
Huh, in just a week?

09 P1 [hai6
係
is
Yes.

10 P1 |jat1 baai3
一拜
one week
A week.
|((P1 nods))

tion format. The change-of-state token in line 04 that displays a change to a state of knowing or understanding is formatted with [↑ngo::] (featuring a rising pitch contour). The continuer in line 06 is formatted as [↓aa1] (featuring a falling pitch contour), and the repair initiator in line 08 is also formatted with [↑aa1], this time featuring a rising pitch contour. Given this discrepancy in production format, can the discourse functions of these utterances be annotated by analysing their prosodic-phonetic form alone? I argue that, while production format is an important factor in the constitution of different functions, additional structural dimensions have to be taken into account to annotate these utterances accurately. Existing approaches to the functional annotation of these minimal yet important monophthongic and nasal utterances largely focus on analysing their production format, especially pitch contour, and propose to annotate functions directly mapped to specific production formats. Before discussing the results of the annotation efforts, I briefly review related work on particles and existing annotation taxonomies for Chinese.

1.1 Related work

Turn-initial particles are well-studied in both Mandarin and Cantonese and various reference grammars of spoken Chinese have described these utterances, referring to them as particles, interjections or non-lexical utterances (for Mandarin see, for instance, Chao (1965), Hu (1987) Li and Thompson (1989) and for Cantonese Killingley (1993), Cheung (2007), Matthews and Yip (2013)).

Only few studies, however, focus on turn-initial particles in naturally-occurring talk-in-interaction in particular. Studies coming out of interactional linguistics and conversation analysis have mainly focused on Mandarin Chinese, but nonetheless provide important insights in the work that turn-initial particles do in both languages. These studies have described a range of functions that turn-initial particles are involved in, mostly exploring a specific action that may be formatted using minimal particles.

Xudong (2008) examines continuers in Mandarin conversation and describes several uses of turn-initial particles under the topic of "listener responses". Oralova (2016) examines "minimal response tokens" in Mandarin and, focusing on *en* (嗯), shows that this particle can format continuers and positive response tokens. Also focusing on *en* (嗯) in turn-initial position, Xu (2009) describes "resumptive openers" in Mandarin.

Existing literature on repair in Mandarin also deals with turn-initial particles. Wu (2006) and Tseng (2013) show how various minimal particles including *en* (嗯) and *aa* (啊) can format repair initiators in Mandarin.

To the author's knowledge, no previous work on turn-initial particles in Cantonese has been done in the fields of interactional linguistics or conversation analysis.

Besides the above interactional linguistic studies that mainly explore the formation of a specific action in talk-in-interaction, a comprehensive annotation guideline for particles in Mandarin Chinese speech has been developed by Ping et al. (2014) (see also Lee et al. 2017). The study provides a detailed guide of how to annotate discourse-interactional functions of turn-initial particles based on the examination of production format and pitch contour. By mapping a specific format to one or more functions, particles can this way be annotated with reasonable accuracy. For instance, all three turn-initial particles previously shown in Figure 1 can be distinguished by examining their production format. However, this 'form-to-function mapping' approach is not well-suited to annotate the use of similar production formats that constitutes different functions in different sequential environments. Consider Figure 2 that illustrates this limitation.

Here two participants use a turn-initial particle of similar production format "*aa1*" (啊), but the utterance formats a positive response token in Data Excerpt 1 (Line 03) and a repair initiator in Data Excerpt 2 (Line 02). Notably, the participants do not treat this 'ambiguity' as problematic and follow up with a turn that displays action ascription respectively. This illustrates that, in order to annotate different functions of similar particles, another structural dimension has to be taken into account: sequential position.

2 Methodology

This study is situated in the field of interactional linguistics (Couper-Kuhlen and Selting, 2018) and

Figure 2: Example of use of 啊 (*aa1*) as positive response token and as repair initiator

```
Data excerpt (1): Positive response token 啊 (aa1)
MYCANCOR 021  (04:24-04:29)

01  P1   kei4(.)keoi5 hai6 mi1 ceot1 hoi1 ceot1 hoi1 kei4 taa1 dei6 fong1.
         其 佢係咪出開 出開其它地方
         actually 2SG is NEG go out go out other place
         Actu ((ally)), didn't she leave, leave ((that workplace)) for another place.

02  P2   >bin1 go3< ji3 mai5?
         邊個 薏米
         who person name
         Who, Barley.

03⇒P1    aa1:
          啊
          INT
          Uh.

04  P2   m4 zi1 ak1.
         唔知喔
         NEG know PAR
         I don't know.



Data excerpt (2): Repair initiator 啊 (aa1)
MYCANCOR 009  (00:45-00:52)

01  P1   nei5 dei2 zung6 jau5 cyun4 bin1 go3.
         你哋仲有傳邊個
         2SG also have convey which one
         Who else did you guys talk about?

02⇒P2    aa1:
          啊 (0.2)
          INT
          Huh?

03  P1   ceoi4 zo2 aa3 ciu1 zung6 jau5 cyun4 bin1 go3.
         除咗阿超仲有傳邊個
         despite Ah-Ciu also have convey which one
         Despite Ah-Ciu who else have you guys been talking about?

04  P2   mou5 aa1.
         冇啊
         not have PAR
         Nobody.
```

examines the use of minimal (monophthongic and nasal) utterances in turn-initial position as part of processes of action formation and ascription in talk-in-interaction (Levinson, 2013).

I present the results of a manual annotation of 20 hours of corpus data from a video corpus of naturally-occurring everyday talk. Around 484 instances of the production of turn-initial particles using monophthongic and nasal utterances were annotated in the corpus. The data was annotated according to interactional linguistic principles, employing the next-turn proof procedure to distinguish five discourse-interactional functions that participants commonly format using monophthongic (and nasal) utterances: continuers, positive response tokens, change-of-state tokens, turn-management tokens and repair initiators. All utterances were then transcribed using the International Phonetic Alphabet (IPA) for Cantonese (Zee, 1991) and a pitch contour analysis was conducted.

All data excerpts are from the 'MYCanCor' corpus of colloquial Malaysian Cantonese, a video corpus of 20 hours of naturally-occurring everyday conversation (Liesenfeld, 2018). The corpus data is transcribed in accordance with common practice in the field of interactional linguistics, using a four-line format consisting of Jyutping romanisation, Chinese characters including the Hong Kong Supplementary Character Set (HKSCS), word-by-word translation and English translation. This is a corpus of Cantonese Chinese as spoken in contemporary Malaysia. While their are differences between this variety of Cantonese and, for instance, Cantonese spoken in Hong Kong, the authors expect that the findings presented in this paper with regards to turn-initial particles are applicable across different Cantonese speech communities.

3 Results

The question that this study addresses is what discourse-interactional functions do participants format when uttering turn-initial minimal particles and what production formats do they commonly use to produce them. The aim is to examine the relationship between sequential position and the prosodic-phonetic properties of these utterances, and, by doing so, to contribute to a better understanding of how to annotate discourse-interactional functions of these particles in colloquial conversation.

For each of the five functions identified in the data set, I show an overview of their smoothed pitch contour and phonetic transcription based on IPA, and briefly discuss the relationship between the two properties.

3.1 Continuers

Continuers (or receipt tokens) are utterances that format a continuer action, i.e. that invite an interlocutor to go on talking. These utterances commonly appear at transition-relevance places (TRPs) and are free-standing, they usually do not constitute the beginning of a larger turn (Gardner 2001, Couper-Kuhlen and Selting 2018). Previous studies on this type of action in Mandarin Chinese have found that these utterances may also be produced to invite more talk and to format displays of information receipt and listener status (Oralova 2016, Gao 2007, Zheng 2007).

A closer look at the 318 instances of the use of monophthongic and nasal particles that format continuers in the corpus utterances shows that these utterances commonly feature a constant or falling pitch contour, and that [ɐ] is the most frequently used phonetic format (Figure 3). Notably, [a], [ɔ] and [m] as well as other formats are also used to format the action.

3.2 Positive response tokens

Positive response tokens (also affirmative tokens) are utterances that, in contrast to continuers, not only display information receipt and listener status, but also constitute a display of affirmation or agreement Couper-Kuhlen and Selting (2018). In Cantonese, monophthongic utterances such as *aa1* (啊) or *ng2* (嗯) can constitute both continuers and affirmative tokens, i.e. these utterances can format interactionally complete affirmative responses.

Figure 4 shows the 41 instances of this use in the data set. Positive response tokens are commonly formatted with a constant or falling pitch contour, using [œ], [ɐ] and [a]. In contrast to continuers, the use of nasal utterances was not observed.

3.3 Change-of-state tokens

Change-of-state tokens format displays that the speaker has moved from a position of unknowing

Figure 3: Continuers: smoothed pitch contour and IPA production format; total n=318

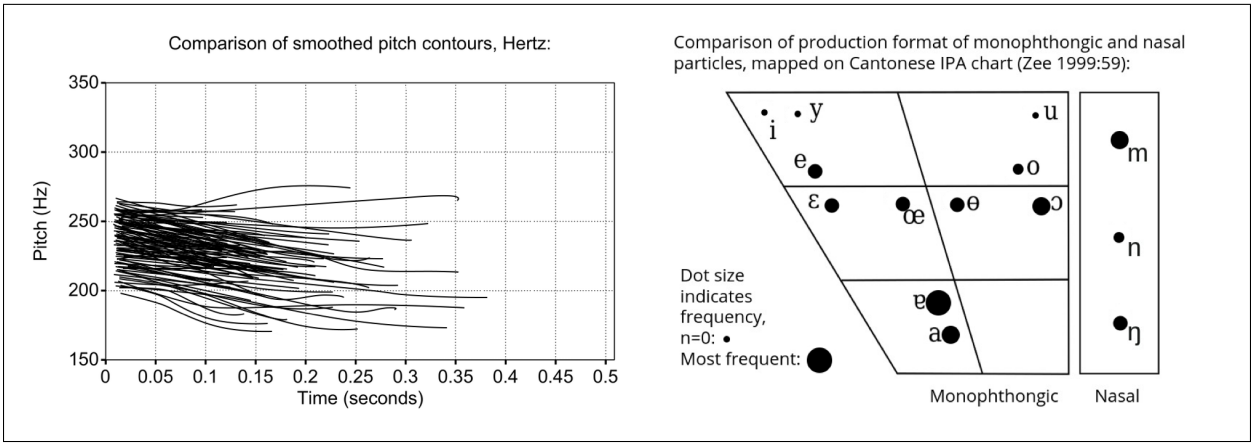
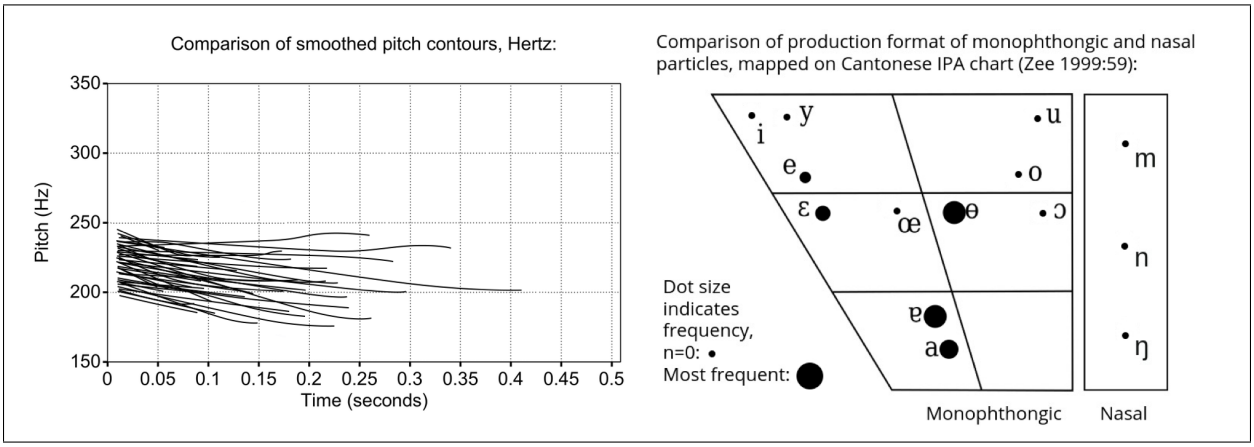


Figure 4: Positive response tokens: smoothed pitch contour and IPA production format; total n=41



to a claimed state of knowing, i.e. they format actions that display understanding or insight (Heritage 1984, 2012). Figure 5 shows 30 instances of the use of monophthonic utterances to format this action. Change-of-state tokens commonly feature a rising pitch contour, and [œ], [o] and [u] are used most frequently in the data set.

3.4 Turn management tokens

Turn management tokens (also turn uptake or turn stalling tokens) format displays of hesitation, reluctance or word search. In the data set (n=68) these utterances commonly feature a constant or falling pitch contour and are formatted using a range of phonetic formats [ɐ],[œ],[θ],[o],[u] and [ɔ]. Notably, neither pitch nor phonetic format appears to be a distinctive feature here, indicating that these actions may be routinely formed by relying on other (possibly sequential) properties.

3.5 Repair initiators

Repair initiators (also trouble tokens or troublesome hearing tokens) are utterances that format displays of a troublesome hearing, doubt or surprise. In the data set (n=27) these actions are commonly formatted featuring a rising pitch contour using [a] or [ɐ].

4 Discussion

Five discourse-interactional functions that can be formatted using minimal (monophthongic and nasal) utterances in turn-initial position have been identified. Based on the analysis of sequential position and

Figure 5: Change-of-state tokens: smoothed pitch contour and IPA production format; total n=30

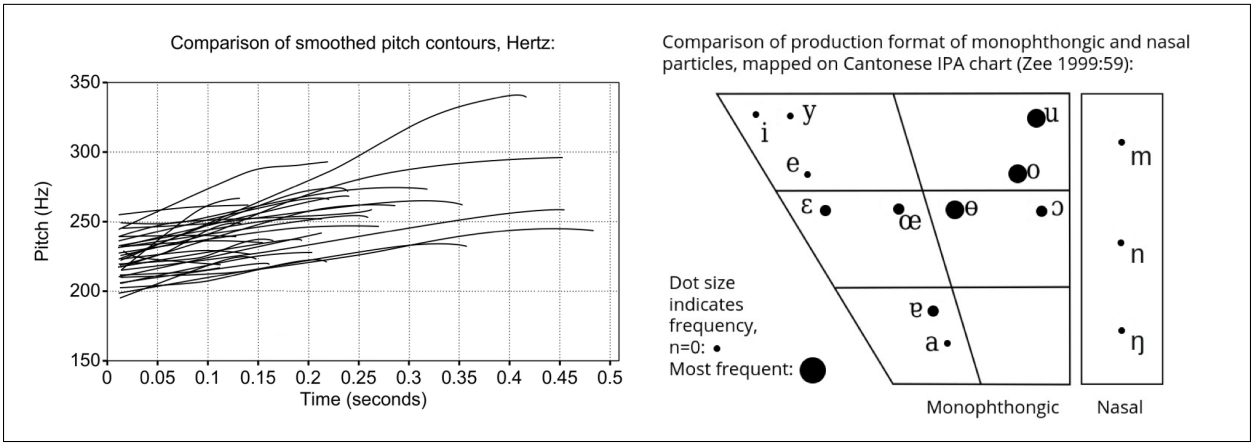


Figure 6: Turn management tokens: smoothed pitch contour and IPA production format; total n=68

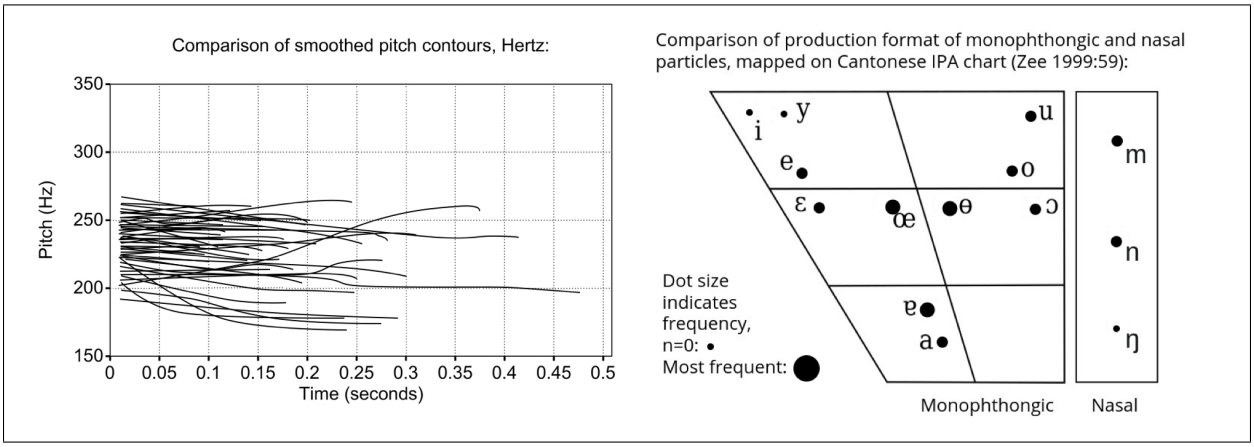
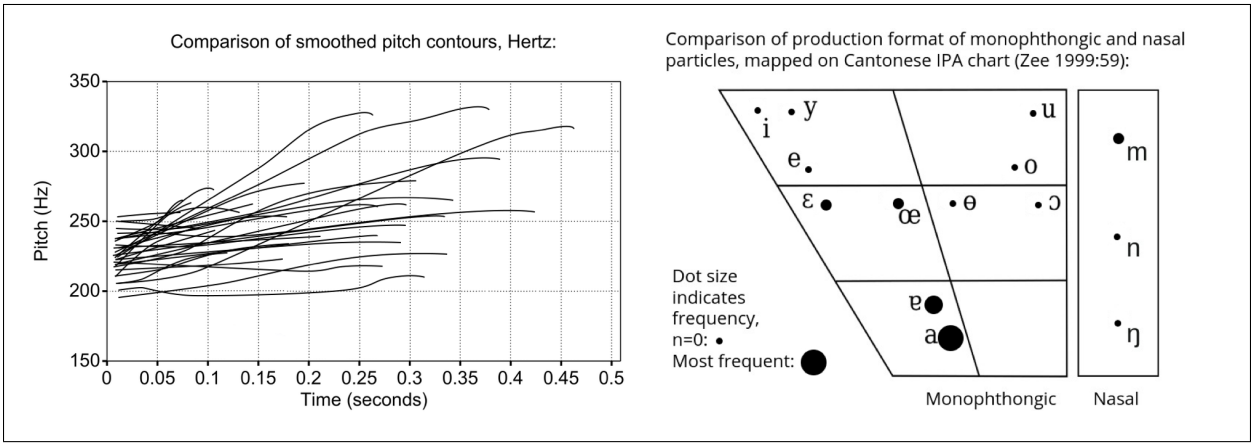


Figure 7: Repair initiators: smoothed pitch contour and IPA production format; total n=27



production format of these utterances in 20 hours of corpus data, I show that the participants format each function using a range of prosodic-phonetic formats that varies in scope, some more constrained than others. I provide an overview of the functions that these utterances can constitute that draws on previous work in interactional linguistics (e.g. Couper-Kuhlen and Selting 2018) and that, in contrast to existing annotation guidelines, is directly grounded in naturally-occurring data. I present an overview of a discourse-functional annotation of 484 usage instances and, focusing on pitch contour and phonetic format, examine the in-situ relationship between action formation and utterance format (Levinson, 2013):

(1) Continuers are produced using a relatively wide range of pitch and phonetic formats, ranging from falling to constant pitch contour and including a range of phonetic formats, with [ɐ] being the most frequent. Continuers are also the most frequent utterance in the data set, making up around 65% of all annotated particles.

(2) Positive response tokens also exhibit constant and falling pitch contours but are more constrained to fewer phonetic formats ([ɐ],[a] and [θ]).

(3) Change-of-state tokens appear to commonly feature a rising pitch contour and are relatively constrained to [θ], [o] and [u].

(4) Turn management tokens feature constant and falling pitch contours but are less constrained in terms of phonetic format, they cover a relatively large range of vocalic and nasal utterance formats.

(5) Repair initiators commonly feature a rising pitch contour and are relatively constrained to [a] and [ɐ].

5 Conclusion

Based on the analysis of recordings of real-world everyday talk-in-interaction, I show that minimal turn-initial utterances of a similar production format can constitute different discourse-interactional functions in different usage environments - a crucial limitation of annotation approaches that rely on direct form-to-function mapping. I conclude that, in order to annotate discourse-interactional functions of minimal turn-initial utterances with reasonable accuracy, at least three structural dimensions have to be taken into account: sequential position, pitch contour and (phonetic) production format. If only pitch contour and production format are considered, good results can be achieved for some functions that appear to be more constrained in their format (such as change-of-state tokens and repair initiators). Other functions, however, appear to not feature strong distinctive prosodic-phonetic properties, which requires their sequential position to be taken into account in order to accurately annotate their respective function (such as turn-management tokens and continuers). The data set shows that participants produce different functions by jointly relying on a range of structural dimensions that are (at least) both sequential and prosodic-phonetic in nature, and that differ in scope for each discourse-interactional function. I hope that this preliminary study provides a useful starting point for further explorations of minimal particles and their involvement in the intricate processes of formation and ascription that participants routinely rely on in natural conversation.

References

Chao, Y. R. (1965). *A grammar of spoken Chinese*. University of California Press.

Cheung, S. H. N. (2007). *A Grammar of Cantonese Spoken in Hong Kong*. Chinese University Press.

Couper-Kuhlen, E. and Selting, M. (2018). *Interactional linguistics*. Cambridge University Press.

Gao, L. (2007). Cong gongneng yuyanxue jiaodu tanxi “en” de shengdiao moshi yu qibiao yi gongneng zhi guanlian (Exploring the tone pattern of ‘en’ and the relationship with its ideographic function from functional linguistic perspectives). *Anhui Literature*, (10):158–159.

Gardner, R. (2001). *When listeners talk: Response tokens and listener stance*, volume 92. John Benjamins Publishing.

Heritage, J. (1984). A change-of-state token and aspects of its sequential placement. *Structures of Social Action: Studies in Conversation Analysis*, pages 299–345.

Heritage, J. (2012). The epistemic engine: Sequence organization and territories of knowl-

edge. *Research on Language & Social Interaction*, 45(1):30–52.

Hu, M. (1987). *An Exploration of Beijing Mandarin (*北京话初探*)*. The Commercial Press (商务印书馆).

Killingley, S.-Y. (1993). *Cantonese*, volume 6. Lincom Europa.

Lee, J. W., Tao, H., and Lu, P. (2017). Transcribing Mandarin Chinese Conversation: Linguistic and Prosodic Issues. *Asia-pacific Journal of Multimedia Services Convergent with Art, Humanities, and Sociology (*예술인문사회융합멀티미디어논문지*)*, 7:787–799.

Levinson, S. C. (2013). Action formation and ascription. In *The handbook of conversation analysis*, pages 103–130. Wiley-Blackwell.

Li, C. N. and Thompson, S. A. (1989). *Mandarin Chinese: A functional reference grammar*. Univ of California Press.

Liesenfeld, A. (2018). MYCanCor: A Video Corpus of spoken Malaysian Cantonese. In *11th Edition of the Language Resources and Evaluation Conference (LREC), 7-12 May 2018, Miyazaki, Japan*.

Matthews, S. and Yip, V. (2013). *Cantonese: A comprehensive grammar*. Routledge.

Oralova, G. (2016). *Minimal Response Token en in Mandarin Conversation*. PhD thesis, University of Alberta.

Ping, L., Won, L. J., and Hongyin, T. (2014). Discourse Properties of Some Special Sound Elements and Their Transcription Treatment (现代汉语口语中特殊话语语音成分的转写研究). *Linguistic Sciences (*语言科学*)*, 13(2):113–130.

Tseng, S.-C. (2013). Lexical coverage in Taiwan Mandarin conversation. *International Journal of Computational Linguistics & Chinese Language Processing, Volume 18, Number 1, March 2013*, 18(1).

Wu, R.-J. R. (2006). Initiating repair and beyond: The use of two repeat-formatted repair initiations in Mandarin conversation. *Discourse Processes*, 41(1):67–109.

Xu, J. (2009). *Displaying overt recipiency: Reactive tokens in Mandarin task-oriented conversation*. PhD thesis, University of Nottingham.

Xudong, D. (2008). The use of listener responses in Mandarin Chinese and Australian English conversations. *Pragmatics. Quarterly Publication of the International Pragmatics Association (IPrA)*, 18(2):303–328.

Zee, E. (1991). Chinese (Hong Kong Cantonese). *Journal of the International Phonetic Association*, 21(1):46–48.

Zheng, Y. (2007). "En" de huayu gongneng fenxi (An analysis of the discourse function of "en"). *Nanfang lunkan*, (10):56–57.

Are TERRORISM and *kongbu zhuyi* translation equivalents? A corpus-based investigation of meaning, structure and alternative translations

Lily Lim
School of Languages and Translation
Macao Polytechnic Institute
llim@ipm.edu.mo

Abstract

This paper reports an investigation using large scale corpora to contrast a pair of translation equivalents – TERRORISM in English and 恐怖主义 *kǒngbù zhǔyì* in Chinese. Close similarities between the two words manifested in the lexical profile produced by Word Sketch, e.g., in terms of their top collocates and syntactic roles. However, we also observed notable differences between the two words – e.g., 恐怖主义 occurs far more frequently than TERRORISM in noun-noun constructions, in particular in the 'X+noun' construction (X=恐怖主义/TERRORISM). Based on evidence from the corpora, 恐怖主义 entails a relatively narrower range of semantic meaning than that of TERRORISM, and is more readily joined by another noun to convey more specific meaning. Given that the two words are not translation equivalents in certain situations, we identified a number of methods that effectively retrieved several lexical candidates from comparable corpora for alternative translations in these situations.

1 Introduction

The use of the word TERRORISM has been rapidly rising from the 1970s, and this probably reflects the social issues we are facing in the modern time. Authoritative English-Chinese dictionaries give the term 恐怖主义 *kǒngbù zhǔyì* 'terror ism' as an explanation of TERRORISM as well as its equivalent item in the Chinese language (e.g., online Cambridge English-Chinese dictionary, https://dictionary.cambridge.org/us/dictionary/english-chinese-simplified/terrorism). Indeed, the two words – i.e., TERRORISM and 恐怖主义 – appear to be good translation equivalents in the two languages. Examples of the words with the suffixes –ism and –主义 *zhǔyì* can be easily accessible, e.g., 'capitalism' and 资本主义 *zīběn zhǔyì* 'capital ism', 'socialism' and 社会主义 *shèhuì zhǔyì* 'society ism', and 'nationalism' and 民族主义 *mínzú zhǔyì* 'ethnic nationalities ism'. However, a closer look at the –ism words and their corresponding –主义 words does suggest some difference between the two. For example, TERRORISM in English can refer to the ideology of using terror to attain goals, the acts or means by which people bring about terror, and the organizations that devise or carry out terrorist attacks. By contrast, 恐怖主义 in Chinese mainly refers to terrorist ideology, and does not effectively convey the meaning expressed by TERRORISM in certain translation situations. In examples (1) to (4), which were retrieved from online English-Chinese parallel corpora (e.g., BCC 北語雙語語料庫: http://bcc.blcu.edu.cn/lang/bi), the instances of TERRORISM denote terrorist acts or organizations, and were translated into various expressions other than 恐怖主义 (underlines added in examples):

(1a) The exact suite of technologies in PROTECT, which stands for Program for Response Operations and Technology Enhancements for Chemical/Biological Terrorism, is not made public.

(1b) 保护系统（即生化恐怖袭击的应对操作和技术强化方案）的确切技术套件尚未公诸于世。

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 480-487, Hakodate, Japan, September 13-15, 2019

'terror(ist) attack/s'
(gloss translation added for the underlined words in Chinese)

(2a) He said he has no ties to terrorism.
(2b) 他自称和恐怖组织并无关连。
'terror(ist) organization/s'

(3a) We have liberated the whole country from LTTE terrorism.
(3b) 我们把整个国家从'猛虎'恐怖组织中解放了出来。
'terror(ist) organization'

(4a) E-mail terrorism
(4b) 利用电子邮件进行的恐怖活动
'terror(ist) activities'

The Chinese sentences in (1b) to (4b) would sound either awkward or imprecise if 恐怖主义 were used in lieu of the underlined words. The examples strongly suggest that there are situations in which 恐怖主义 does not serve as a good translation equivalent of TERRORISM. At this juncture, large scale English and Chinese corpora potentially provide evidence on the differences between the two words in terms of their semantic meaning and the syntactic structures they tend to take part, and may contain lexical items that can serve as alternative translations.

This study utilizes Sketch Engine (https://www.sketchengine.eu/) with the large scale corpora of both English and Chinese, and also some major web-based English-Chinese translation databases to answer the following research questions:

a) To what extent are TERRORISM and 恐怖主义 different from each other in their lexical profiles in terms of the semantic meaning and grammatical structures they construct?
b) In what circumstances, if any, is 恐怖主义 no good translation for TERRORISM?
c) When 恐怖主义 does not translate TERRORISM well, what are alternative translations?

We will use the functions such as Word Sketch, Concordance and Thesaurus in Sketch Engine (SkE) to gather information for answering the research questions (see 3.1 and 3.2), and devise specific methods to identify alternative translations (see 3.3). Two large-scale monolingual corpora accessible in SkE are selected for this study – i.e., enTenTen15 that consists of 15.7 billion words from English web 2015, with advanced genre classification and sophisticated spam removal, as the corpus for the English language (using Penn TreeBank part of speech tagset), and zhTenTen11 with 1.7 billion words crawled in 2011 mainly from the web of the Chinese Mainland for the Chinese language (using Chinese Penn Treebank standard models, Stanford Log-linear Part-of-Speech Tagger).

2 Trends and dictionary definitions

The word TERRORISM exhibited a sharp rise from the 1970s, according to Google Books Ngram Viewer (https://books.google.com/ngrams). A number of social events probably contributed to this – e.g., the Israeli-Palestinian conflict in the 1970s, the attacks on 11 September 2001 and the 2002 Bali bombings. The surge of 恐怖主义 in the Chinese language came around the late 1990s, following the rise of TERRORISM in English (see Figures 1 and 2).

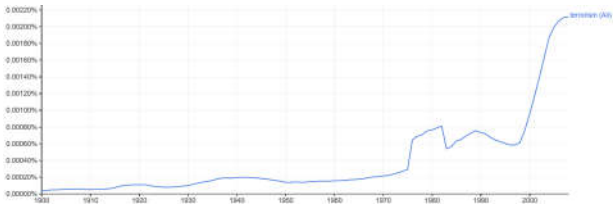


Figure 1: The frequency of TERRORISM in English: Google Books Ngram Viewer (1990-2008)

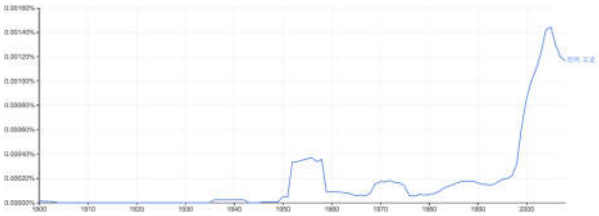


Figure 2: The frequency of 恐怖主义 in Chinese: Google Books Ngram Viewer (1990-2008)

TERRORISM is defined by English Oxford Living Dictionaries (https://en.oxforddictionaries.com/) as:

> Noun [mass noun]. The unlawful use of violence and intimidation, especially against civilians, in the pursuit of political aims.

A similar definition for 恐怖主义 is given by the authoritative 现代汉语词典 'The Dictionary of Modern Chinese' (2005: 781):

> 恐怖主义：蓄意通过暴力手段（如制造爆炸事件、劫持飞机、绑架等）造成平民和非战斗人员伤亡和财产损失，以达到某种政治目的的行为和主张。(underlines added for emphasis)
>
> *Kongbu zhuyi*: the acts or ideologies that rely on deliberate use of violent means (e.g., carrying out explosion, aircraft hijacking, kidnapping and so on) to cause the casualties of civilians and non-combatants and the loss of property, in order to attain certain political aims (English translation by the investigator)

Both definitions point out the use of violence on civilians for achieving political aims. The Chinese definition also specifies that 恐怖主义 entails both 行为 'the act or deed' and 主张 'ideology' or 'proposition'. While the Chinese definition gives a very reasonable explanation of what 恐怖主义 is about in social lives, large-scale language corpora would demonstrate the actual use of the word in context, providing rich information about the syntactic structures in which it occurs and the semantic meanings it conveys, which we will examine in the following section.

3 The results

In this section, we investigate the evidence from large scale language corpora – i.e., enTenTen15 for English and zhTenTen11 for Chinese – to contrast TERRORISM and 恐怖主义 both semantically and syntactically.

3.1 Similarities in Word Sketch results

We first produced the lexical profile of both TERRORISM and 恐怖主义 using the Word Sketch function (https://app.sketchengine.eu/#wordsketch) in SkE. The profiles reveal predominant similarities between the two words in terms of their lexical collocates and grammatical relations (namely 'gramrel' in SkE). In terms of the words occurring in the "and/or" positions in relation to the two words, items such as 'extremism', 'crime', 'violence' and their Chinese equivalents are ranked top. Both TERRORISM and 恐怖主义 tend to modify nouns such as 'act', 'offences', 'threat, 'financing' in English, and, similarly, 行为 'behavior', 犯罪 'crime', 威胁 'threat', 融资 'financing' in Chinese. In addition, both TERRORISM and 恐怖主义 occur as the object of verbs such as 'combat', 'counter', 'defeat, 'condemn' and 'eradicate' in English, and 打击 'combat', 反击 'fight back/counter', 胜过 'overcome', 谴责 'condemn', 根除 'eradicate' in Chinese, exhibiting close similarities between the two. We also observed remarkable similarities between the two words in terms of the verb predicates of the subject X (X=TERRORISM in English and 恐怖主义 in Chinese), the modifiers of X, and the X's Y structure.

3.2 Differences in X+noun construction and its alternative structures

Apart from the similarities, we noted a conspicuous difference between the two words in terms of their tendency to take part in the 'X + NP' structure. In terms of lineal syntactic structure, 恐怖主义 in Chinese is very frequently followed by a noun to form a noun-noun construction, e.g., 恐怖主义犯罪 'terrorism crime', 恐怖主义活动 'terrorism activity', 恐怖主义威胁 'terrorism threat', 恐怖主义行为 'terrorism behavior'. This X+noun construction accounts for 31.4% of all the occurrences of 恐怖主义 (n=5,067) in zhTenTen11, nearly triple the frequency (11.6%) of X+noun construction of all the occurrences of TERRORISM (n=435,996) in enTenTen15 (X = TERRORISM / 恐怖主义: see Table 1). Examples of this type in English are such as 'terrorism act', 'terrorism financing', 'terrorism charges'.

A closer look at the top NPs occurring in the 'X+noun' construction in enTenTen15 and zhTenTen11 reveals that 恐怖主义 is most frequently joined by NPs such as 活动 'activity' (3.91%), 袭击 'attack' (1.87%) and 犯罪 'crime' (4.91%), at frequencies far higher than those similar nouns joining TERRORISM in English, e.g., 'activities' (0.07%), 'attacks' (0.04%) and 'offences' (0.15%). Our intuition as native speakers of Chinese suggests that 恐怖主义

mainly denotes ideology or way of thinking, while TERRORISM entails both ideology and practice (i.e., the acts and deeds). If 恐怖主义 is narrower than TERRORISM in terms of semantic meaning, then joining another noun enables 恐怖主义 to extend the range of meaning so as to denote the acts or deeds.

	n=	%	example
English			
X+noun	50,429	11.6	*terrorism* act
noun+of+X	64,090	14.7	act of *terrorism*
X's+noun	109	0.03	*terrorism*'s root
total	114,628	26.3	
Chinese			
X+noun	1,589	31.4	恐怖主义犯罪 'terrorism crime'
X+的+noun	569	11.2	恐怖主义的威胁 'terrorism *de* threat'
total	2,158	42.6	

Note: X stands for TERRORISM in English and 恐怖主义 in Chinese

Table 1: X+noun construction and its alternative structures in English and Chinese

This hypothesis is largely supported by the evidence from zhTenTen11. Instances such as 恐怖主义活动猖獗 'terrorism activities (are) rampant' and 恐怖主义势力猖獗 'terrorism power (is) rampant' frequently occur in Chinese, and these complex NPs would sound rather redundant if translated word for word into English. More succinct expressions are much more preferred in English, according to our data, e.g., 'terrorism is rampant', rather than 'terrorism/terrorist activities are rampant'. No instance of 'terrorism acts/deeds are/were […] rampant' occurs in enTenTen15, while there is only one instance of 'the acts of terrorism were rampant'. Our data showed that the word TERRORISM alone clearly entails the meaning of the acts or deeds in English.

Although X+noun construction – e.g., 恐怖主义活动猖獗 'terrorism activities (are) rampant', 恐怖主义势力猖獗 'terrorism power (is) rampant' – frequently occurs in zhTenTen11, there are also many instances in which 恐怖主义 stands alone as the NP, e.g., in 恐怖主义(日益/尤爲)猖獗 'terrorism (increasingly/particularly) rampant'. Using 恐怖主义 without a noun following it tends to be interpreted as referring to ideology, e.g., 恐怖主义思维(模式)/思潮猖獗 'terrorism thinking (mode)/trend of thoughts rampant'. Having said this, we also found a few instances in which 恐怖主义 is used to refer to the acts and deeds, e.g.,

(5) 9・11 恐怖袭击事件发生后，恐怖主义被国际社会视为针对全人类的严重犯罪，完善、加强打击恐怖主义立法成为世界范围的立法潮流 (from znufe.edu.cn).

After the event of terrorist attacks of 9.11, terrorism has been considered a serious crime against the whole humanity by the international community, and perfecting and strengthening anti-terrorism legislation has become a worldwide legislative trend (translation into English by the investigator)

Example (5) suggests that the meaning of 恐怖主义 extends to denote terrorist acts, a move of converging to the scope of meaning of TERRORISM in English.

Our finding that 恐怖主义 is more frequently joined by a noun to form complex NP than does TERRORISM will not be convincing if we overlook the alternative structures to the X+NP construction, e.g., NP+of+X in English and X+的+noun in Chinese. Table 1 includes such alternatives, with the total numbers tallied for both words in English and Chinese. The results were obtained from Corpus Query Language (CQL) queries under the Concordance tab, e.g., [tag="N.*"] [word="of"] [word="terrorism"] for the 'noun of X (terrorism)' structure.

The overall results indicate that 恐怖主义 (42.6%) occurs in these complex NPs 1.6 times as frequently as TERRORISM does (26.3%). Table 1 also shows that, to form complex NPs, TERRORISM tends to occur in the 'noun of X' structure more than in the 'X+noun' structure, while in Chinese, 恐怖主义 occurs predominantly in the 'X+noun' structure.

It is noteworthy that we obtained the key finding that 恐怖主义 constructs the complex NPs 1.6 times as frequently as TERRORISM does based on the percentages these NPs account for in the total

occurrences of 恐怖主义 and TERRORISM respectively. At this point, we need to be aware of fact that TERRORISM occurs 9.8 times as frequently in enTenTen15 (23.71 per million, n=435,996) as 恐怖主义 occurs (2.41 per million, n=5,067) in zhTenTen11. To confirm the relatively much lower frequency of 恐怖主义 compared to that of TERRORISM, we also investigated Chinese web 2017 (i.e., zhTenTen17, simplified Chinese), which is the most recent and largest Chinese corpus (13.5 billion words) accessible at SkE, in which 恐怖主义 occurs at a moderately increased frequency – i.e., 3.53 per million (n= 58,634). However, our key findings remain valid – i.e., (a) 恐怖主义 in Chinese still occurs far less frequently than TERRORISM does in English, and (b) when 恐怖主义 does occur, it exhibits a markedly stronger tendency to be complemented by a noun to form a complex NP than TERRORISM does.

The primary reason for TERRORISM being used much more frequently than 恐怖主义, we would argue, lies in the fact that TERRORISM is a rather versatile word that readily entails a wide range of meaning and domains – e.g., terrorist acts, behaviors and organizations – while 恐怖主义 is much less so. Another reason for the low frequency of 恐怖主义 has to do with the blending and shortening of the lexical chunks that contain 恐怖主义, since concise expressions tend to be much preferred in Chinese. For example, the shortened expressions such as 恐怖袭击 'terror attack' and the blended form 恐袭, are more commonly used than the full expression 恐怖主义袭击 'terror ism attack'. Similarly, 恐怖组织 'terror organization' is far more frequently used and sounds more idiomatic than the full expression 恐怖主义组织 'terror ism organization'. In general, expressions with the '恐怖主义+(的)+noun' construction tend to be shortened into '恐(怖)(主义)+noun', reducing 恐怖主义 to a lesser translation equivalent of TERRORISM in terms of frequency and versatility.

To sum up, we have observed differences between 恐怖主义 and TERRORISM in their tendencies to construct certain complex NPs, probably reflecting their different ranges of semantic meaning. Given the differences between the two words, we examine the occasions on which 恐怖主义 does not serve as good translation equivalent for TERRORISM, and what (alternative) translation/s can be used in 3.3.

3.3 Alternative translations

In our quest for alternative translations for TERRORISM, we first tried to identify the occasions in which the instances of TERRORISM are not translated into 恐怖主义. We searched for examples in major online English-Chinese translation databases. For example, at the portal of BCC (北語雙語語料庫), we queried the word 'terrorism' and searched through the instances of English-Chinese translation to identify examples similar to (1) to (4). These examples are crucial for highlighting the situations in which TERRORISM and 恐怖主义 do not stand as good translation equivalents to each other.

In (1), for example, 'chemical/biological terrorism' should better not be rendered as 生化恐怖主义 'bio-chemical terrorism' in Chinese, which would obscure the meaning. Since 生化恐怖主义 does not translate 'chemical/biological terrorism' well, we need to ask if there are other expressions that do. In other words, we are interested in discovering the alternative expressions that Chinese speakers would naturally use when they talk about 'chemical/biological terrorism' matters. Large scale Chinese corpora are reasonable resources that potentially contain such expressions. We selected zhTenTen11 at SkE, which stands as a comparable corpus to enTenTen13 in this study.

To construct the textual context about 'bio-chemical terrorism', we attempted the following CQL query on zhTenTen11 under the Concordance tab in SkE:

```
[word="生物"] []? [word="恐怖"]
```

The query returned 312 results, on which we performed KWIC search under the Frequency tab, revealing 生物恐怖 'biology terror' (n=305) as the predominant expression and five other much less frequently used expressions, e.g., 生物化学恐怖 'biology chemistry terror' (see Table 2).

Based on the Chinese collocations we have identified in Table 2 on the 'chemical/biological terror(ism)' matters, we further investigated the words occurring immediately to the right of the

KWIC (see Table 3), from which we removed the noise such as conjunctions and punctuations. Examining the Chinese expressions in Tables 2 and 3 together allows longer lexical chunks on the topic to emerge.

	KWIC (with gloss)	Frequency
1	**生物 恐怖** ‘biology terror’	**305**
2	生物 化学 恐怖 ‘biology chemistry terror’	4
3	生物 的 恐怖 ‘biological terror’	1
4	生物 武器 恐怖 ‘biology weapon terror’	1
5	生物 拥有 恐怖 ‘biology has terror’	1

Table 2: Query results of “生物” and “恐怖”

	Word (with gloss)	Frequency
1	**袭击 ‘attack’**	**109**
2	**事件 ‘event’**	**55**
3	活动 ‘activity’	9
4	防范 ‘prevention’	7
5	威胁 ‘threat’	6
6	制剂 ‘preparation’	5
7	材料 ‘material’	3
8	攻击 ‘attack’	3
9	因子 ‘factor’	3
10	分子 ‘agent’	3
11	防护 ‘protection’	2
12	行动 ‘action’	2

Table 3: The First word to the right of the KWIC of the query of “生物” and “恐怖”

We can see that the most commonly used expressions in Chinese on the topic include (cf. Tables 2 and 3) 生物(化学)恐怖袭击/事件/活动, ‘biology (chemistry) terror attack/event/activity’. These expressions are valuable lexical candidates for translating ‘chemical/biological terrorism’ into Chinese. On the other hand, it is noteworthy that 生物(化学)恐怖主义 ‘biology (chemistry) terror ism’ never occurred in zhTenTen11, strongly suggesting that it is not a likely translation equivalent of ‘chemical/biological terrorism’.

Given that 生物化学 ‘biology chemistry’ is often shortened into 生化 ‘bio-chem(istry)’ in Chinese, we performed a CQL (corpus query language) query on “生化” and “恐怖”:

```
[word="生化"] []? [word="恐怖"]
```

obtaining 21 results similar to those in Tables 2 and 3. Queries into the comparable corpus therefore led to the discovery of the frequently-used expressions in Chinese for discussing the subject matter.

Like the two CQL queries above, we discovered that querying “verb + noun (恐怖)” under the Concordance tab can be fruitful as well. For example, the CQL query

```
[word="打击"][]{1,2}[word="恐怖"]
```

returned 107 results. The KWIC list generated by this query and the list of the first word on the right of the KWIC contain translation alternatives that closely overlap with the expressions in Table 3.

Coming to back to the question of whether TERRORISM has other translation equivalents in Chinese apart from 恐怖主义, we sorted out a method for gathering potential candidates as follows. We first queried the synonyms and similar words of 恐怖主义 using the Thesaurus tab in SkE, and identified 极端主义 ‘extreme ism’ as the top synonym, which tends to be present in the co-text of 恐怖主义. We proceeded to run the following two CQL queries,

```
[word="极端主义"][] {1,3}[word="恐怖"]
[word="恐怖"] [] {1,3} [word="极端主义"]
```

using 极端主义 to construct the co-text in which expressions with 恐怖____ occur. The queries returned candidates including 恐怖活动/袭击/手法/组织 ‘terror activity/attack/means/organization’, which would potentially serve as alternative translations for TERRORISM.

4 Discussion and future studies

This study uses large scale corpora to contrast a pair of translation equivalents – TERRORISM and 恐

怖主义 – across languages, revealing similarities between the two words, and more importantly, pinpointing some fine differences between the two, semantically and syntactically.

Word-level (non-)equivalence has been a central topic in translation studies (Baker 2018: Chapter 2). In translation practice, translators also need to identify translation equivalents and be aware of the extent to which the equivalents entail the same meaning in context. The advancement in language corpora and concordancing techniques to date has brought unprecedented amount of materials and tools to translators and lexicographers, making available information about word meaning and usage far richer that than what conventional dictionaries can offer (cf. Section 2). From the results of the present study, we advocate the potential and value for translators and lexicographers to use corpus-based tools (e.g., SkE) for resolving translation problems and sorting out lexical puzzles, e.g., about near synonyms.

The present study echoes previous studies that use corpus tools to uncover unarticulated differences between near-synonyms (e.g., Chief et al. 2000; Tsai et al. 1998; Wang and Chu-Ren 2017; Xiao and McEnery 2006), and extends this line of study to examine translation equivalents across languages (cf. Li, Dong and Wang forthcoming).

This study shows that TERRORISM and 恐怖主义 are the closest translation equivalents to each other in a wide range of situations, based on evidence gathered by the sophisticated concordancing tools working on the large scale corpora in SkE. Both words occur in the structures such as the 'and/or', 'subject-verb', 'modifier-NP' constructions, in which they collocate with words of similar meaning between English and Chinese.

However, the most interesting and perhaps important findings of our investigation rest on the differences between the two words pinpointed by the corpus evidence. We noted from English-Chinese parallel corpora that TERRORISM is not translated into 恐怖主义 from time to time, suggesting that the semantic meanings of the two words do not totally overlap. We also found their difference in semantic meaning contributes to the grammatical relations in which they participate, e.g., 恐怖主义 is heavily present in the X+noun construction (cf. 3.2).

Finally, we were able to identify alternative translations of TERRORISM other than 恐怖主义. Using Corpus Query Language (CQL), translators/investigators can create new queries to track down potential candidates for translation equivalents through different paths (cf. 3.3). Corpus-based queries allow the investigator to form new hypotheses, test his/her language intuition, drawing on vivid examples and distribution patterns to attempt new findings.

In terms of the balance between (or the integration of) corpus-based evidence and the investigator's original thinking and intelligence, we hold that the latter should still play a central role. Modern concordancing platforms such as Sketch Engine provide very comprehensive, sophisticated (built-in) lexical profiling devices, generating a huge amount of (preliminary) analysis results. The staggering information can be extremely valuable, but, unfortunately, can also be overwhelming to (novice) investigators. We underscore the importance that investigators should exercise their agency, use their own language intuition and draw on their experience as translator (cf. Wang and Lim 2017), interpreter or lexicographer, to identify the key issues to pursue and sort out (new) paths of investigation meaningful to their professional practice or intellectual interests (cf. Lim 2019). In terms of methodology, triangulating the evidence from both comparable and parallel corpora was fruitful for the present study, and would be worth attempting in subsequent studies.

Acknowledgments

The author thanks the anonymous reviewers of this paper for their valuable comments and suggestions, and acknowledges research grant (RP-ESLT-01/2016) from the Macao Polytechnic Institute.

References

Baker, Mona. 2018. *In Other Words: A Coursebook on Translation*. London: Routledge.

Chief, Lian-Cheng, Huang, Chu-Ren, Chen, Keh-Jiann, Tsai, Mei-Chih and Chang, Li-Li. 2000. What can near synonyms tell us. *International Journal of Computational Linguistics & Chinese Language Processing: Special Issue on Chinese Verbal Semantics* 5(1): 47-60.

Li, Longxing, Dong, Sicong and Wang, Vincent X. forthcoming. Gaige and Reform: A Chinese-English Comparative Keywords Study (Ch.22). In Qi Su and Weidong Zhan (ed.) *From Minimal Contrast to Meaning Construct*. Springer/Peking University Press.

Lim, Lily. 2019. A corpus-based study of braised dishes in Chinese-English menus. In Stephen Politzer-Ahles, Yu-Yin Hsu, Chu-Ren Huang, Yao Yao (ed.) *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation: 25th Joint Workshop on Linguistics and Language Processing*. Hong Kong: Association for Computational Linguistics.

Tsai, Mei-Chih, Huang, Chu-Ren, Chen, Keh-Jiann and Ahrens, Kathleen. 1998. Towards a Representation of Verbal Semantics-An Approach Based on Near-Synonyms. Paper presented at the International Journal of Computational Linguistics & Chinese Language Processing, Volume 3, Number 1, February 1998: Special Issue on the 10th Research on Computational Linguistics International Conference.

Wang, Shan and Chu-Ren, Huang. 2017. Word sketch lexicography: new perspectives on lexicographic studies of Chinese near synonyms. *Lingua Sinica* 3(1): 1-22.

Wang, Vincent X. and Lim, Lily. 2017. How do translators use web resources? Evidence from the performance of English-Chinese translators. In Dorothy Kelly (ed.) *Human Issues in Translation Technology*. London: Routledge, pp. 63-79.

Xiao, Richard and McEnery, Tony. 2006. Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective. *Applied Linguistics* 27(1): 103-29.

中国社会科学院语言研究所词典编辑室. 2005. 现代汉语词典（第 5 版）. 北京市: 商务印书馆.

L1 and L2 Processing of Chinese Separable VO Compounds

Junghwan Maeng
Department of EALC
University of Illinois, Urbana-Champaign
U.S.A.
jmaeng3@illinois.edu

Abstract

This research explores the processing of Chinese separable VO compounds by L1 and L2 speakers. While Chinese separable VO compound verbs are categorized as words, they also occur as phrases via syntactic reanalysis. However, such syntactic reanalysis is not applicable to most VV compound verbs due to lack of syntactic relation between the two constituent morphemes. Given this, this research employs a lexical decision task to examine the underlying structure of VO compounds in L1 and L2 lexicon based on their response times for VO and VV compounds. The results of the analysis suggest that while both VO and VV compounds are processed as whole-word by L1 speakers due to word superiority effect in Chinese, L2 speakers seem to make distinctions between VO and VV compounds via decomposition as the effect of morpheme frequency was more pronounced for VO compounds.

1 Introduction

There has been an ongoing debate on whether the underlying structure of Chinese separable VO compounds should be identified as word or phrase due to its unique structure that licenses syntactic reanalysis. While Chinese separable VO compound verbs (帮忙 , *bangmang*, 'to help') are listed in dictionaries as words, they can also be analyzed into phrases as they allow insertion of a suffix on the head verbal morpheme or insertion of a modifier on the object nominal morpheme as in (1).

(1) Separable VO word
帮了 他 的 忙
Bang-le ta de mang.
to help-LE he-DE business
'to help him'

Except for highly lexicalized compounds such as *danxin* (担心, 'to worry') and *touzi* (投资, 'to invest), separable VO compounds cannot be followed by another nominal phrase because their argument structure is fulfilled internally by the thematic relation between the two constituent morphemes. As in (1), what is interpreted as the object of the compound is realized as a possessive nominal phrase to the second constituent, which is syntactically analyzed as a theme argument to the head verbal morpheme. With the second constituent functioning as a theme argument to the main predicate, having another object nominal phrase at the end of the compound would result in ungrammaticality because more than one constituent cannot occur in the same phrase as suggested by Phrase Structure Condition (Huang, 1984).

Regarding the identity of separable VO compounds, Huang (1984) favors the point of view of the underlying structure as phrases based on Phrase Structure Hypothesis. Because PSC assumes that verbs can be followed by one constituent at most, VO compounds should be identified as 'idiomatic phrases', and they are analyzed as words only when followed by a noun phrase. While Huang (1984) identifies VO compounds as phrases that are analyzed as words only when filtered out by PSC, Packard (2000) proposes a differing view which treats VO compounds as underlyingly words but reanalyzed as phrases when required by context. Packard (2000) claims that VO compounds are stored as words once entered into lexicon, and they are only being reanalyzed into phrases depending on syntactic and semantic requirements.

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 488-495, Hakodate, Japan, September 13-15, 2019

Unlike separable VO compounds, inseparable VV compound verbs are perceived clearly as words because separation of constituents cannot be licensed on them. Suffixation on VV is attached at the end of the entire compound, and the word order is relatively straightforward for VV compound words as they follow the canonical SVO sentence structure.

(2) VV word
帮助 他 了
Bangzhu *ta* *le.*
to help him LE
'to helped him'

If VO compounds are stored as phrases in L1 mental lexicon as suggested by Huang (1984), it can be assumed that VO and VV compounds may have different representations in L1 mental lexicon because the latter would be recognized strictly as words by native speakers. On the other hand, if the underlying structure of VO compounds is stored as word in the mental lexicon, it is expected that both VO and VV compounds would be stored as words, and hence would be processed in virtually the same pattern regardless of different internal structures of the compounds (VO vs VV).

The unique identity of Chinese VO compound verbs presents difficulties to L2 learners because most learners are not aware of syntactic constraints associated with VO compounds. Because they have insufficient knowledge with respect to the syntactic constraints of VO compounds, they often create ungrammatical sentences by placing another noun phrase after a separable VO compound (Yang & Han, 2016). The L2 error pattern is illustrated in (3). The main reason for the high error rates by L2 learners is that they are inclined to treat them exclusively as words and apply the general suffixation rule and word order as they would do for inseparable VV compounds.

(3) Frequent L2 error on VO
*我帮忙了他。
Wo bangmangle ta.
I help-ASP him

Shallow Structure Hypothesis (Clashen, 2006) claims that L2 learners rely more on semantic and lexical cues than syntactic information during sentence processing, but a similar tendency can also be observed in the L2 processing of structurally complex compound words. Clashen (2015) found that L2 learners are more dependent on semantic information than native speakers when dealing with morphologically complex compound words and prefer a simpler structural reading of the compounds. Combined with the frequent errors made by L2 learners on VO compounds, it can be assumed that L2 learners may tend to analyze VO compounds based on the meaning of the whole compound and choose to process them as structurally simpler forms, which are fully lexicalized inseparable compounds.

In order to provide empirical evidence regarding how VO compounds are processed by L1 and L2, a lexical decision task experiment was conducted to compare separable VO and inseparable VV compounds are processed by L1 and L2 speakers. This study first attempts to examine whether different compound types (separable VO vs inseparable VV) provide L1 and L2 speakers with different processing costs, and then the current study also delves into the detailed investigation of which factors has significant impact on the processing of VO compounds for L1 and L2 speakers.

2 Hypotheses and Research Questions

Following questions will be addressed in this paper based on the results from lexical decision task:

a) Do separable VO and inseparable VV compounds present different processing costs for L1 and L2 speakers?

b) Would separable VO and inseparable VV compounds be processed as decomposition or as whole word during lexical decision task?

For the first research question, L1 and L2 reading times for VO and VV compounds would be compared using a mixed random effect model with subject as a random effect. As for the second question, a mixed random effect model will be run individually on L1 and L2 reading times data to measure the effect of various factors on the processing of VO and VV compounds to examine whether decomposition occurs on each compound type for L1 and L2 speakers. VO and VV test items were selected from HSK (standardized Chinese proficiency test)

vocabulary list.

It is not yet clear how VO compounds are stored in L1 and L2 lexicon. As mentioned earlier, there are two possible scenarios as to how separable VO compounds are stored in L1 mental lexicon. The one explanation is that the underlying structure of VO compounds is stored as phrases in the mental lexicon based on Huang (1984), and the other view based on Packard (2000) proposes that VO compounds are stored as word by default in the lexicon. The current study proposes that if VO compounds are processed as underlyingly phrases, it is possible that L1 speakers may demonstrate different reactions for VO compounds from VV compounds because it may require native speakers to perform both syntactic and lexical readings of VO compounds while only lexical reading is needed for VV compounds. On the contrary, if VO compounds are processed strictly as words by L1 speakers, there should be no difference in reading time between VO and VV compounds during lexical decision task when all other variables are controlled because only lexical reading would be activated for both types of compounds.

Furthermore, representations of VO compounds in mental lexicon can be tested by whether they are decomposed or processed as whole words during lexical decision task. In order to test whether VO compounds are decomposed or processed as full-form, the effects of morpheme frequency and whole-word frequency on response time need to be examined. Previous studies on the processing of compound words suggest that while effects of morpheme frequency should be taken as evidence of decomposition, effects of the whole-word frequency should be interpreted as full-form access of compounds words (Andrews, Miller, & Rayner, 2004; Baayen, Dijkstra, & Schreuder, 1997; Taft & Forster, 1976).

If Chinese VO compounds are stored as phrases in lexicon as suggested by Huang (1984), it can be assumed that VO compounds may be more susceptible to decomposition if processed as phrases because L1 processors may opt for compositional route for which they have to retrieve meaning from each individual morpheme. If compositional processing is actually employed by L1 processors on VO compounds during lexical decision task, reading times on VO compound would be modulated mainly by individual morpheme frequency while reading times for VV compounds whose meaning is less transparent would be affected by the whole word frequency (hereby bigram frequency).

However, if VO compounds are processed strictly as words during lexical decision task, it is expected that decomposition may not be operated on VO compounds at the conceptual level because the whole word representation is enforced on VO compound so strongly that decomposition will be suppressed at this level. Given this, it is expected that whole word representations rather than decomposition would be activated or both VO and VV compounds at the conceptual level, and L1 reading times on the two types of compounds would be modulated mainly by bigram frequency during lexical decision task.

L2 speakers, on the other hand, are expected to prefer simpler reading of VO compound structure and rely heavily on available semantic cues as suggested by Clashen (2015). Given this, it is expected that L2 learners may rely more on the meaning of VO compounds and process them as inseparable compounds which are the most common compound type in Chinese. Hence, L2 learners are expected to demonstrate no difference in their reading times between VO and VV compounds when all other variables controlled. Furthermore, L2 reading times for both VO and VV compounds would be affected by bigram frequency rather than individual morpheme frequency because L2 speakers are expected to recall only whole word representations for both compound types.

3 Methods

In order to compare the effect of VO structure for the processing of Chinese words by L1 and L2 groups, a lexical decision task was conducted to measure the L1 and L2 participants' response times to VO and VV compound verbs. The task was administered using Paradigm software.

3.1 Participants

12 native speakers of Mandarin Chinese were recruited in Beijing as the control group. All L1 participants were 18 years or older. 24 L2 learners were also recruited from universities located in Beijing as the experiment group. All L2 participants were advanced level Chinese learners certified by HSK (Level 6: 21 participants, Level 5: 3 participants) to avoid the chance level accuracy rate. L1 backgrounds were various among the L2 group with 17

L1 Korean, 4 L1 Turkish, 1 Burmese and 2 L1 Arabic. All L2 participants are currently enrolled in an undergraduate or a graduate program in China. The mean age of the L2 participants were 27.79, and their total length of residence in the country was 7.5 years on average.

3.2 Materials

24 separable VO words and 24 inseparable VV words were selected from the HSK vocabulary list as test items with controlled bigram frequency, individual morpheme frequency, stroke number and neighbor size. Each pair of separable VO and VV compound verbs share the identical first constituent, but only differed in their second constituent. Following are the examples of test items.

(4) Test Items

VO	**VV**
离婚	离开
li-hun	*li-kai*
to leave-marriage	to leave-to open
'to divorce'	'to leave'

Based on the frequency measures from BCC Chinese Corpus, VV and VO items were controlled in terms of their bigram frequency, individual morpheme frequency and stroke numbers as well as neighborhood size. Two sample independent t-tests were conducted to ensure that all factors are controlled. The mean bigram frequency for VO items was 53,775.79 (SD = 89.957.74), and it was 47,799.96 (SD = 65,599.67) for VV items; $t(42.07) = 0.26$, $p = 0.79 > 0.05$). As for the individual morpheme frequency, only the frequency of the second constituents were considered because all test pairs have the same first constituent. As a result, the mean frequency of the individual morpheme for the VO compounds was 543,823.38 (SD = 808,011.9), and it was 437,275.71 (SD = 574,478.9) for the VV compounds; $t(41.52) = 0.53$, $p = 0.60 > 0.05$. In order to control for the visual complexity, stroke numbers were also measured for the second constituent of each test item. The mean stroke number of the VO items was 8.75 (SD = 2.82), and it was 8.29 (SD = 2.68) for the VV items; $t(45.89) = 0.58$, $p = 0.57 > 0.05$).

Furthermore, the lexical neighborhood size (the total number of words sharing either the first or the second constituent) was also controlled to prevent the number of lexical neighbors from affecting the processing of test items to a significant extent. Because each pair of test items shares the same initial constituent, only the number of words sharing the second constituent was controlled. The mean lexical neighborhood size for VO and VV compounds were 38.88 (SD = 27.00) and 30.88 (SD = 15.97), and the difference across compound types was not significant in a two sample independent t-test ($t(37.33) = 1.25$, $p = 0.22 > 0.05$).

48 nonwords were also included in a full list of items as distractors. Along with the nonwords, 24 real words were also included in the list. It was made sure that all real word fillers are nouns to avoid possible confounding effects that may be caused by the overlapping word class. With test items and nonwords combined, a full list of experiment items contains 120 items in total, which will be presented in 12 separate blocks in randomized orders. The VO and VV test items will allow us to examine whether syntactic processing of VO compounds has any significant impact on the processing of compounds for L1 and L2 by comparing their reading times to VV compounds. Furthermore, the effects of various factors such as bigram frequency, morpheme frequency and stroke number will be measured to examine whether VO and VV compounds are processed as whole words or decomposition by L1 and L2 speakers. Also, it should be noted that the effect of morpheme frequency only refers to the effect of the second constituent morpheme because VO and VV conditions share the identical first constituent morpheme.

3.3 Procedures

Participants performed a lexical decision task administered by Paradigm Software installed on a personal laptop at a closed space. Subjects were instructed to read a two-character Chinese compound on the screen and judge as fast as possible whether it is a legitimate Chinese word or not by pressing the left and right arrow keys representing real word and nonword respectively. For each trial, a fixation point ("##") was presented for 1000ms, and then a two-character Chinese compound appeared on the screen. Before the test session begins, participants completed a practice session with 8 sample items to become familiar with the task format. The entire test session lasted for less than 10 minutes for both L1 and L2 participants.

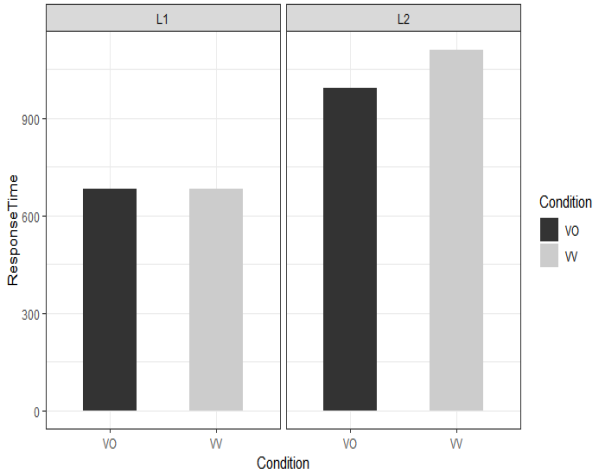
4 Results

Test items answered by L1 and L2 participants (1728 observations) were examined for the analysis. RT outliers were removed from each dataset if they were out of 2.5 SD from the mean RT. For the RT analysis, incorrect responses were further removed. As a result, 2.7% of data was excluded from the experiment dataset for the accuracy rates analysis, and only 5.4% (91 observations) of the accuracy rates data was further removed for the RT analysis. Both accuracy rates analysis and response times analysis were conducted using R.

4.1 Descriptive Analysis

First, the mean accuracy rates and the mean response times for the L1 group was calculated. The mean accuracy rates by the L1 group for VO and VV conditions were 99.30% (SD = 0.08) and 98.61% (SD = 0.12), and the mean response times for VO and VV compounds were 681.12ms (SD = 257.21) and 681.14ms (SD = 220.22). As for the L2 group, the mean accuracy rates for VO and VV compounds were 93.35% (SD = 0.25) and 91.29% (SD = 0.28), and the mean response times for VO and VV compounds were 991.76ms (SD = 422.04) and 1111.53ms (SD = 468.94) respectively. The descriptive data for the mean RT is plotted in Figure 1.

Figure 1: Descriptive RT Data



4.2 Inferential Statistics

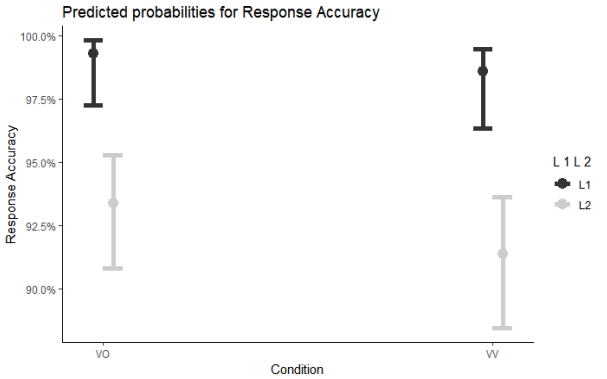
A binomial logistic regression with subject as a random effect was run by using the *glmer*() function of the *lme4* package to examine if the effects of compound type and L1/L2 status. Models with and without an interaction were compared based on the *anova*() function fo the *lme*4 package and their AIC value, and the one including the interaction was selected as best fitting model for the analysis. The results of the binomial logistic regression model are summarized in Table 1.

Table 1: Binomial Logistic Results for Accuracy Rates

	Estimate	SE	z value
Intercept	5.03 ***	0.72	6.98
VV condition	-0.71	0.87	-0.82
L2 Status	-2.35 **	0.73	-3.23
ConditionVV:L2	0.42	0.90	0.47
Signif. codes: 0 '***' , 0.001 '**', 0.01 '*', 0.05 '.', 0.1, ' ' 1			

The results revealed that the effects of L1/L2 status (Estimate = -2.35, SE = 0.73, z = -3.23, $p < 0.05$) was significant while the effect of condition (Estimate = -0.71, SE = 0.87, z = -0.82, $p > 0.05$) was not. However, no interaction of compound type and L1/L2 status (Estimate = 0.42, SE = 0.90, z = 0.47, $p > 0.05$) was found in the results. As in the plot in Figure 2, while there appears to be a noticeable difference between L1 and L2 groups, the effect of compound type was not significant for both participant groups.

Figure 2: Predicted Accuracy Rates



A mixed random effects model with subjects as a random effect ((1 | Subject)) was run to measure the effect of compound type and L1/L2 status for a broader population using the *lmer* () function of the *lmerTest* package. In order to select the best fitting model for the analysis, linear models with and

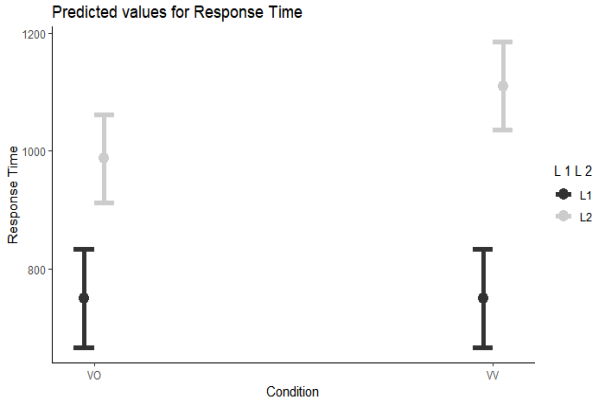
without an interaction of the two variables were compared based on the result of the *anova*() function in the *lme4* package and the r squared value. As a result, a model including the interaction of compound type and bigram frequency was selected for the analysis because it did not only yield a significant ANOVA result ($p < 0.05$) but also had a higher R^ value compared to the non-interaction model.

Table 2: Mixed Random Effects Model Results for Response Time

	Estimate	SE	Df	t-value
Intercept	750.95 ***	42.28	50.29	17.76
VV condition	0.363	29.04	1,558.61	0.01
L2	236.39 ***	30.73	1,522.20	7.69
Condition:L2	122.25 ***	36.24	1558.75	3.37
Signif. codes: 0 '***' , 0.001 '**', 0.01 '*', 0.05 '.', 0.1, ' ' 1				

The results of the mixed random effects model are summarized in Table 2. The mixed random effects model found a significant effect of L1/L2 status (Estimate = 236.39, SE = 30.73, df = 1,522.20, t = 7.69, $p < 0.001$) while the effect of compound type did not reach statistical significance (Estimate = 0.363, SE = 29.04, df = 1,558.61, t = 0.01, $p > 0.05$). The model predicts that L2 speakers' overall mean response times for the test items would be delayed by 236.39ms compared to the baseline condition. A significant interaction of compound type and L1/L2 status (Estimate = -1.228e-02, SE = 3.323e-04, df = 1,317, t = -4.37, $p < 0.001$) was also observed in the model.

Figure 3: Predicted Response Times



Furthermore, mixed random models with subject as a random effect were run on L1 and L2 RT data individually to examine the effect of following factors: Compound type, Bigram frequency, Morpheme frequency, lexical neighbor size and stroke number.

Table 3: Analysis of L1 RT

	Estimate	SE	Df	t-value
Intercept	911.45 ***	91.75	472.61	9.93
Condition:VV	-77.43	109.72	537.03	-0.71
Bigram Frequency	-53.92 ***	16.14	537.05	-3.34
Morpheme Frequency	-10.90	13.67	537.17	-0.80
Neighbor Size	0.11	0.34	537.05	0.31
Stroke	3.07	3.03	537.00	1.01
Condition*Bfreq	31.20	23.60	537.03	1.32
Condition*Stroke	-4.77	4.32	537.01	-1.10
Signif. codes: 0 '***' , 0.001 '**', 0.01 '*', 0.05 '.', 0.1, ' ' 1				

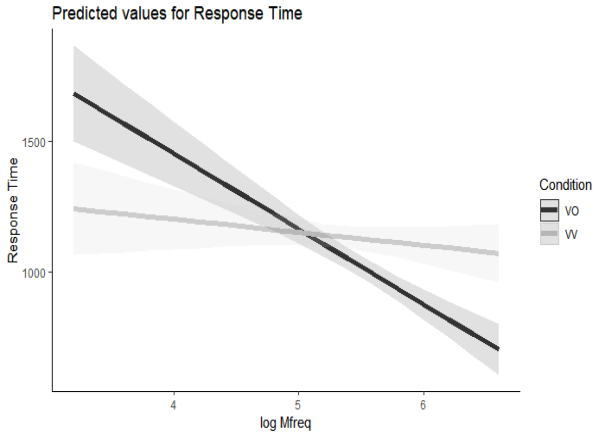
First, only the main effect of bigram frequency (Estimate = -53.92 SE = 16.14, df = 537.17, t = -3.34, $p < 0.001$) was found from the result of the L1 random effect model while the compound type, morpheme frequency, stroke number and neighbor size were not significant predictors for the L1 RT. Also, there was no interaction between different factors found from the L1 RT data analysis.

Table 4: Analysis of L2 RT

	Estimate	SE	Df	t-value
Intercept	2933.15 ***	261.18	1022.82	11.23
Condition:VV	-1118.68 ***	344.16	1004.68	-3.25
Bigram Frequency	-236.51 ***	38.98	1004.22	-6.10
Morpheme Frequency	-192.20	48.21	1004.84	-4.00
Neighbor Size	0.97	0.81	1004.52	1.20
Stroke	17.31	7.17	1004.37	2.41
Condition*Mfreq	171.91	57.14	1004.81	3.01
Condition*Bfreq	90.71	58.28	1004.23	1.56
Condition*Stroke	-10.68	10.32	1004.23	-1.04
Signif. codes: 0 '***' , 0.001 '**', 0.01 '*', 0.05 '.', 0.1, ' ' 1				

Unlike the L1 data, various effects were found from the L2 random effect model on the RT data. In the L2 RT data, main effects of the compound type (Estimate = -1118.68, SE = 344.16, df = 1,004.68, t = -3.25, p = 0.001), bigram frequency (Estimate = -236.51, SE = 38.98, df = 1,004.22, t = -6.10, p < 0.001), morpheme frequency (Estimate = -192.20, SE = 48.21, df = 1,004.84, t = -4.00, p < 0.001) and stroke number (Estimate = 17.31, SE = 7.17, df = 1,004.37, t = 2.41, p < 0.05) were found. Furthermore, there was a significant interaction of compound type and morpheme frequency (Estimate = 171.91, SE = 57.41, df = 1,004.81, t = 3.01, p < 0.01) meaning that the effect of morpheme frequency became more pronounced when the compound type changed from the VO to the VV structure as shown in the figure below.

Figure 4: Interaction of condition and morpheme frequency for L2 RT



The interaction graph shows that while the effect of morpheme frequency has little influence on the response times for VV compounds, the increase in the morpheme frequency facilitates the L2 processing of VO compound to a significant degree.

5 Discussion

The results of accuracy rates analysis found no interaction of compound type and L1/L2 status as the effect of compound type was not significant for both L1 and L2 groups while a significant between-group difference was observed in the overall mean accuracy rates. The lack of compound type effect for both experiment group suggests that VO and VV test items were processed with equal amount of difficulty for L1 and L2 participants. Based on the accuracy rates analysis, it is difficult to tell which compound type presented more processing cost for each participant group.

Contrary to the prediction, the result of the RT data analysis shows that the effect of compound types was significant for L2 speakers while it was virtually absent in the L1 RT data. The absence of the compound type effect from the L1 RT data seems to provide supporting evidence that VO compounds are stored exclusively as words in the L1 mental lexicon because both VO and VV compounds were answered at virtually the same speed. Also, the results of the mixed random model on the L1 RT data provides further supporting evidence for VO compounds as words because the reading times of VO and VV compounds were modulated by bigram frequency while the effects of morpheme frequency were not found in the analysis. The result of L1 RT analysis also provides supporting evidence for the word superiority effect in Chinese (Mattingly & Xu, 1994; Shen & Li, 2012), in which the whole word frequency was found to be predominant independently of morpheme frequency. The fact that L1 processors relied less on morpheme frequency in the processing of both VO and VV compounds in the current study lends support for the lexical processing of VO compounds suggested by Packard (2000).

However, it is still questionable as to why L2 speakers were faster to process the VO compounds than the VV compounds considering that nonnative speakers are assumed to be less sensitive to the difference in the internal structures of the compounds due to insufficient L2 knowledge. While L1 response times provided evidence for the word superiority effect, L2 speakers failed to demonstrate the same effect as they engaged in more analytic reading of compound words written in a less familiar language. One possible explanation is that L2 speakers may have been more sensitive to differences in semantic properties between nominal and verbal morphemes contained in VO and VV compounds. While word superiority effect prevents L1 speakers from further analysis of individual constituents of VO and VV compounds, L2 speakers may have made distinctions between VO and VV compound based on the decompositional processing. Previous studies on the processing of different parts of speech have reported that nouns may be processed with more ease compared to verbs due to

factors such as higher imaginability and simple meaning structure (Cordier et al., 2013; Kauschke and Stenneken 2008; Sereno 1999). Given this, it is possible that L2 speakers may have benefited from the processing advantage presented by a nominal constituent in VO compounds, and hence, the effect of morpheme frequency appeared more pronounced for VO compounds than VV compounds. While the word superiority effect prevents compounds from decomposition in the L1 processing, L2 speaker may have opted for compositional processing and make distinctions between VO and VV compounds based on the semantic information presented by different parts of speech (noun vs verb).

6 Conclusion

The results of the analysis of L1 response time data provide supporting evidence for word superiority effect in Chinese as their response times for VO and VV compounds were affected only by bigram frequency. However, despite being advanced level Chinese speakers, L2 speakers in this study tend to be more analytic of reading VO and VV compounds and distinguish the two compound types based on the difference in the second constituent morpheme (noun vs verb).

References

Andrews, S., Miller, B., & Rayner, K. (2004). Eye movements and morphological segmentation of compound words: There is a mouse in mousetrap. *European Journal of Cognitive Psychology*, 16: 285-311.

Baayen, R. H., Dijkstra, T., & Schreuder, R. (1997). Singulars and plurals in Dutch: Evidence for a parallel dual-route model. *Journal of Memory and Language*, 37(1): 94-117.

Clahsen, H., & Felser, C. (2006). Continuity and shallow structures in language processing. *Applied Psycholinguistics*, 27(1): 107-126.

Clashen, H., Gerth, S., Heyer, V. & Schott, E. (2015). Morphology constrains native and non-native word formation in different ways, *The Mental Lexicon*, 10(1): 53-87.

Cordier, F., Croizet, J-C., & Rigalleau, F. (2013). Comparing nouns and verbs in a lexical task. J Psycholinguist Res, 42: 21-35.

Huang, J. (1984). Phrase structure, lexical integrity and Chinese compounds. *Linguistics Inquiry*, 15: 530-78.

Kauschke, C., & Stenneken, P. (2008). Differences in noun and verb processing in lexical decision cannot be attributed to word form and morphological complexity alone. *Journal of Psycholinguistic Research*, 37: 443–452.

Mattingly, I., & Xu, Y. (1994). Word superiority in Chinese. In H-W. Chang, J-T. Huang, C-W. Hue & O. J. L. Tzeng (Eds.), *Advances in the study of Chinese language processing, Vol I: Selected Writing from the sixth international symposium on cognitive aspects of Chinese language* (pp. 101-111). Taipei: Department of Psychology, National Taiwan University.

Packard, J. (2000). *The Morphology of Chinese*. Cambridge, UK: Cambridge University Press.

Sereno, J. A. (1999). Hemispheric differences in grammatical class. *Brain and Language*, 70: 13–28

Shen, W & Li, X. (2012). The uniqueness of word superiority effect in Chinese reading (in Chinese). Chin Sci Bull (Chin Ver), 57: 3414-3420

Taft, M., & Forster, K. I. (1976). Lexical storage and retrieval of polymorphemic and polysyllabic words. *Journal of Verbal Learning and Verbal Behavior*, 15: 607-620

Yang, F. (杨帆), Han, W. (韩威) (2016). V-O Detachable words based on interlanguage corpus: error types and countermeasures (基于汉语中介语语料库的述宾式离合词偏误类型分析与对策). *East Forum (东方论坛)*, 2016(1): 116-122.

Syntax and Semantics of Adjectives in Cape Verdean Creole: A View from Markedness

Chigusa Morita
Toita Women's College
2-21-17 Shiba, Minato-ku,
Tokyo, 105-0014, Japan
morita@toita.ac.jp

Miki Obata
Hosei University
3-7-2 Kajino-Cho, Koganei
Tokyo, 184-8584, Japan
obata@hosei.ac.jp

Abstract

The main goal of this work is to describe some grammatical properties observed in Cape Verdean Creole (CVC), which is a less-studied language. In particular, we focus on the semantic restriction of adjectives observed in CVC and suggest that marked structures in individual languages need to satisfy additional semantic requirements at the interface by extending Pires and Taylor's (2007) common ground requirement for wh-in-situ phenomena in overt wh-fronting languages.

1 Introduction

This work aims to describe some grammatical properties observed in Cape Verdean Creole (henceforth CVC), which is a Portuguese-based creole, through comparison with other languages. The specific case we examine here is the semantic restriction of adjectives.

The paper is organized as follows: After presenting some basic properties of adjectives in CVC, Section 2 discusses two types of semantic restriction in CVC; one is gained through agreement while the other is triggered by movement, the latter of which we focus on in this work. Section 3 presents the syntactic structure which generates the semantic restriction, based on Cinque (2010), and also considers how semantic differences are obtained between unmarked structures and marked structures in a single language. We also present some similarities between the semantic restriction of adjectives in CVC and wh-in-situ phenomena observed in overt wh-fronting languages, based on Pires and Taylor (2007), and propose that marked structures, but not unmarked ones, in a given language need to satisfy additional semantic requirements. The final section clarifies some consequences obtained from the proposed analysis and concludes our work.

2 Semantic Restrictions on Adjectives

In this section, we demonstrate some cases in which attributive adjectives are interpreted differently depending on their position. We show that postnominal adjectives are semantically restricted relative to prenominal ones in CVC. We also present similar phenomena observed in English.

2.1 Two Types of Semantic Restriction: Adjectives in Cape Verdean Creole

Phi-agreement between an adjective and its modifying noun optionally takes place in CVC and the default form (without agreement) is a masculine ending, according to Baptista (2002). If the head noun is human, for example, adjectives optionally agree with their modifying nouns for gender as in (1), but not for number as in (2).

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 496-502, Hakodate, Japan, September 13-15, 2019

(1) Gender agreement
   a. un mininu bunit**u**
      a boy handsome.MAS
      'a handsome boy'
   b. un minina bunit**a**
      a girl beautiful.FEM
      'a beautiful girl'
   c. uns mininu bunit**u**
      some boy handsome.MAS
      'some handsome boys'
   d. un minina bunit**a**
      some girl beautiful.FEM
      'some beautiful girls'

(2) Number agreement
   a. Ano *e* animadu.
      NONCL COP courageous
      'We are courageous.'
   b. *Ano *e* animadu**s**.
      NONCL COP courageous.PLU
      'We are courageous.'

(Baptista 2002: 66)

Interestingly, phi-agreement triggers a semantic restriction, as given in (3).

(3) a. Elsa *e* un **bon** mudjer.
      Elsa COP a good woman
      'Elsa is a good woman.'
   b. Elsa *e* un **boa** mudjer.
      Elsa COP a good woman
      'Elsa is an attractive woman'

(Baptista 2002: 68)

The adjective *bon* is interpreted as 'good,' as in (3a). When the adjective *bon* changes to *boa*, which is a feminine ending, as in (3b), it is interpreted as 'attractive.' Based on Baptista's (2002) analysis, Obata and Morita (2018, 2019) claim that the adjective *boa* is derived from *bon* through (optional) agreement, which specifies/ limits the meaning of the adjective *bon*/*boa*.[1]

[1] Semantic restriction by agreement can also be observed in Japanese. Japanese has two morphological forms of adnominal adjectives: the *-i* form ('stem-*i*') and the *-na* form ('stem-*na*'). Although those two forms are not interchangeable in most cases, there are some adjectival stems to which both the morphemes *-i* and *-na* can attach. But a semantic difference is observed between the *-i* and *-na* forms derived from the same stem, as given in (i).

(i) a. ooki-i ani
      big-I elder.brother
      '(physically) big elder brother'
      'older elder brother'
   b. ooki-na ani
      big-NA elder.brother
      '(physically) big elder brother'
      #'older elder brother'

The *-i* form *ooki-i* is ambiguous between '(physically) big elder brother' and 'older elder brother,' while the *-na* form lacks the latter interpretation. Obata and Morita (2018, 2019) propose that the *-na* form is obtained by phi-agreement, which triggers the semantic restriction. See Obata and Morita (2018, 2019) for more details.

CVC has another way of restricting the interpretations of adnominal adjectives. Some attributive adjectives in CVC can appear either prenominally or postnominally, which is also observed in French. Baptista (2002) observes that prenominal and postnominal adjectives differ in interpretation. [2] Notice that the position of adjectives does not affect the agreement pattern.

(4) a. João ten un **noba mudjer**.
      John has a new wife
      'John has a new wife.'

[2] One might wonder if the semantic restriction we are discussing is observed not only in CVC but also in French. It can also be observed in French that attributive adjectives have different interpretations depending on their positions. However, the semantic restriction does not take place by changing the order of adjectives and nouns in French. Consider the adjective *faux*, for example.

(i) a. de **faux** pianos
      some false pianos
      'false (fake) pianos'
   b. des pianos **faux**
      some pianos false
      'pianos that are out of tune'

(Bouchard 2002: 74)

According to Bouchard (2002), the adjective *faux* means 'false, fake' in the prenominal position, as in (ia). When the adjective *faux* appears postnominally, as in (ib), it is interpreted as 'out of tune.' Following Bouchard, we assume that the difference in meaning between the prenominal adjective *faux* in (ia) and the postnominal one in (ib) is due to the difference in modification: the prenominal adjective modifies a subpart of the noun, while the postnominal adjective modifies the entirety of the noun.

b. João ten un **mudjer noba**.
John has a wife young
'John has a young wife.'

(5) a. Tenha un **grandi omi** ki ta txoma
was a big man COMP TMA call
Nho Djiku.
Nho Djiku
'There was an big man whose name was Nho Djiku.'

b. Tenha un **omi grandi** ki ta txoma
was a old man COMP TMA call
Nho Djiku.
Nho Djiku
'There was an old man whose name was Nho Djiku.'

(Baptista 2002: 70, Marlyse Baptista p.c.)

The adjective *noba*, meaning 'new' in English, is interpreted differently depending on where it appears. The prenominal adjective *noba* in (4a) is interpreted as 'new,' while the postnominal one in (4b) is interpreted as 'young.' Given that 'young' belongs to the semantic range of 'new,' the semantic range of the adjective *noba* is narrowed down and limited to 'young' when the adjective appears in the postnominal position. Also, the adjective *grandi* meaning 'great' in English is interpreted as 'big/tall/great' in the prenominal position as in (5a) while the postnominal one in (5b) is interpreted as 'old'. Since 'big' includes the meaning of 'old' e.g. in *big brother*, the semantic range of *grandi* is narrowed down and limited to 'old' in this case, too. In other words, a semantic restriction is observable in the case of postnominal adjectives.

To sum up, two semantic restriction cases are observed in CVC: semantic restriction occurs through agreement, and it can be also observed by changing the order of adjectives and nouns. In this paper, we especially focus on the latter case, in which semantic restriction occurs by changing word order.

2.2 Adjectives in English

The same type of semantic restriction can be also found in English. According to Cinque (2010), prenominal and postnominal adjectives differ in meaning in English. The adjective *possible*, for example, is interpreted either as 'potential' or as 'can be done or achieved' in the prenominal position, while it only has the latter interpretation in the postnominal position.

(6) Mary interviewed every **possible** candidate.
a. 'Mary interviewed every potential candidate.'
b. 'Mary interviewed every candidate that was possible for her to interview.'

(7) Mary interviewed every candidate **possible**.
a. #'Mary interviewed every potential candidate.'
b. 'Mary interviewed every candidate that was possible for her to interview.'

(Cinque 2010: 8)

The prenominal adjective *possible* in (6) is ambiguous between two readings: (6a) and (6b) while the postnominal one in (7) has only one reading, (7b), in which the reading is disambiguated. Again, a semantic restriction occurs when attributive adjectives appear in the postnominal position, just like in CVC.

Notice that the two interpretations of the prenominal adjective *possible* in (6a) and (6b) do not result from a difference in how the adjective modifies the noun. The above case should be distinguished from the following case, in which the prenominal adjective *beautiful* in (8) is ambiguous between two readings in (8a) and (8b).

(8) Olga is a beautiful dancer.
a. 'Olga is a dancer who dances beautifully.'
b. 'Olga is a dancer and she is beautiful.'

(Larson 1999)

Given that the meaning of the noun *dancer* is paraphrased as 'a person who dances,' the adjective *beautiful* is interpreted differently depending on whether the adjective modifies the individual or the hidden event denoted by the noun. When the adjective modifies the event 'dance,' it is interpreted as in (8a). The adjective, on the other, is interpreted as in (8b) when it modifies a person who habitually dances. In this paper, we do not deal with this case, in which semantic ambiguity is caused by a difference in modification of nouns by adjectives.

In sum, a semantic restriction results from changing the order of adjectives and nouns in CVC as well as in English.

3 Semantic Restriction in Marked Structures

In order to explain why and how a semantic restriction occurs by changing word order, we consider the following two questions. First, which of the orders AP-NP and NP-AP is unmarked in CVC and English? Second, why are postnominal adjectives semantically restricted in CVC and English?

In this section, we demonstrate that the order AP-NP is unmarked in both CVC and English, and semantic restriction occurs only in the marked order NP-AP. We also show another case where a semantic restriction occurs in the marked structure: wh-in-situ in overt wh-fronting languages. By extending Pires and Taylor's (2007) common ground requirement for wh-in-situ, we claim that syntactic representations including marked structures can be interpreted properly at the semantic interface by satisfying an additional semantic requirement, which forces postnominal adjectives to be disambiguated among several possible interpretations.

3.1 Structural Analysis of the Order of Adjectives and Nouns

Let us consider the research question of which of the orders AP-NP and NP-AP is unmarked in CVC and English. Following Cinque (2010), we assume that there is only one structure available for all languages, and all attributive adjectives are merged into the prenominal position. As in (9), the functional head F is first merged with NP, and then adjectives are merged. F is the functional head which constructs a modifying relation between AP and NP, in accordance with Cinque (2010):

(9) $[_{FP}$ AP F $[_{FP}$ AP F $[_{FP}$ AP F NP]]]

Based on Cinque's structural analysis, the order of NP-AP is derived from the order of AP-NP by applying syntactic movement/Internal Merge. The order NP-AP is obtained when NP moves to the specifier of FP, as shown in (10b). If syntactic movement does not take place, the order AP-NP is available, as in (10a).

(10) a. AP-NP: $[_{FP}$ F $[_{FP}$ AP F NP]]
b. NP-AP: $[_{FP}$ NP F $[_{FP}$ AP F <NP>]]

According to Cinque (2005, 2010), syntactic movement of NP to the specifier of FP is triggered by a nominal feature. He proposes that each phrase has a nominal feature to be licensed, and this can be satisfied either by movement of NP or by merging a nominal feature, which agrees with NP without movement.

It is a matter of parametric variation across/within languages whether syntactic movement of NP takes place or not: some languages employ the movement strategy for a nominal feature to be licensed, others employ the non-movement strategy, and still others employ both. Cinque (2005, 2010) claims that many attributive adjectives in Romance languages such as French and Italian obligatorily appear postnominally, and thus the derived order NP-AP is unmarked in these languages. In Germanic languages such as English and German, on the other hand, the base-generated order AP-NP is unmarked, since adjectives usually appear in the prenominal position. That is, the derived order/structure is marked in some languages, while it is unmarked in other languages.

Recall that a semantic restriction can be observed by changing the order of adjectives and nouns in English, whose data in (6) and (7) are repeated as (11) and (12).

(11) Mary interviewed every **possible** candidate.
a. 'Mary interviewed every potential candidate.'
b. 'Mary interviewed every candidate that was possible for her to interview.'

(12) Mary interviewed every candidate **possible**.
a. #'Mary interviewed every potential candidate.'
b. 'Mary interviewed every candidate that was possible for her to interview.'

Since the order NP-AP is allowed only with some specific adjectives, it is reasonable to assume that in English the order AP-NP is unmarked while the order NP-AP is marked. Under this assumption, the adjective *possible* is interpreted more restrictedly in the marked structure (12) than in the unmarked structure (11). That is, we can make the generalization that the semantic restriction occurs only in the marked structure.

We have seen that the same pattern of semantic restriction can be also observed in CVC, as repeated in (13).

(13) a. João ten un **noba mudjer**.
John has a new wife
'John has a new wife.'
b. João ten un **mudjer noba**.
John has a wife young
'John has a young wife.'
(Baptista 2002: 70)

Similar to French, many adjectives in CVC appear postnominally, and some adjectives, including *noba* ('new'), *grandi* ('great'), and *bon* ('good'), occur prenominally. In this sense, CVC employs both orders as unmarked and either of the orders is chosen depending on the type of adjective. That is, we can say that the AP-NP order is unmarked in (13), just as in English. The semantic range of adjective *noba* is restricted in the marked NP-AP structure. Under this view, the semantic restriction is present in the marked structure in CVC.

To sum up, we have demonstrated that the interpretations of attributive adjectives are restricted in the marked structure: the NP-AP order is marked both in English and in some adjectives of CVC, so that postnominal adjectives have more restricted interpretations than prenominal ones.

3.2 Wh-in-situ in Wh-Fronting Languages: The Common Ground Requirement

In the above section, we presented the generalization that the marked structure causes a semantic restriction in the case of adjectives in English and CVC. In fact, the same type of semantic restriction can be also observed in the case of wh-in-situ in overt wh-fronting languages such as English.

According to Pires and Taylor (2007), wh-in-situ is allowed in a single wh-question in a language like English, and is not limited to echo-questions, as shown below:

(14) *Wh-question with overt wh-fronting:*
a. What did you eat?
b. *Did you eat what?
(15) *Echo-question*
A: Mary ate a skunk.
B: Mary ate WHAT ↑ ?
(16) *[+specific] question*
A: I made desserts.
B: You made [what ↑ kind of desserts ↓] ?
(17) *Expect-question*
A: [employee]: I made many different kinds of desserts.
B: [manager]: So, you made [how many cookies ↓]?
(Pires and Taylor 2007: 202-203)

The wh-phrase standardly undergoes overt movement in English, as in (14). As in (15)-(17), however, there are some cases in which the wh-phrases can stay in-situ in single wh-questions. (15) is an example of an echo-question, which repeats a part or all of the sentence which has been just uttered. (16) is an example of a [+specific] question, requesting more specific information about the utterance which has been just given. In (16), that is, B is asking for more specific information about desserts (e.g. ice cream, chocolate cakes, etc.) (17) is an example of an expect-question, asking for further new information. In (17), B is expecting that A made several desserts, including cookies, and requesting the number of cookies. Under these specific environments, wh-in-situ is allowed even in an overt wh-fronting language like English.

Considering possible answers to each of the question sentences in (14)-(17), we can find clear differences between (14) and (15)-(17). In (14), everything you ate can be a possible answer. In (15)-(17), on the other hand, the range of possible answers is more restricted. In (15), for example, what A said Mary ate is the only possible answer. In (16)-(17), the utterance by A which has been just given limits the possible answers to B's question. Pires and Taylor (2007) suggest that those wh-in-situ examples need to satisfy the common ground requirement, which requires the set of possible answers to those questions to be part of the common ground defined in Stalnaker (1978: 704):

(18) Common Ground:
"Presuppositions are what is taken by the speaker to be the common ground of the participants of the conversation, what is treated [by the speaker, AP&HT] as their common knowledge or mutual knowledge."
(Pires and Taylor 2007: 205)

In other words, the possible answers are restricted by the common ground requirement, and only when this requirement can be satisfied, wh-in-situ is allowed in overt wh-fronting languages. In this sense, this semantic-pragmatic requirement makes the marked structure (i.e. wh-in-situ) interpretable properly at the interface.

Interestingly, the common ground requirement needs to be satisfied only in the marked structure. Since English is an overt wh-fronting language, wh-phrases undergo overt movement, which is the unmarked structure in this language. Remember that adjectives in English and CVC also show a semantic restriction only in the marked structure: the NP-AP order. Thus, we can say that wh-in-situ phenomena in English and the NP-AP order in English and CVC behave in the same manner with respect to semantic restrictions in the marked structures, so the generalization we presented in Section 3.1 gains additional empirical support from wh-in-situ phenomena.

3.3 Semantic Disambiguation of Adjectives at the Conceptual-Intentional (CI) interface

As discussed in the last section, the marked structure (i.e. wh-in-situ in overt wh-fronting languages) can be properly interpreted at the interface by satisfying the common ground requirement, which is an additional semantic(-pragmatic) requirement. Based on this view, we can say that the marked NP-AP structure can be ruled in and interpreted properly at the interface by satisfying an additional semantic requirement. This is why adjectives cannot be ambiguous among several possible interpretations in the marked structure, unlike in the unmarked structure. That is, the additional semantic requirement in this case is the semantic disambiguation requirement, which limits several possible interpretations of adjectives to a single interpretation. By satisfying this additional requirement, the NP-AP structure in English and CVC can be properly interpreted at the interface.

Although we did not discuss in detail another type of semantic restriction triggered through syntactic agreement in (3), which is studied in Obata and Morita (2018, 2019), the meaning of *bon/boa* meaning ‘good’ is limited to the specific meaning ‘attractive’ if agreement takes place. In CVC, gender agreement between an adjective and its modifying noun is optional and the default foam, a masculine ending, is assigned if agreement does not take place. Considering the optionality of gender agreement and the existence of the default form, we can say that the option of applying syntactic agreement is marked. The proposed analysis can thus be extended to another type of semantic restriction triggered by agreement.

4 Consequences and Conclusion

In this paper, we have seen how a semantic restriction occurs in the case of adjectives and proposed a theory of why it happens. We demonstrated that markedness matters: in the marked structure, but not in the unmarked structure, the additional semantic requirement, i.e. the semantic disambiguation requirement, needs to be satisfied, so that keeping adjectives ambiguous is not allowed in the marked structure. We extended Pires and Taylor’s (2007) common ground requirement for wh-in-situ phenomena in overt wh-fronting languages. If the proposed analysis is on the right track, we can find semantic and syntactic commonalities between adjectives and wh-questions.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP17K02823. We are very grateful to Marlyse Baptista for insightful comments and suggestions.

References

Baptista, Marlyse. 2002. The Syntax of Cape Verdean Creole. John Benjamins Publishing Company, Amsterdam, Netherland and Philadelphia, PA.

Bouchard, Denis. 2002. Adjectives, Number and Interfaces: Why Language Vary. Elsevier, Oxford, UK.

Cinque, Guglielmo. 2005. Deriving Greenberg’s Universal 20 and Its Exceptions. Linguistic Inquiry 36 (3), 315-332.

Cinque, Guglielmo. 2010. The Syntax of Adjectives: A Comparative Study. MIT Press, Cambridge, MA and London, UK.

Larson, Richard K. 1999. Semantics of Adjectival Modification. Lecture Notes, LOT Winter School, Amsterdam.

Obata, Miki and Chigusa Morita. 2018. Syntactic Agreement as a Disambiguation Task: Evidence from Japanese Adjectives. Ambiguity: Perspectives on Representation and Resolution, presented at Workshop at ESSLLI 2018, University of Sofia.

Obata, Miki and Chigusa Morita. 2019. Three Types of Adjectives in Japanese: A View from Cape Verdean Creole, presented at Society for Pidgin and Creole Linguistics.

Pires, Acrisio and Heather Taylor. 2007. The syntax and Wh-In-Situ and Common Ground. *The Proceedings of the Annual Meeting of Chicago Linguistics Society 43*, Vol. 2, 201-215.

Stalnaker, R. 1978. Assertion. *Pragmatics, Syntax and Semantics Vol. 9*. ed. by P. Cole. New York: Academic Press.

Japanese Daily Utterance Styles: A Factor Analysis based on Balanced Corpus

Hajime Murai
Future University Hakodate
116-2 Kamedanakano-cho, Hakodate,
Hokkaido, Japan
h_murai@fun.ac.jp

Abstract

It is often considered more difficult to extract fundamental utterance styles in everyday conversation than in fictional utterances. This is because the characteristics of utterance styles are exaggerated in fictional utterances. However, by referring to a large-scale corpus of daily conversations, it is possible to identify the fundamental patterns of everyday Japanese utterance styles. This study employs the statistical method of factor analysis to identify the characteristics of utterance styles within the Corpus of Everyday Japanese Conversation - a gender and age balanced corpus. Eight factors ("Neutral style," "Dialect style," "Frank style," "Polite style," "Feminine style," "Crude style," "Series style," and "Parallel style") were extracted quantitatively. The results suggest that "Series style" and "Parallel style" are unique to everyday conversation. On the other hand, "Aged style," "Interrogative style," "Approval style," and "Dandy style" (found in utterances in written fiction) were not found. Unlike previous studies, these results are based on a balanced corpus.

1 Introduction

Utterance styles are affected by various factors, such as gender, age, context, cultural setting, social background, personalities of the characters, and the mood of the moment.

Elsewhere in the literature, characterized written styles is applied to texts of various kinds for text categorization and author identification tasks (Zheng et al., 2006; Stamatatos, 2009). Previous research has analyzed the characteristics of utterance styles mainly on the basis of gender, or age (Argamon et al., 2006; Schwartz et al., 2013; Goswami et al., 2009).

In the case of Japanese fictional utterances (in novels or essays), an important way characters can be differentiated is on the basis of utterance style. This popular technique helps readers understand each character's personality (Kinsui, 2003). These utterance styles can be detected by comparing the frequency of function words in utterances. Furthermore, fundamental patterns of utterance styles composed of particles and auxiliary verbs can be identified by conducting a factor analysis of a fictional corpus (Murai, 2018A).

In the field of Japanese everyday conversation, the main research topics tended to focus on grammatical characteristics and pragmatic semantics (Seto and Kishi, 2015), as well as the relationships between single attributes (such as politeness and gender) and utterance styles (Kurosawa, 2010). It is considered more difficult to extract fundamental utterance styles in real, daily conversation than in fiction writing, because distinct utterance styles are often exaggerated in conversations between fictional characters (particularly in entertainment contents). Therefore, explorations of Japanese utterance styles in daily conversation have tended to employ case studies (Miyazaki et al., 2014) and psychological experimental approaches (Shen et al., 2012).

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 503-510, Hakodate, Japan, September 13-15, 2019

Previous attempts to extract utterance styles in daily conversation (Murai, 2018B) have yielded mixed results because of an unbalanced corpus. A total pattern of Japanese utterance styles identified using a quantitative analysis of a real-life, balanced corpus has so far been lacking. However, by drawing on a large-scale, balanced corpus of daily conversation – in this case, the trial version of the new Corpus of Everyday Japanese Conversation (Koiso et. al., 2016) – it has been possible to identify fundamental patterns using the statistical method of factor analysis. In addition, the present study examined the difference between utterance styles in fiction writing and in daily conversation. The difference of utterance styles had been predicted by previous linguistic studies (Kinsui, 2003) but it had not been examined quantitatively.

2 Corpus used in Analysis

The trial version of the Corpus of Everyday Japanese Conversation (CEJC) is composed of 126 real-life Japanese dialogues. The duration of these recordings is 3015 minutes. A total of 463 speakers were recorded their real-life natural dialogues (some speakers were included in several dialogues). The speakers are native Japanese speakers of various ages. The dialogues take place in various situations, such as home, school, workplace, restaurant and leisure. In addition, speakers have different relationships with each other, such as family, friends, co-workers, teachers, and students.

The factor analysis grouped utterances by speaker in 126 conversation scenes. In total, 435 utterance sets were identified in the CEJC, (excluding very reticent speakers who spoke less than ten words, for statistical reasons).

The attributes of the speakers in the 435 utterance sets are given in Table 1.

Age	Male	Female	Total
10 to 29	51	40	91
30 to 49	54	92	146
50 to 69	48	73	121
Over 70	19	18	37
Total	172	223	395

Table 1: Speaker details for utterance sets in the CEJC

Table 1 does not include 40 speakers whose attributes are unknown. These are accidental participants such as restaurant employees. In the analysis, these unknown speakers were included as the 435 utterance data set, except for very reticent speakers as mentioned above.

3 Characteristics of Utterance Styles

In this study, the frequency with which function words occur in utterances was adopted as a characteristic of utterance style, because in many Japanese novels, different usage patterns of function words are used to express characters' personalities (Kinsui, 2003). In the Japanese language, function words usually correspond to particles and auxiliary verbs. Therefore, the statistical significance of the frequency with which particles and auxiliary verbs occur was analyzed using factor analysis (Murai, 2018A) for a fictional corpus. The CEJC provides morphologically analyzed data sets for the conversation texts; therefore, particles and auxiliary verbs in utterances could be extracted from the 435 data set units and counted. The frequency with which the top 30 particles and auxiliary verbs appeared is shown in Table 2.

4 Factor Analysis for Utterance Styles

4.1 Factor Analysis for Daily Conversation

To extract the typical utterance styles of Japanese daily conversation, a factor analysis was carried out to establish the frequency with which particles and auxiliary verbs were used. Owing to statistical limitations, 83 function words (where the frequency of those particles and auxiliary verbs exceeded 10) were selected, and 83 dimensional word frequency vectors were extracted for each speaker, in each scene. The Promax rotation method was used and a parallel analysis performed to determine the number of factors involved. After the factor analysis, less significant words (with a maximum factor loading of < 0.3) were eliminated and a factor analysis was repeatedly performed for the eliminated data set. Finally, after performing the factor analysis four times, eight factors were identified. The resultant factor loadings are shown in Table 3. The bold font signifies cells where the absolute value of factor scores exceeded 0.3.

Table 4 depicts average factor scores for each age / gender category in the CEJC (as in Table 1). Table 5 depicts another average factor score for

each situational category. “Speakers” signifies the number of speakers who participated in each dialogue of that category. In both tables 4 and 5, bold font designates cells whose absolute value of factor scores exceeded 0.2.

Word and part of speech	Frequency
Auxiliary verb "Da"	29708
Final particle "Ne"	17250
Auxiliary verb "Ta"	14493
Connective particle "Te"	13466
Quasi-particle "No"	13335
Case particle "No"	10080
Case particle "De"	9852
Auxiliary particle "Ka"	9689
Case particle "Ga"	9496
Incidental particle "Mo"	9318
Incidental particle "Wa"	9112
Final particle "Yo"	8954
Case particle "Ni"	8870
Case particle "To"	8381
Auxiliary verb "Teru"	7947
Auxiliary particle "Tte"	7316
Auxiliary verb "Nai"	6711
Auxiliary verb "Desu"	6424
Final particle "Ka"	5919
Connective particle "Kara"	5784
Connective particle "Keredo"	4118
Final particle "No"	3756
Final particle "Na"	3589
Final particle "Sa"	3371
Case particle "Wo"	3253
Auxiliary verb "Masu"	2881
Auxiliary verb "Chau"	1957
Connective particle "To"	1794
Auxiliary verb "Tuu"	1652
Case particle "Kara"	1482

Table 2: Top 30 frequently appearing words

Eight factors corresponded with utterance patterns that frequently appeared in daily conversation in the CEJC. The characteristics of each, as well as an explanation of how each was named, are provided below:

Factor 1: Included the most frequently used general function particles and auxiliary verbs, such as the case particles “Wo,” “To,” “Ni,” “Ga,” “No,” “De,” and “Kara”. However, Factor 1 did not include words that indicated specific attributes. In other words, it represented a ‘Neutral style’ of utterance.

This utterance style seems to be commonly used by middle-aged females (Table 4). It is also generally employed in situations that are not particularly intimate, such as school, business or service situations, like shopping (Table 5).

Factor 2: Included particles and auxiliary verbs such as “Hen,” “Yan,” “De,” and “Nen”. These words are frequently used in Japanese dialects such as Kansai-ben. Therefore, Factor 2 is referred to as “Dialect style”.

This utterance style is often used by middle-aged males (Table 4). The reason would be that the category of middle aged men includes more people from Kansai region than other categories.

Factor 3: Included final particles such as “Jan,” “Yo”, “Mono,” “Ke,” and “Sa.” These are characteristic of informal, frank communication styles. Therefore, Factor 3 is referred to as “Frank style.”

Table 4 shows that this factor is strongly associated with males between the ages of 10 and 29. In addition, Table 5 shows that this utterance style is frequently used in family relationships and service situations. This suggests, for instance, that customers often use frank utterance styles when speaking to sales clerks in shopping situations.

Factor 4: Included the auxiliary verbs “Desu” and “Masu.” These are clearly related to Japanese honorific utterance styles. Therefore, Factor 4 is referred to as “Polite style.”

This style is generally used in less intimate situations such as schools or businesses, or in service settings like shopping (Table 5). It is similar to Factor 1 in this respect. This factor is common amongst young males (see Table 4), it is also often used in dialogues at school (which could also include young male students in Table 5).

Factor 5: Included feminine characteristic particles (e.g. “Wa,” “Kashira,” and “No”) and is thus referred to as “Feminine style”. This utterance style is related to middle-aged and elderly females as expected (in Table 4). However, young women do not appear to use this traditional utterance style.

	F1	F2	F3	F4	F5	F6	F7	F8
Case particle "Wo"	**1.04**	-0.01	-0.12	0.04	-0.09	0.00	0.10	-0.26
Case particle "To"	**0.99**	-0.01	-0.19	-0.02	0.00	-0.04	-0.12	**0.31**
Auxiliary particle "Ka"	**0.92**	-0.05	-0.13	-0.06	0.06	0.00	**-0.33**	**0.36**
Auxiliary particle "Tari"	**0.90**	0.01	**-0.32**	0.00	-0.06	-0.02	-0.15	0.11
Case particle "Ni"	**0.85**	-0.02	-0.09	0.00	0.09	0.07	0.12	0.07
Case particle "Ga"	**0.84**	-0.02	0.06	0.09	-0.01	0.03	0.11	-0.07
Incidental particle "Mo"	**0.80**	-0.01	0.01	0.01	0.13	-0.04	-0.08	0.18
Connective particle "To"	**0.80**	-0.03	-0.18	0.08	0.09	0.12	-0.06	0.02
Connective particle "Keredo"	**0.79**	-0.04	0.05	0.01	0.02	-0.02	-0.10	0.21
Incidental particle "Wa"	**0.78**	-0.01	0.09	0.09	0.01	0.00	0.03	-0.05
Connective particle "Te"	**0.78**	0.03	-0.13	0.03	0.05	-0.02	0.26	0.16
Case particle "No"	**0.78**	0.01	0.13	0.05	-0.09	-0.01	0.16	-0.05
Auxiliary particle "Tte"	**0.76**	0.01	0.25	-0.04	-0.10	-0.02	0.01	0.08
Auxiliary verb "Da"	**0.70**	-0.07	**0.34**	-0.08	0.20	0.05	-0.13	-0.02
Case particle "De"	**0.68**	0.00	0.10	0.09	0.11	0.00	0.08	0.05
Quasi-particle "No"	**0.64**	-0.04	0.16	0.29	0.07	-0.01	-0.04	0.02
Final particle "Na"	**0.62**	0.29	0.14	0.07	0.12	0.00	-0.12	-0.10
Auxiliary verb "Teru"	**0.59**	0.05	**0.31**	-0.08	0.03	-0.01	0.04	0.16
Auxiliary verb "Seru"	**0.57**	-0.03	0.00	-0.11	-0.05	0.02	-0.05	-0.15
Auxiliary verb "Ta"	**0.55**	0.01	0.18	-0.03	0.17	0.05	0.09	0.14
Final particle "Ne"	**0.54**	-0.09	0.15	0.09	**0.53**	-0.09	-0.17	-0.20
Auxiliary verb "Rareru"	**0.54**	-0.01	0.19	-0.07	-0.22	0.01	0.09	-0.02
Connective particle "Kara"	**0.54**	0.04	0.26	-0.13	0.06	-0.01	0.16	0.15
Connective particle "Nagara"	**0.52**	0.01	0.10	0.06	-0.05	-0.04	-0.09	-0.04
Auxiliary verb "Reru"	**0.51**	0.03	**0.31**	-0.08	-0.20	0.02	0.04	0.13
Connective particle "Shi"	**0.50**	0.01	0.08	-0.02	0.00	0.04	-0.12	**0.33**
Case particle "Kara"	**0.48**	-0.02	0.02	-0.03	0.08	-0.02	**0.33**	0.08
Auxiliary verb "Nai"	**0.47**	-0.04	**0.34**	0.03	0.03	0.00	0.10	0.14
Auxiliary particle "Kurai"	**0.43**	0.01	-0.10	0.22	0.12	-0.01	0.09	0.20
Connective particle "Ba"	**0.41**	-0.02	0.04	0.23	-0.14	0.01	0.27	0.08
Auxiliary particle "Dake"	**0.40**	0.08	0.12	0.12	0.02	-0.08	0.10	0.06
Auxiliary verb "Hen"	0.01	**1.00**	0.00	-0.02	0.00	-0.14	0.03	0.02
Final particle "Yan"	0.01	**0.96**	0.01	-0.02	-0.01	0.06	-0.02	0.02
Final particle "De"	-0.02	**0.93**	0.00	0.01	0.03	-0.16	0.08	-0.01
Final particle "Nen"	-0.04	**0.91**	-0.07	0.00	0.04	0.27	-0.02	-0.01
Final particle "Jan"	0.15	-0.03	**0.62**	-0.27	-0.19	0.07	0.07	0.17
Final particle "Yo"	0.15	-0.01	**0.59**	0.19	**0.31**	-0.01	0.05	-0.18
Final particle "Mono"	-0.08	0.03	**0.55**	0.15	0.14	0.05	0.00	0.04
Auxiliary verb "Tuu"	**0.34**	-0.01	**0.48**	0.10	-0.04	0.05	0.08	-0.17
Final particle "Ke"	0.12	0.02	**0.47**	0.08	0.04	0.05	-0.24	0.01
Final particle "Sa"	**0.42**	-0.04	**0.45**	**-0.32**	-0.01	-0.04	-0.02	0.01
Auxiliary particle "Sura"	0.04	-0.01	**0.40**	0.06	-0.09	-0.04	-0.01	-0.11

Table 3-1: Results of factor analysis of frequently appearing function words in the CEJC

	F1	F2	F3	F4	F5	F6	F7	F8
Auxiliary verb "Desu"	-0.03	0.01	-0.02	**1.11**	-0.28	0.00	0.04	0.02
Auxiliary verb "Masu"	0.28	-0.01	-0.28	**0.80**	-0.10	0.02	0.07	-0.10
Final particle "Ka"	**0.34**	-0.03	0.12	**0.61**	0.03	-0.01	-0.05	-0.05
Connective particle "Tutu"	0.00	-0.01	0.20	**0.34**	-0.03	-0.04	-0.11	-0.05
Final particle "Kashira"	-0.02	-0.06	-0.12	-0.06	**0.59**	-0.02	0.07	-0.17
Final particle "Wa"	0.00	0.10	0.01	-0.14	**0.57**	0.00	-0.01	-0.09
Final particle "No"	0.20	-0.01	**0.38**	**-0.33**	**0.39**	-0.03	0.06	0.04
Auxiliary verb "Chau"	0.26	-0.09	0.13	0.05	**0.36**	0.01	0.13	0.00
Auxiliary verb "Yagaru"	-0.13	-0.02	-0.01	0.04	-0.04	**0.76**	-0.08	0.08
Auxiliary verb "Beshi"	-0.05	-0.03	-0.04	-0.04	0.08	**0.63**	-0.07	0.00
Connective particle "Ga"	0.30	0.05	-0.09	-0.01	0.04	**0.49**	-0.04	-0.18
Final particle "Zo"	0.00	0.00	0.17	-0.04	-0.13	**0.34**	0.04	-0.05
Case particle "He"	0.03	0.03	-0.06	0.01	0.06	-0.06	**0.50**	-0.11
Connective particle "Tatte"	-0.04	0.00	0.01	-0.04	0.00	-0.02	**0.40**	0.03
Auxiliary verb "Teku"	**0.33**	-0.01	0.01	0.00	0.02	-0.10	**0.35**	0.03
Auxiliary verb "Toku"	0.00	-0.01	-0.09	0.11	0.22	0.26	**0.34**	0.05
Auxiliary verb "Saseru"	0.24	0.00	-0.07	-0.07	-0.15	-0.03	-0.01	**0.37**
Auxiliary verb "Tai"	**0.31**	0.02	0.07	0.19	-0.07	-0.04	-0.08	**0.34**
Auxiliary particle "Shika"	0.14	0.01	0.25	0.12	-0.13	0.11	0.06	**0.32**

Table 3-2: Results of factor analysis of function words frequently appearing in the CEJC

	Age	F1	F2	F3	F4	F5	F6	F7	F8
Male	10 to 29	0.09	0.00	**0.42**	**0.22**	**-0.23**	**0.43**	-0.12	0.19
	30 to 49	0.12	**0.39**	0.06	0.11	-0.09	0.10	-0.15	-0.08
	50 to 69	-0.18	-0.09	-0.13	-0.16	**-0.23**	-0.11	-0.15	**-0.30**
	Over 70	-0.04	-0.07	-0.12	-0.16	-0.03	-0.07	**0.69**	-0.11
Female	10 to 29	**-0.21**	-0.06	-0.17	-0.17	**-0.32**	-0.08	-0.12	-0.01
	30 to 49	**0.26**	-0.08	0.19	0.14	**0.45**	-0.02	0.07	**0.27**
	50 to 69	0.18	-0.01	0.08	0.11	**0.29**	-0.05	0.05	0.16
	Over 70	**-0.26**	-0.09	**-0.27**	-0.14	0.18	-0.12	**0.54**	**-0.27**

Table 4: Average factor scores for each gender / age category in the CEJC

	Speakers	F1	F2	F3	F4	F5	F6	F7	F8
Family	117	0.00	0.04	**0.24**	**-0.28**	-0.09	**0.22**	**0.21**	0.08
Family and relatives	58	**-0.32**	-0.09	**-0.22**	-0.15	-0.02	-0.16	**0.20**	-0.14
Friends	156	0.07	0.05	0.08	0.11	0.15	-0.02	-0.09	0.11
Teachers and students	6	**1.03**	-0.07	-0.03	**1.83**	0.15	0.01	**-0.54**	**0.20**
Business relationship	14	**0.36**	-0.09	**-0.27**	**0.52**	-0.01	-0.07	0.19	**-0.25**
Co-worker	22	**-0.20**	-0.07	-0.09	-0.05	0.00	-0.19	**-0.43**	-0.06
Sales clerk and customer	7	**0.28**	-0.09	**0.26**	**0.26**	**0.20**	-0.04	-0.03	**0.29**

Table 5: Average factor scores for each situation category in the in the CEJC

Factor 6: Included relatively crude expressions such as “Yagaru,” and “Zo” and connective particle “Ga” which also has a crude nuance. Therefore, it is labelled “Crude style.”

This utterance style is common amongst young males (Table 4). It is not suitable in formal situations; and is therefore only associated with the family situation (Table 5).

Factor 7: Included the case particle “Kara,” “He,” and connective particle "Tatte". “Kara” and “Tatte” are often used to signify logical connections, such as cause and effect. It is therefore associated with a series of connected utterances using particles that represent logical relationships. And in consequence is referred to as “Series style.” This style is used mainly by elderly males and females (Table 4).

Factor 8: Included case particle "To," auxiliary particle "Ka," and connective particle "Shi". Both particles are used to juxtapose sentences or phrases, much like the English words “and” or “or”. In contrast with Factor 7, this utterance style indicates that utterances are connected in a parallel fashion using particles for juxtaposition. Therefore, it is labelled “Parallel style.” This utterance style is most often used by young males and middle age females (Table 4).

Those eight factors and factor scores seem to reflect daily use of utterance styles as mentioned in explanations of factors. However, in some case, detailed differences were not discriminated. For instances, utterance styles in shops include both sales clerk and customer (in Table 5), and the polite utterance style of clerks and the frank utterance style of customers would be combined in the result. Because the utterances were tagged according to situation.

4.2 Comparison with Factor Analysis for Fiction Writings

In order to establish which characteristics might be unique to utterance styles in real daily conversation, the results of the factor analysis were compared with similar results for utterance styles within fiction writing (Murai, 2018A).

The analysis of utterance styles in novels are based on a random sampling of dialogues in Japanese novels. These dialogues are drawn from a subset of texts within the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014). These texts from Japanese novels, are included under the Nippon Decimal Classification number 913. One hundred texts were randomly sampled from Japanese novels included in the library-based corpus in the BCCWJ. In addition, a speaker database provides gender and age attributes for each speaker who appears in the novel texts. Each data entry used in the factor analysis consists of a 100-dimensional vector for all utterances attributed to a single fictional character. Those dimensions indicate the frequency with which 100 types of commonly used particles and auxiliary verbs are used. Because of statistical limitations, 7576 utterance data sets (utterance sets of 7576 speakers in fiction writings) with total frequencies higher than 20 were selected from 11860 data sets.

As a result, ten factors were identified in fiction writing utterance data: “Neutral style,” “Frank style,” “Dialect style,” “Polite style,” “Feminine style,” “Crude style,” “Aged style,” “Interrogative style,” “Approval style,” and “Dandy style” (Murai, 2018A).

The first six factors (“Neutral style,” “Frank style,” “Dialect style,” “Polite style,” “Feminine style,” and “Crude style”) are nearly as common in the utterance styles of daily conversation between real people, and the utterance styles used in fictional conversation between fictional characters. These six utterance styles are therefore clearly characteristic of Japanese speech in general. Examples are shown in Table 6. The sentences used in the example all have the same meaning, “What are you doing?” in Japanese. The difference in nuance cannot be expressed in the English language.

Style	Example	Japanese
Neutral	Nani wo shite iru?	何をしている？
Frank	Nani shi teru?	何してる？
Dialect	Nani shi ten nen?	何してんねん？
Polite	Nani shi te i masu ka?	何していますか？
Feminine	Nani shi te iru no?	何しているの？
Crude	Nani shi te yagaru?	何してやがる？

Table 6: Examples of six common utterance styles

Figure 1 depicts the relationships between daily conversation and fiction writing utterances through common particles and auxiliary verbs. The left side of the figure shows factors for daily conversation utterances, and right side shows factors for utterances in fiction writing. Common particles and auxiliary verbs are shown in the middle with lines connecting them to related factors.

In Figure 1, most of the factors have some words in common with a factor from the other side. However, "Aged style" in fictional writing has no common words with other factors. "Aged style" includes the auxiliary verb "Ja" and final particle "Nou." This style is often used when representing aged people in Japanese fiction, but it would be unrealistic utterance style to use in real life.

On the other hand, some words of "Serial style" and "Parallel style" used in daily conversation are combined in "Neutral style" in fiction writing. Moreover, some words that are included in "Frank style" in fiction writing are combined in "Neutral style" and "Feminine style" in daily conversation. Therefore, in daily conversation, "Neutral style" tends to be more "frank" (or forthright) than in fiction writing. This result may be influenced by the fact that the CEJC includes more family and friend situations than other situations (Table 5).

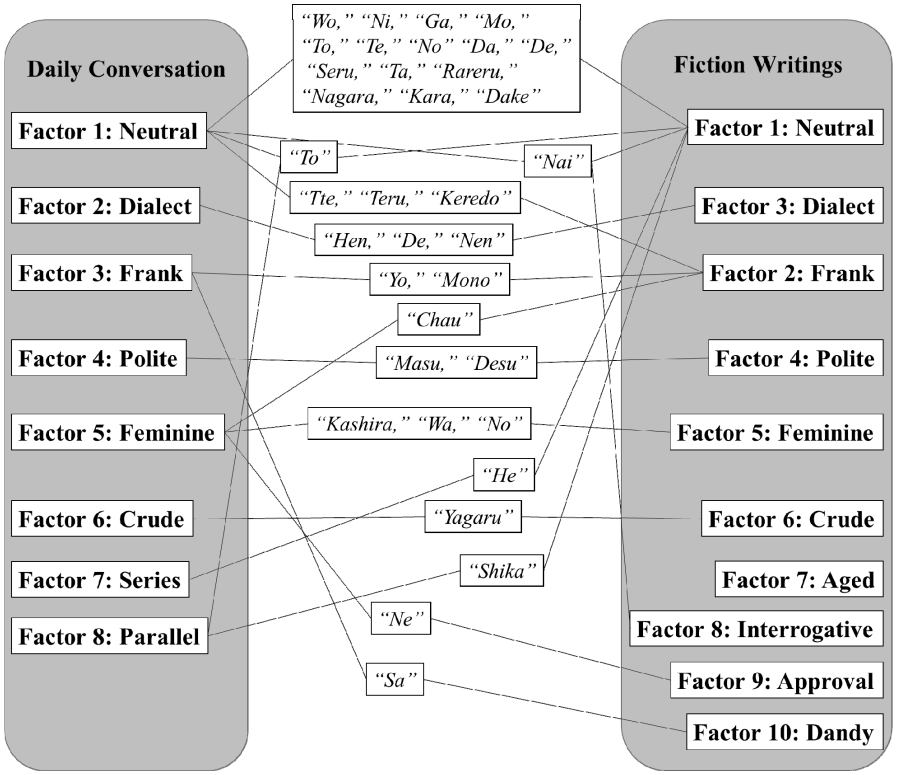


Figure 1: Relationship between daily conversation styles (CEJC) and fiction writing styles (BCCWJ)

5 Conclusion and Future Work

Japanese utterance styles in daily conversation were identified on the basis of a factor analysis for particles and auxiliary verbs in the CEJC. The relationships between these factors on the one hand and the speakers' attributes (ages and genders) and conversation settings on the other, were analyzed. As a result, eight factors were extracted: "Neutral style," "Dialect style," "Frank style," "Polite style," "Feminine style," "Crude style," "Series style," and "Parallel style".

In addition, a comparison with the utterance styles of characters in fiction writings was done. The results show that six factors are almost

identical between real daily conversation and fiction writings. However, “Neutral style” in daily conversation is more “frank” than in fiction writings. In addition, “Aged style” in fiction writing is a distinctly fictionalized or imaginary style that has no real-life use.

Because the utterances were tagged according to situation, some detailed characteristics of utterance styles were combined in the result of this research (in Table 5). If detailed relationships between speaker and listener for each utterance (e.g. parents to children, superior to subordinate, sales clerk to customer, or teachers to students) were to be added to the CEJC, more detailed utterance styles could be extracted.

Based on six fundamental utterance styles, it would be capable to generate more natural utterances automatically by utilizing natural language processing techniques in the future.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 26730168, the NINJAL collaborative research project ‘A Multifaceted Study of Spoken Language Using a Large-scale Corpus of Everyday Japanese Conversation', and the NINJAL project ‘Corpus of Everyday Japanese Conversation'.

References

Akiko Kurosawa. 2010. The sentence-final forms used in Meidai Dialogue Corpus: Does the plain style differ from the polite style? Yamagata University Working Papers in International Education, 2:3–11. (In Japanese)

Chiaki Miyazaki, Toru Hirano, Ryuichiro Higashinaka, Toshiro Makino, Yoshiro Matsuo, and Satoshi Sato. 2014. Fundamental Analysis of Linguistic Expression that Contributes to Characteristics of Speaker. In the Proceedings of the Association for Natural Language Processing, pp. 232−235. (In Japanese)

Efstathious Stamatatos. 2009. A Survey of Modern Authorship Attribution Methods. Journal of the Association for Information Science and Technology, 60(3): 538-556.

Hajime Murai. 2018A. Factor Analysis of Utterances in Japanese Fiction-writing Based on BCCWJ Speaker Information Corpus. Advances in Human-Computer Interaction, vol. 2018, Article ID 5056268, 9 pages.

Hajime Murai. 2018B. Factor Analysis of Japanese Daily Utterance Styles. LREC 2018 Joint Workshop LB-ILR2018 and MMC2018 Proceedings, 26-29.

Hanae Koiso, Tomoyuki Tsuchiya, Ryoko Watanabe, Daisuke Yokomori, Masao Aizawa, and Yasuharu Den. 2016. Survey of Conversational Behavior: Towards the Design of a Balanced Corpus of Everyday Japanese Conversation. Proceedings of the 10th edition of the Language Resources and Evaluation Conference, 4434-4439.

Hansen Andrew Schwartz, Johannes C. Eichstaedt, Margaret L. Kern, Lukasz Dziurzynski, Stephanie M. Ramones, Megha Agrawal, Achal Shah, Michal Kosinski, David Stillwell, Martin E. P. Seligman, and Lyle H. Ungar. 2013. Personality, Gender, and Age in the Language of Social Media: The Open-Vocabulary Approach. PLoS ONE, 8(9): 1-16.

Kazuma Seto, and Yoshiki Kishi. 2015. Construction of a Dialogue System Using a Speech Type of Estimation by Adjacency. Proceedings of Information Processing Society of Japan 2015, 131—132. (In Japanese)

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced corpus of contemporary written Japanese. Language Resources and Evaluation, 48(22): 345−371.

Raymond Shen, Hideaki Kikuchi, Katsumi Ohta, and Takeshi Mitamura. 2012. Towards the text-level characterization based on speech generation. Journal of Information Processing Society of Japan, 53(4):1269–1276. (In Japanese)

Rong Zheng, Jiexun Li, Hsinchun Chen, and Zan Huang. 2006. A Framework for Authorship Identification of Online Messages: Writing Style Features and Classification Techniques. Journal of the Association for Information Science and Technology, 57(3): 378-393.

Satoshi Kinsui. 2003. Virtual Japanese: Mystery of Functional Words. Iwanami Shoten, Tokyo. (In Japanese)

Shlomo Argamon, Moshe Koppel, Jonathan Fine, and Anat Rachel Shimoni. 2006. Gender, Genre, and Writing Style in Formal Written Texts. Text - Interdisciplinary Journal for the Study of Discourse, 23(3): 321-346.

Sumit Goswami, Sudeshna Sarkar, and Mayur Rustagi. 2009. Stylometric Analysis of Bloggers' Age and Gender. Proceedings of the Third International ICWSM Conference, 214-217.

A Speaker Accent Recognition System for Filipino Language

Batman Odulio*, Karl Adrian Cruz, Justin Raphael Ariaso, Mico Ian Orjalo
Angelica Dela Cruz, Ramon Rodriguez and Manolito Octaviano Jr.
College of Computing and Information Technologies
Manila, Philippines
National University
*oduliobc@students.national-u.edu.ph

Abstract

This paper presents the development of an accent recognition system for the native speakers of Bikol and Tagalog using deep learning. The results of the work serve as baseline for the advancement of recognizing speakers with Tagalog and Bikol accents in Filipino language. A monologue written in Filipino is prepared as script for the development of the speech corpus. The script is used to capture the Bikol accent and Tagalog accent in the recordings. The corpus was validated, cleaned and divided into 80:20 ratios for training and testing. Afterwards, Praat is utilized to analyze and extract prosodic features such as F1 and energy of speech. The model was tested and yields 79.28% and 78.33% accuracy for Tagalog and Bikol accent, respectively.

1 Introduction

Accent can be defined as the pronunciation style in a language. In the Philippine setup, the accent of a speaker can be highly influenced by other speakers in an approximate geographical location. For this reason, people can speak the same language but with a different accent resulting to a language (e.g. Filipino) used with multiple accents. This can provide information about the status, age, gender, dialect and ethnicity of a speaker when analyzed intensively (Tjalve, 2007). For the past years, the value of recognizing an accent of a speaker has begun to receive attention in the field of computing. Its influence was acknowledged as foundation in developing different large-scale speech applications such as automatic speech recognition (Petkar, 2016), (Zheng et al., 2005). However, automatic accent recognition is a challenging research task since a language can have multiple style pronunciation.

Automatic accent recognition, also known as accent identification, is based on the consistency of acoustic patterns that can be identified in speaking style that leads to the distinction of pronunciation on the same accent cluster. The work of (Lazaridis et al., 2014) categorized accent recognition system into foreign and regional. Foreign accent classification is characterized by the distinct difference in utterances of a foreign language as spoken by a non-native speaker. On the other hand, regional accent classification is characterized by the changes in pronunciation mainly in speaking styles among native speakers of the language.

There have been numerous research efforts to develop accent recognition system in different languages. However, as of the writing of this paper, the work of (Danao et al., 2017) is the only existing study that explored accent recognition in Philippine languages.

In order to expand the research of accent recognition, this paper focused on the development of accent recognition between Bikol and Tagalog speakers using Filipino monologue. The work shows a baseline work for recognizing speakers with Tagalog and Bikol accents in Filipino language.

2 Related Works

There are different efforts made to develop accent recognition in different regions using various approaches. Various models created from acoustic features, but also deep neural networks were explored. Gaikwad et al. (2013) focused on the English pronunciation of native speakers of Marathi and Arabic language. Acoustic features such as energy, pitch, and formant frequency were

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 511-515, Hakodate, Japan, September 13-15, 2019

extracted and used in the experiment. It was noted that formant frequency feature gives a promising result for accent of Marathi speakers while energy feature for Arabic speakers. In another study, Pham et al. (2016) explored the combination of Mel Frequency Cepstrum and F0 for Gaussian Mixture Model in recognizing Vietnamese dialects. Combining formants and bandwidths with normalized F0 boost the baseline dialect identification of the language from 58.6% up to 72.2%. While the study of Biadsy (2011) described a variety of approaches that make use of different acoustic features of a speech signal such as frame-based acoustic, phonetic, phonotactics features and high-level prosodic features in building a system that recognizes the regional dialect and accent of a speaker. The best approach of the study was tested in four broad Arabic dialects, ten Arabic subdialects, American English vs. Indian English accents, American English Southern vs. Non-Southern, American dialects at the state level plus Canada and three Portuguese dialects. The approach introduced by the study was able to achieve an Equal Error Rate (EER) of 4% for four broad Arabic dialects, an EER of 6.3% for American vs. Indian English accents, 14.6% for American English Southern vs. Non-Southern dialects, and 7.9% for three Portuguese dialects. To further test the approach, it was applied to an automatic speech recognition system that was significantly improved by 4.6%.

On the other hand, approaches based on deep neural networks were also explored. Astrid et al. (2017) analyzed speech accents on videogames using deep learning. The intuition is that characters from a videogame have traits, such as appearance and speech accent which can determine their characteristics. A model was trained using AlexNet to differentiate American, British, and Spanish accents in a videogame. Reported result shows a low accuracy of 61% that opens for various questions for further research extension. Similarly, Jiao et al. (2016) experimented on the fusion of deep neural networks and recurrent neural networks. The fusion of network performed better as compared to individual networks tested on a speech corpus with 45-second utterances.

3 Methodology

3.1 Data Collection

A 3-page monologue written in Filipino is prepared for the speakers that was examined by a Filipino linguist to ensure that the accent and emotion when read will be emphasized (Hatzidaki et al., 2015). Speech collection is done in Bikol University, Legazpi to collect Bikol accent and National University, Manila to collect Tagalog accent. The participants selected lived in the region for at least 10 years to ensure their Tagalog and Bikol accents. The speech was recorded in a closed room using Audacity[1], a free, open source, cross-platform audio software and headset with digital stereo sound and noise canceling microphone. A total of 106 Bikol and 51 Tagalog speech data were collected from each university.

3.2 Data Processing

A native speaker of the collected language is sought to validate the speech data. During the validation, it was noted that there exist five (5) variations of Bikol accent as shown in Table 1.

Accent Variations	Male	Female
Bikol-Buhi	1	4
Bikol-Daraga	14	10
Bikol-Legazpi	28	25
Bikol-Masbate	2	5
Bikol-Sorsogon	6	11
Tagalog	25	26

Table 1. Collected speech data

In order make the Bikol accent comparable to the Tagalog speech data, the dominant accent (Bikol-Legazpi) in the corpus is used. The speech data from the two languages were cleaned by removing the unnecessary background noises using noise profiling using Audacity. Speech signals were divided into frames to help in determining the uniqueness of a certain accent. Each sentence was slice into 42 parts. Praat, a free computer software package for the scientific analysis of speech, was used to extract prosodic features necessary for

[1] https://www.audacityteam.org/

training using 25ms frame length and 10ms window. The extracted features were F0, Mean Energy, Duration, Minimum Pitch, Maximum Pitch, Minimum Energy and Maximum Energy. Afterwards, extracted features from the speech data were split into two: 80% for training data and 20% for testing data.

3.3 Modelling

The features were extracted from the speech signals and fed in a 1-Dimensional Convolutional Neural Networks (1D-CNN) using Keras[2]. Below is the architecture of the model (see Figure 1 for the visualized neural network architecture):

- hidden layers: convolutional layer (4)
- activation function (hidden layers): ReLU
- dropout layer: 0.5
- dense layer (fully-connected layer): softmax activation function
- output layer: Sigmoid activation function

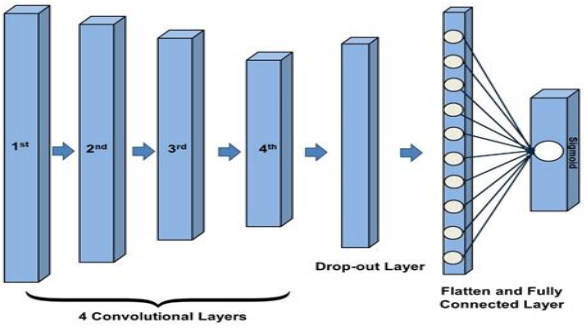


Figure 1. Convolutional Neural Network Architecture

Different parameters were explored by changing the values of output filters, batch size and epoch which yields different training and testing accuracy.

3.4 Feature Selection

The different features were experimented in varying combinations to find the suitable features that will be used to train the model of the accent recognition. These experiment setups were fed in the 1D-CNN. See Table 2 for the experiment setups used in the study.

Upon experimentation, the features that yielded highest accuracy is experiment #6 that has an accuracy of 82.89%, while the features that yielded the lowest accuracy is experiment #4 that has an accuracy of 63.27%. Experiment numbers 1-5 used the duration feature that affected the accuracy since the duration per data is different from one to another due to the uneven length of slice per data. Based on the previous studies and this experimentation, Energy and F0 were used as features for the accent recognition.

Experiment #	Features	Acc. (in %)
1	F0, Mean Energy, Duration, Minimum Pitch, Maximum Pitch, Minimum Energy and Maximum Energy.	76.75
2	F0, Mean Energy, Duration, Minimum Pitch, Maximum Pitch and Minimum Energy.	77.78
3	F0, Mean Energy, Duration, Minimum Pitch and Maximum Pitch.	77.31
4	F0, Mean Energy, Duration and Minimum Pitch.	63.27
5	F0, Mean Energy and Duration.	77.99
6	F0 and Mean Energy.	82.89

Table 2. Set of features for different experiments

3.5 Evaluation

The evaluation metrics used are accuracy and F1 score based from previous works in accent recognition. Accuracy is one of the evaluation metrics for assessing classification models appropriate to this study.

[2] https://keras.io/

4 Results and Discussion

Exp #	Output Filters	Batch Size	Epoch	Train Acc. (in %)	Testing Acc. (in %)
1	16/32	42	200	80	77
2	32/64	42	200	82	78
3	64/128	42	200	83	79
4	64/128	32	100	81	77
5	64/128	22	100	81	80
6	64/128	12	100	80	65
7	64/128	32	50	82	76
8	64/128	22	50	81	69
9	64/128	12	50	82	77

Table 3. Results of the experiments for parameter tuning

The convolutional neural network was trained using the determined hyperparameters (shown in table 3) to generate the accent classifier model. The generated model was then evaluated using the evaluation metrics. Table 4 shows the results of evaluation of the model.

Accent	Accuracy	F1 score	Recall	Precision
Tagalog	78.33	79.00	79.00	79.00
Bikol	79.28	78.50	78.00	79.00

Table 4. Results of evaluation

In addition, the researchers implemented the model by developing a user-friendly prototype in python that follows Input-Process-Output (IPO) scheme. The prototype is named "PARS: Philippine Accent Recognition System". Philippine Accent Recognition System (PARS) aims to distinguish the Accent of Bikol and Tagalog languages through utilizing the prosodic features of speech using the developed model developed. The researchers tested the prototype by having 840 testing data set and utilized the developed model and the result is as shown in the confusion matrix on Table 5.

Out of 420 Bikol speech data, the model correctly recognized 329 Bikol accent and out of 420 Tagalog speech data, the model correctly recognized 333 Tagalog accents. The results have shown the performance of the developed model and it reflects that the amount of correctly recognized input is more than the misrecognized input.

There are different factors that affect the performance of the model. The amount of data in this study is way less than the amount of data used by another study that also used deep learning in recognizing accents in speech.

It was also observed that during data cleaning or noise removal stage, words in low volume were removed in noise reduction activity.

Accent	Bikol	Tagalog
Bikol	329	91
Tagalog	87	333

Table 5. Confusion matrix

5 Conclusion and Future Work

A speaker accent recognition model for Filipino was developed using a 1D-CNN architecture. The study focused on two accents, Bikol and Tagalog. The model was tested and yields 79.28% and 78.33% accuracy for Tagalog and Bikol accent, respectively. The model was also implemented by developing a prototype named PARS. For future work, an increase in speech data and the inclusion of a wider scope of regions from the selected area is highly recommendable. The balance distribution of the participants with regards to age-group, gender and language can also be considered in order to improve the speaker accent recognition model. The researchers suggest to robust the model by considering the noise and making the model noise resistant. The researchers also suggest that in order to maximize the speech patterns and optimize the algorithm, adding more prosodic and other speech features can be explored.

References

Astrid Ensslin, Tejasvi Goorimoorthee, Shelby Carleton, Vadim Bulitko, Sergio Poo Hernandez 2017 Deep Learning for Speech Accent Detection in Videogames. In Thirteenth Artificial Intelligence and Interactive Digital Entertainment Conference.

Biadsy, F. 2011. Automatic dialect and accent recognition and its application to speech recognition Doctoral dissertation. Columbia University.

Glorianne Danao, Jolea Torres, Jamila Vi Tubio, and Larry Vea 2017. Tagalog Regional Accent

Classification in the Philippines. 2017 IEEE 9th International Conference on Humanoid, Nanotechnology, Information Technology, Communication and Control, Environment and Management (HNICEM).

Harshalata Petkar. 2016 A Review of Challenges in Automatic Speech Recognition, 151-No.3. International Journal of Computer Applications.

Hatzidaki, A., Baus, C., & Costa, A. (2015). The way you say it, the way I feel it: emotional word processing in accented speech. *Frontiers in psychology*, *6*, 351. doi:10.3389/fpsyg.2015.00351.

Michael Tjalve 2007. Accent features and idiodictionaries: on improving accuracy for accented speakers in ASR. Dissertation. University College London.

Lazaridis, Alexandros and Khoury, Elie and Goldman, Jean-Philippe and Avanzi, Mathieu and Marcel, S´ebastien and Garner, Philip N 2016 Swiss French regional accent identification. In Proceedings of odyssey 2014: The speaker and language recognition workshop.

PhamNgocHung,TrinhVanLoan,NguyenHongQuang 2016 Automatic identification of vietnamese dialects, V.32, N.1 (2016). Journal of Computer Science and Cybernetics.

Santosh Gaikwad, Bharti Gawali, Kale, K.V. 2013 Accent Recognition for Indian English using Acoustic Feature Approach International Journal of Computer Applications.

Yishan Jiao, Ming Tu, Visar Berisha, and Julie Liss 2016 Accent Identification by Combining Deep Neural Networks and Recurrent Neural Networks Trained on Long and Short Term Features Interspeech 2016.

Zheng, Yanli and Sproat, Richard and Gu, Liang and Shafran, Izhak and Zhou, Haolang and Su, Yi and Jurafsky, Daniel and Starr, Rebecca and Yoon, Su-Youn. 2005 Accent detection and speech recognition for shanghai-accented mandarin. Ninth European Conference on Speech Communication and Technology.

A corpus-based investigation of collexemes for active-passive alternation in the English part of an English-Japanese parallel corpus

Masanori Oya
Meiji University
masanori_oya2019@meiji.ac.jp

Abstract

This study conducted a corpus-based investigation of collexemes for active-passive alternation found in the English part of an English-Japanese parallel corpus as an attempt to use them as metrics for distinguishing native English and non-native English. The results show that some verbs in the data are used more often in the active voice than the passive voice, and vice versa, and the differences are statistically significant. However, these verbs are not the same as those found in a previous study. This fact supports the claim that active-passive alternation constitutes a lexico-semantic phenomenon that is sensitive to various factors, such as differences in genres and type of the authors of the text.

1. Introduction

This study conducts a corpus-based investigation of *collexemes* (Gries & Stefanowitsch, 2004; Stefanowitsch & Gries, 2003) for active-passive alternation found in the English part of an English-Japanese parallel corpus. Collexemes are a set of words that are attracted to certain types of *constructions* in the sense of the term used in Goldberg (1995) and Lakoff (1987). Collexemes are used in a certain construction more often than other words, and the difference between the frequency of their use and that of non-collexemes in the same construction is statistically significant. Investigation of collexemes in constructions is expected to facilitate a deeper understanding of the relationship between the lexicon and syntax. This is because such investigations allow us to have a critical viewpoint about the mainstream syntactic investigations of today, which presuppose that words are inserted into certain syntactic structures arbitrarily without considering other factors such as semantics of the word and discourse or genre of the text wherein the sentence is used. Rather, an investigation of collexemes is expected to suggest that the lexicon and syntax are closely related to each other, and different syntactic constructions necessarily attract certain words because of their semantic properties and other factors dependent on the characteristics of each construction.

This study is an attempt to use collexemes as one of the metrics for distinguishing native English and non-native English. If a collexeme in English texts generated by native speakers of English is found as non-collexeme in English texts generated by non-native speakers of English, then that collexeme indicates the difference between these two groups of speakers of English. In addition to this, the information of collexemes is expected to be of educational value; learners can learn different collexemes for different constructions and that will lead them to more natural use of the language.

The remainder of this paper is structured as follows: Section 2 reviews the rationale of collexeme investigations in contrast with collocational analyses. Section 3 reviews previous studies on collexemes, with special attention on the works of Gries and Stefanowitsch. Section 4 describes the data used in this study. The method, results, and discussion on the results are reported in Sections 5, 6, and 7, respectively. Lastly, Section 8 concludes.

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 516-521, Hakodate, Japan, September 13-15, 2019

2. Collocation and Collexemes

One important aspect of corpus linguistics is *collocational* analysis, in which the semantic and syntactic properties of a word or phrase are analyzed in terms of the context in which the word or phrase appears. Context here refers to the words before and after the word or phrase to be investigated, and they are called *collocates*. The span of collocates varies across different researchers with different research interests; for example, it is ±1 in Kennedy's study of *between* and *through* and ±5 in Church and Hunk's analysis of *doctor* (as cited in Stefanowitsch & Gries, 2004).

The problem with collocational analysis is that it only focuses on the linear order of the target word or phrase and its collocates, and it ignores their syntactic relationships. In this respect, Stefanowitsch and Gries (2003) point out that collocational analysis cannot capture the deeper relationship between the target word or phrase and the words associated with it, or the relationship between the target word and certain *construction pairs* that are considered as alternates. The so-called key word in context (KWIC) cannot capture the difference between construction pairs. For example, it cannot capture the context of where the target word appears in the double-object construction and where it appears in the direct-object-to-object construction (e.g., Sarah has given David some books vs. Sarah has given some books to David). In linguistics, these are called *dative alternations*, and it is virtually impossible to capture such alternations in KWIC, because they appear across words in such sentences, and the linear order of these words does not contain enough information to represent each alternation.

Stefanowitsch and Gries (2003) first introduced the idea of *collostructure* and applied it to corpus data, in order to overcome the shortcomings of KWIC-style research stated above. Their research aimed to apply the idea of *construction* (Goldberg, 1995; Lakoff, 1987) into the investigation of significant associative relationships between vocabulary and grammatical structure. They assumed that 1) lexicon and grammar are fundamentally similar and 2) multi-word expressions create links between the lexicon and grammar. Their assumptions can be paraphrased as follows: the so-called alternating constructions such as *active-passive alternations* (e.g., Sarah has broken some dishes vs. Some dishes have been broken by Sarah) and dative alternations do not alternate from one construction to the other, but they are actually two different constructions that are independent from each other. Stefanowitsch and Gries argue that this means these construction pairs should not be treated syntactically but lexico-semantically; some words are attracted to one of the construction pair, while others to the other. In other words, certain sets of verbs are used more often in the active voice, while another set of verbs are used more often in the passive voice, and the difference in their frequencies is statistically significant. Collexemes are such words that are attracted to certain constructions.

Inspired by their investigations, this study explores the possibility that English texts with limited focus on a topic contain certain collexemes for certain constructions. In particular, this study conducts a corpus-based investigation of collexemes for active-passive alternations found in the English part of an English-Japanese parallel corpus, which was constructed by translating a Japanese original text into English (the details are described in Section 4). These data are selected because it is expected that the collexemes for active-passive alternations reflect the characteristics of non-native English in terms of the collexemes for the alternation, which are different from the collexemes found through research using the corpus data generated by native speakers of English. As mentioned in the previous section, this study constitutes an attempt to use collexemes as one of the metrics for distinguishing native English and non-native English, with their educational value in mind.

3. Previous Studies

Stefanowitsch and Gries (2003) investigated constructions such as *cause N* (nouns that are attracted to the verb *cause*), X think nothing of V *gerund*, into-causative (e.g., Sarah tricked David into employing her), ditransitives, progressives, the imperatives, and past tense, and they found collexemes for each of these constructions. Based on the same assumption, Gries and Stefanowitsch (2004; G&S henceforth) conducted further research on the constructions' *active-passive alternations* and future tense as *will* and *be going*

to. They found that each construction is associated with a set of collexemes, and the association is so strong that there is a statistically significant difference between the frequency of these verbs in the active voice and the passive voice. The same is true for the pair of *will* and *going to*.

Investigations of collexemes are extended to the study of *semantic prosody*. Semantic prosody is a phenomenon where a certain word is associated with a positive or negative connotation because of its frequent occurrence with certain other words (Sinclair 1991). For example, Tang (2017) showed that the verb *cause* appears in various constructions which also contain words with negative semantic connotations, and therefore the verb *cause* has the tendency to be accompanied with negative semantic prosody.

4. Data

The corpus used in this study is the Japanese-English Bilingual Corpus of Wikipedia's Kyoto Articles, v.2.01 (National Institute of Information and Communications Technology, 2011). This corpus includes approximately 500,000 Japanese-English translation pairs of Wikipedia articles on 15 topics related to Kyoto, and each topic comprises one subcorpus. The articles are translated from the original Japanese text into English manually by Japanese translators, and then these are proofread by native English speakers. These translations are then edited by Japanese professionals, with special attention paid to the technical terms. This study uses the subcorpus of the topic related to Buddhism, which contains 26,890 Japanese-English translation pairs. These data are chosen with the assumption that English sentences translated from Japanese sentences are one of the genres of non-native English.

5. Method

In this study, the English sentences in the data are parsed by the Stanford Dependency Parser (de Marneffe & Manning, 2008), and the parsed results are used to calculate the number of verbs associated with their subject and object (active transitive verbs) and with their subject in the passive voice (passivized transitive verbs). We can count the number of passive verbs in the corpus by counting the number of dependency-type nominal subject of passivized verbs (NSUBJPASS in the parsed output) in the parse output through a simple regular-expression search. As for active verbs, on the other hand, we can determine their number by counting the dependency-type direct objects (DOBJ in the parsed output) in the same parse output through the same search method as for passive verbs. This means that this study ignores active verbs that are used without their direct objects with the assumption that they are used as intransitive verbs and therefore should not be counted as transitive verbs.

To show that the difference is larger than a coincidence between the probability that a verb *v* is used in the active voice in the corpus data and that all the verbs other than *v* are used in the active voice in the same corpus data, we conduct Fisher's exact test (1922, 1954), which was developed to examine the significance of the association between the two groups. This test has the following characteristics: it can be used when 1) the sample size is small and 2) the data are not distributed normally. This study uses this test because of these characteristics, as was the case in G&S.

In addition, to show exactly how large the difference between these two probabilities is, we calculate Cohen's *h* (Cohen 2013), which G&S did not. Cohen's *h* is employed to measure the differences between proportions in relation to hypothesis testing. The difference between two proportions is "statistically significant" when it seems that the population proportions are different. However, it is also possible that this difference can be too small to be meaningful. In other words, the "statistically significant" result does not indicate how large the size of the difference is. In this context, Cohen's *h* indicates the size of the difference and allows us to decide how meaningful the difference is.

Cohen's *h* is calculated in the following procedure. First, each probability is transformed through an "arcsine transformation" as follows:

$$\varphi = 2\arcsin\sqrt{p} \qquad (1)$$

When we have two probabilities, *p1* and *p2*, Cohen's *h* is the difference between their arcsine transformations:

$$h = \varphi 1 - \varphi 2 \qquad (2)$$

Cohen's *h* is interpreted as follows through a rule of thumb:

h = 0.20, "small effect size"; *h* = 0.50, "medium effect size"; *h* = 0.80, "large effect size."

In this study, $\varphi 1$ is the probability that a given verb *v* is used in the active voice, and $\varphi 2$ is the probability that all the verbs other than *v* is used in the active voice. For each verb in the data of this study, Fisher's exact test and Cohen's *h* are calculated by using js-STAR ver. 9.2.5j (http://www.kisnet.or.jp/nappa/software/star/freq/2x2.htm#). If Cohen's *h* is larger than 0.8 for a verb, the verb is more likely to be used in the active voice, while if it is smaller than -0.8, the verb is more likely to be used in the passive mood. If Cohen's *h* is between -0.2 and 0.2 for a verb, the verb has no preference of being used either in the active or passive voice. We ignored such verbs that appear less than 15 times in the data, either in the active or passive voice, so we can concentrate on frequently used verbs.

6. Results

This study found that the data contain 4,751 active verbs and 2,765 passive verbs. These total 7,516 verbs belong to 960 types, of which 306 are used in either the active or passive voice, 500 only in the active voice, and 154 only in the passive voice. Among the 806 types of active verbs (306+500), 56 are used more than 15 times in the corpus data, while among the 460 types of passive verbs (306+154), 34 are used more than 15 times in the same data.

The verbs used more often in the active voice than all the other verbs are listed in Table 1. Their Cohen's *h* is larger than 0.8, except for the verb "attain."

	Active	Passive	p	Cohen's *h*
have	240	0	** p <.01	1.467
enter	127	0	** p <.01	1.377
study	84	0	** p <.01	1.347
mean	43	0	** p <.01	1.320
follow	29	0	** p <.01	1.311
play	19	0	** p <.01	1.305
learn	70	1	** p <.01	1.099
visit	47	1	** p <.01	1.032
assume	38	1	** p <.01	0.995
receive	98	5	** p <.01	0.907
reach	33	2	** p <.01	0.829
attain	25	2	** p <.01	0.755

Table 1: Verbs used more often in active voice in the data

This table includes the verbs "have" and "mean"; they are also included in the result of G&S as these verbs tend to be used in active voice. On the other hand, this table does not contain all the other verbs that tend to be used in active voice in the result of G&S, since they are not frequent enough (used only 14 times or less in either the active or passive voice) or their Cohen's *h* is not larger than 0.8.

The verbs used more often in passive voice than all the other verbs are listed in Table 2. Their Cohen's *h* is lower than -0.8.

	Active	Passive	p	Cohen's *h*
say	9	176	** p <.01	-1.609
refer	1	26	** p <.01	-1.480
believe	5	51	** p <.01	-1.294
locate	4	42	** p <.01	-1.292
bear	10	85	** p <.01	-1.277
base	2	19	** p <.01	-1.239
know	16	105	** p <.01	-1.122
bury	3	17	** p <.01	-1.068
destroy	10	39	** p <.01	-0.947
think	5	20	** p <.01	-0.939
assign	5	18	** p <.01	-0.894

Table 2: Verbs used more often in passive voice in the data

This table includes the verbs "bear" and "base"; they are also included in the result of G&S as they tend to be used in the passive voice. However, this table does not contain all other verbs that tend to be used in passive voice in the result of G&S, since they are not used frequently enough in our data (used only 14 times or less in either the active or passive voice) or their Cohen's *h* is larger than -0.8.

This table includes the verbs "believe," "think," "say," and "know" as they are used more often in

the passive voice than the active voice. This result is in contrast with the result of G&S, wherein these verbs are used more often in the active voice than the passive voice.

The verbs whose Cohen's *h* falls between -0.2 and 0.2 are listed in Table 3; they are called "neutral" verbs.

	Active	Passive	p	Cohen's *h*
call	239	182	**p < .01	-0.213
found	51	38	ns	-0.140
describe	24	17	ns	-0.108
write	55	34	ns	-0.041
name	19	11	ns	-0.007
show	21	11	ns	0.042
grant	22	11	ns	0.064
confer	24	10	ns	0.152
send	24	10	ns	0.152
put	22	9	ns	0.160
preach	18	7	ns	0.182
give	123	49	*p < .05	0.193

Table 3: Neutral verbs in the data

None of the verbs in Table 3 are included in the result of G&S.

7. Discussion

This study found that the data include verbs that are used in the active voice more often than the passive voice, and vice versa. This finding suggests that the active-passive alternation is not a purely syntactic phenomenon but rather a lexical-semantic one. The same result was obtained by G&S.

However, the list of the verbs in this study is not identical with that of G&S; although there are some similarities ("have" and "mean" in active voice and "bear" and "base" in passive voice), all the other verbs in Tables 1 and 2 are not included in their study. In addition, we can find contradictory cases between their study and ours as some verbs ("believe," "think," "say," and "know"), which are used in the active voice in their study, are used more often in the passive voice in ours.

This discrepancy is surely the result of different foci on which verbs should be considered in G&S and our study: G&S focused on all the verbs in their data, while we focused on only some frequently used verbs in our data. In addition, the corpus they used contains a variety of genres of text written by native English speakers, while our data contains definitions of terms in a limited area of interest (Buddhism) translated from Japanese into English by non-native English speakers and edited by native speakers.

It can be argued that this discrepancy between G&S and our study supports the claim that the active-passive alternation constitutes a lexico-semantic phenomenon. That is, the difference in text genres is reflected by which verbs tend to be used more often in the active voice than the passive voice. In particular, the verbs "believe," "think," "say," and "know" are used in passive voice, because their passive constructions can express situations wherein a story or incident is accepted by the general public (e.g., "it is believed that...," and "it is said that..."). Moreover, it is natural that these expressions are used more frequently than usual, as the aim of the texts in our data is to provide an introduction to a historical person or historical incident. In addition, we cannot ignore the influence of Japanese phrases that use passive voice verbs such as "*...to shinjirareteiru*" (It is believed that...), "*...to kangaerareteiru*" (It is thought that...), "*...to iwareteiru*" (It is said that...)," and "*...to shirareteiru*" (It is known that...). In future research, we aim to identify such constructions in English translations of Japanese texts that are used more often than usual (possibly) because of the influence of the original, or in English sentences produced by non-native speakers of English, such as Japanese learners of English.

The observation of these passive voice verbs with the possible influence of the original Japanese sentences seems to support the assumption mentioned in Section 4 above that English sentences translated from Japanese sentences are one of the genres of non-native English.

The observation of these passive voice verbs also seems to argue against the claim that the genre of corpus used in this study cannot be employed to address the issue of distinguishing native English and non-native English; that is, the corpus data in this study are English sentences translated from Japanese sentences with proofreading by native speakers of English, and therefore they can be less non-nativelike than other "pure" non-native English sentences, such as essays written by Japanese learners of English. However, the proofreading by native speakers of English does not necessarily render English sentences as nativelike as possible, and therefore they cannot be

"pure" native English sentences, as other types of English sentences produced by non-native speakers of English.

In this context, though, it will be productive to explore the possibility of finding more supportive results through the investigation of collexemes in the corpus data produced by non-native learners of English, with the same method as this study. This will be the goal of future research.

To support the claim that active-passive alternation constitutes a lexico-semantic phenomenon, we need to explain that some verbs can alternate between the active and passive voice without any bias toward either. G&S did not address this issue, since they only reported verbs that are distinctively biased toward the active or passive voice. As reported in Table 3, we found that some verbs are used either in the active or passive, and there is no significant difference between these two usages as far as our data is concerned. This may support the claim that active-passive alternation constitutes a syntactic phenomenon, and any bias toward the active or passive cannot be found, at least for these verbs. In this context, the behaviors of these verbs, which are found unbiased in our data, need to be investigated in different corpora or subcorpora of the same corpus, so that we may verify the possibility that these verbs can also show a tendency to be used in either the active or passive voice. This will reflect the particular characteristics of the corpus data, which will be a research question of future studies.

8. Conclusion

This study conducted a corpus-based investigation of collexemes for active-passive alternation found in the English part of an English-Japanese parallel corpus, as an attempt to use them as metrics for distinguishing native English and non-native English. The results show that some verbs in the data are used in the active voice more often than the passive voice, and vice versa, and the differences are statistically significant. However, these verbs are not the same as those found in a previous study. This fact supports the claim that active-passive alternation constitutes a lexico-semantic phenomenon that is sensitive to various factors, such as differences in genres and type of the authors of the text (e.g., native speakers vs. non-native speakers). Moreover, some verbs are neutral to the alternation, which will be addressed in future studies on the relationships between collexemes and constructions.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number 17K02740.

References

Cohen, J. 2013. Statistical power analysis for the behavioral sciences (2nd ed.). Abington, UK: Routledge.

De Marneffe, M.C., & Manning, C. 2008. The Stanford typed dependencies representation. Proceedings CrossParser '08 Coling 2008: *Workshop on Cross-Framework and Cross-Domain Parser Evaluation*. Manchester, UK: Association for Computational Linguistics.

Fisher, R. A. 1922. On the interpretation of χ^2 from contingency tables, and the calculation of P. Journal of the Royal Statistical Society, *85*(1): 87–94.

Fisher, R. A. 1954. Statistical methods for research workers (12th ed.). Edinburg, UK: Oliver and Boyd.

Goldberg, A. 1995. Constructions: A construction grammar approach to argument structure. Chicago, IL: University of Chicago Press.

Gries, S., & Stefanowitsch, A. 2004. Extending collostructural analysis: A corpus-based perspective on 'alternations'. International Journal of Corpus Linguistics, *9*(1): 97–129.

Lakoff, G. 1987. Women, fire, and dangerous things. Chicago, IL: University of Chicago Press.

National Institute of Information and Communications Technology. 2011. The Japanese-English bilingual corpus of Wikipedia's Kyoto articles, v.2.01. Retrieved from https://alaginrc.nict.go.jp/WikiCorpus/index_E.html

Stefanowitsch, A., & Gries, S. 2003. Collostructions: Investigating the interaction of words and constructions. International Journal of Corpus Linguistics, 8(2): 209–243.

Sinclair, J. M. 1991. Corpus, Concordance, Collocation. Oxford: OUP

Tang, X. 2017. Lexeme-based collexeme analysis with DepCluster. Corpus Linguistics and Linguistic Theory, 13(1): 165–202

Korean-to-Chinese Machine Translation using Chinese Character as Pivot Clue

Jeonghyeok Park[1,2,3] and Hai Zhao[1,2,3,*]

[1]Department of Computer Science and Engineering, Shanghai Jiao Tong University
[2]Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University, China
[3] MoE Key Lab of Artificial Intelligence AI Institute, Shanghai Jiao Tong University
117033990011@sjtu.edu.cn, zhaohai@cs.sjtu.edu.cn

Abstract

Korean-Chinese is a low resource language pair, but Korean and Chinese have a lot in common in terms of vocabulary. Sino-Korean words, which can be converted into corresponding Chinese characters, account for more then fifty of the entire Korean vocabulary. Motivated by this, we propose a simple linguistically motivated solution to improve the performance of Korean-to-Chinese neural machine translation model by using their common vocabulary. We adopt Chinese characters as a translation pivot by converting Sino-Korean words in Korean sentence to Chinese characters and then train machine translation model with the converted Korean sentences as source sentences. The experimental results on Korean-to-Chinese translation demonstrate that the models with the proposed method improve translation quality up to 1.5 BLEU points in comparison to the baseline models.

1 Introduction

Neural machine translation (NMT) using sequence-to-sequence structure has achieved remarkable performance for most language pairs (Bahdanau et al., 2014; Cho et al., 2014; Sutskever et al., 2014; Luong and Manning, 2015). Many studies on NMT have tried to improve the translation performance by changing the structure of the network model or adding new strategies (Wu and Zhao, 2018; Zhang et al., 2018; Xiao et al., 2019). Meanwhile, there are few attempts to improve the performance of the NMT model using linguistic characteristics for several language pairs (Sennrich and Haddow, 2016). On the other hand, Most of the recently proposed statistical machine translation (SMT) systems have attempted to improve translation performance by using linguistic features including part-of-speech (POS) tags (Ueffing and Ney, 2013), syntax (Zhang et al., 2007), semantics (Rafael and Marta, 2011), reordering information (Zang et al., 2015; Zhang et al., 2016) and so on.

In this work, we focus on machine translation between Korean and Chinese, which have few parallel corpora but share a well-known culture heritage, the Sino-Korean words. Chinese loanwords used in Korean are called Sino-Korean words, and can also be written in Chinese characters which are still used by modern Chinese people. Such a shared vocabulary makes the two languages closer despite their huge linguistic difference and provides the possibility for better machine translation.

Because of its long history of contact with China, Koreans have used Chinese characters as their writing system, and even after adopting Hangul(한글 in Korean) as the standard language, Chinese characters have a considerable influence in Korean vocabulary. Currently, the writing system adopted by modern Korean is Hangul, but Chinese characters continue to be used in Korean and Chinese characters used in Korean are called ”Hanja”. Korean vocabulary can be categorized into native Korean words, Sino-Korean words, and loanwords from other languages. The Sino-Korean vocabulary refers to Ko-

* Corresponding author. This paper was partially supported by National Key Research and Development Program of China (No. 2017YFB0304100) and Key Projects of National Natural Science Foundation of China (U1836222 and 61733011).

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 522-530, Hakodate, Japan, September 13-15, 2019

Systems	Sentences
Korean	명령은 아래와 같이 반포되었다.
HH-Convert	命令은 아래와 같이 颁布되었다.
Chinese	命令颁布如下。
English	The command was promulgated as follows.
Korean	양국은 광범한 영역에서의 공동 이익을 확인했다.
HH-Convert	两国은 广范한 领域에서의 共同 利益을 确认했다.
Chinese	两国在广泛的领域确认了共同利益。
English	The two countries have confirmed common interests in a wide range of areas.

Table 1: The HH-Convert is Korean sentence converted by Hangul-Hanja conversion of the Hanjaro. The underline denotes Sino-Korean word and its corresponding Chinese characters in Korean sentence and HH-Convert sentence, respectively.

rean words of Chinese origin and can be converted into corresponding Chinese characters, and considerably account for about 57% of Korean vocabulary. Table 1 shows some sentence pairs of Korean and Chinese with the converted Sino-Korean words. In Table 1, some Chinese words are commonly observed between the converted Korean sentence and the Chinese sentence.

In this paper, we present a novel yet straightforward method for better Korean-to-Chinese MT by exploiting the connection of Sino-Korean vocabulary. We convert all Sino-Korean words in Korean sentences into Chinese characters and take the converted Korean sentences as the updated source data for later MT model training. Our method is applied to two types of NMT models, recurrent neural network (RNN) and the Transformer, and shows significant translation performance improvement.

2 Related Work

There have been studies of linguistic annotation, such as dependency label (Wu et al., 2018; Li et al., 2018a; Li et al., 2018b), semantic role labels (Guan et al., 2019; Li et al., 2019) and so on. Sennrich and Haddow (2016) proved that various linguistic features can be valuable for NMT. In this work, we focus on the linguistic connection between Korean and Chinese to improve Korean-to-Chinese NMT.

There are several studies on Korean-Chinese machine translation. For example, Kim et al. (2002) proposed verb-pattern-based Korean-to-Chinese MT system that uses pattern-based knowledge and consistently manages linguistic peculiarities between language pairs to improve MT performance. Li et al. (2009) improved the translation quality for Chinese-to-Korean SMT by using Chinese syntactic reordering for an adequate generation of Korean verbal phrases.

Since Chinese and Korean belong to entirely different language families in terms of typology and genealogy, many studies also tried to analyze sentence structure and word alignment of the two languages and then proposed the specific methods for their concern (Huang and Choi, 2000; Kim et al., 2002; Li et al., 2008). Lu et al. (2015) proposed a method of translating Korean words into Chinese using the Chinese character knowledge.

There are several attempts to exploit the connection between the source language and the target language in machine translation. Kuang et al. (2018) proposed methods to somewhat shorten the distance between the source and target words in NMT model, and thus strengthen their association, through a technique bridging source and target word embeddings. For other low-resource language pairs, using pivot language to overcome the limitation of the insufficient parallel corpus has been a choice (Habash and Hu, 2009; Zahabi et al., 2013; Ahmadnia et al., 2017). Chu et al. (2013) bulid a Chinese character mapping table for Japanese, Traditional Chinese, and Simplified Chinese and verified the effectiveness of shared Chinese characters for Chinese–Japanese MT. Zhao et al. (2013) used the Chinese character, a common form of both languages, as a translation bridge in the Vietnamese-Chinese SMT model, and improved the translation quality by con-

北 선전매체 “北美관계도 “南北관계처럼 대전환”
3.1운동 100주년 맞아 장병 어깨에 원색(原色) 태극기 부착

Table 2: News headlines with Chinese characters. The underline denotes Chinese characters.

verting Vietnamese syllables into Chinese characters with a pre-specified dictionary. Partially motivated by this work, we turn to Korean in terms of NMT models by fully exploiting the shared Sino-Korean vocabulary between Korean and Chinese.

3 Sino-Korean Words and Chinese Characters

Korea belongs to the Chinese cultural sphere, which means that China has historically influenced regions and countries of East Asia. Before the creation of Hangul (*Korean alphabet*), all documents were written in Chinese characters, and Chinese characters were used continuously even after the creation of Hangul.

Today, the standard writing system in Korea is Hangul, and the use of Chinese characters in Korean sentences is rare, but Chinese characters have left a significant influence on Korean vocabulary. About 290,000 (57%) out of the 510,000 words in the *Standard Korean Language Dictionary* published by the *National Institute of Korean Language* belongs to Sino-Korean words, which were originally written in Chinese characters. Some Sino-Korean words do not currently have corresponding Chinese words and their meanings and usage have changed in the process of introduction, but most of them have corresponding Chinese words. In Korean, Sino-Korean words are mainly used as literary or technical vocabulary and are often used in abstraction concepts and technical terms. The names of people and Korea place are mostly composed of Chinese characters, and newspapers and professional books occasionally use both Hangul and Chinese characters to clarify the meaning. Table 2 shows some news headlines that contain Chinese characters from the Korean news.

Since Korean belongs to alphabetic writing systems and is a language that does not have tones like Chinese, many homophones were created in their vocabulary in the process of translating the Chinese words into their language. Around 35% of the Sino-Korean words registered in the *Standard Korean Language Dictionary* belong to homophones. Thus converting Sino-Korean words into (usually different) Chinese characters will have a similar impact as semantic disambiguation. For example, the Korean word uisa (의사 in Korean) has many homophones and can have several meanings. To clarify the meaning of the word uisa in Korean context, these words are occasionally written in Chinese characters as follows: 医师 (*doctor*), 意思 (*mind*), 义士 (*martyr*), 议事 (*proceedings*).

In addition, There is a difference between Chinese characters (Hanja) used in Korea and Chinese characters used in China. Chinese can be divided into two categories: Traditional Chinese and Simplified Chinese. Chinese characters used in China and Korea are Simplified Chinese and Traditional Chinese, respectively.

4 The Proposed Approach

The proposed approach for Korean-to-Chinese MT has two phases: Hangul-Hanja conversion and NMT model training. We first convert the Sino-Korean words of the Korean input sentences into Chinese characters, and convert the Traditional Chinese characters of the converted Korean input sentences into Simplified Chinese characters to share the common units between source and target vocabulary. Then we train NMT models with the converted Korean sentences as source data and the original Chinese sentences as target data.

For Hangul-Hanja conversion, we use open toolkit Hanjaro that is provided by the *Institute of Traditional Culture*[1]. The Hanjaro can accurately convert Sino-Korean words into Chinese characters and is based on open toolkit UTagger (Shin and Ock (2012) in Korean) developed by the *Korean Language Processing Laboratory* of *Ulsan University*. More specifically, the Hanjaro first obtains tagging information about morpheme, parts of speech(POS) and homophones of a Korean sentence through the Utagger, and converts Sino-Korean words into corresponding Chinese characters by using this tagging

[1]https://hanjaro.juntong.or.kr

Domains	Train	Validation	Test
Society	67363	2,000	2,000
All	258386	5,000	5,000

Table 3: The statistics for the parallel corpus extracted from Dong-A newspaper (The number of sentences).

information and pre-built dictionary. The UTagger is the Korean morphological tagging model which has a recall of 99.05% on morpheme analysis and 96.76% accuracy on POS and homophone tagging. Nguyen et al. (2019) significantly improved the performance Korean-Vietnamese NMT system by building a lexical semantic network for the special characteristics of Korean, which is using a knowledge base of the UTagger, and applying the Utagger to Korean tokenization.

For MT modeling, we use two types of NMT models: RNN based NMT and Transformer NMT models. We train the NMT models on parallel corpus processed through the Hangul-Hanja conversion above.

5 Experiments

There have been many studies on how to segment Korean and Chinese text (Zhao and Kit, 2008a; Zhao and Kit, 2008b; Zhao et al., 2013; Cai and Zhao, 2016; Deng et al., 2017). To find out which segmentation method has the highest translation performance, we tried multiple segmentation strategies such as byte-pair-encoding (Sennrich et al., 2016), jieba[2] , KoNLP[3] and so on. Eventually, we found that character-based segmentation for both languages can give the best performance. Therefore, both Korean and Chinese sentences are segmented into characters for our NMT models.

5.1 Parallel Corpus

We use two parallel corpora in our experiment. The first corpus is a Chinese-Korean parallel corpus of casual conversation and provided by *Semantic Web Research Center*[4] (SWRC). However, the SWRC corpus contained some incomplete data, so we removed the erroneous data manually. The parallel corpus consists of a set of 55,294 pairs of parallel sentences. 2,000 and 2,000 pairs from the parallel corpus were extracted as validation data and test data, respectively.

The second corpus (Dong-A) is collected from the online Dong-A newspaper[5] by us. We collected articles on four domains, Economy (81,278 sentences), Society (71,363), Global (68,073) and Politics (61,208), to build two corpora as shown in Table 3.

Since the sentences in the Dong-A newspaper are relatively long, the maximum sequence length that we used to train the NMT model is set to 200. On the other hand, the maximum sequence length for SWRC corpus is set to 50 because each sentence in the SWRC corpus is short.

5.2 NMT Models

The Torch-based toolkit OpenNMT (Klein et al., 2018) is used to build our NMT models, either RNN-based or Transformer.

As for RNN-based models, we further consider two types of them, one with unidirectional LSTM encoder (uni-RNN) and the other with bidirectional LSTM based encoder (bi-RNN). For both RNN based models, we use 2-layer LSTM with 500 hidden units on both encoder and decoder and use the global attention mechanism as described in (Luong et al., 2015). We use stochastic gradient descent (SGD) optimizer with the initial learning rate 1 and with decay rate 0.5. Mini-batch size is set to 64, and the dropout rate is set to 0.3.

For our Transformer model, both the encoder and decoder are composed of a stack of 6 uniform layers, each built of two sublayers as described in (Vaswani et al., 2017). The dimensionality of all input and output layers is set to 512, and that of Feed-Forward Networks (FFN) layers is set to 2048. We set the source and target tokens per batch to 4096. For optimization, we used Adam optimizer (Kingma and Ba, 2014) with β_1= 0.9, β_2= 0.98 to tune model parameters, and the learning rate is set by the warm-up strategy with steps 8,000 ,and it decreases proportionally as the model training progresses.

All of the NMT models are trained for 100,000

[2]https://pypi.org/project/jieba/

[3]http://konlpy.org

[4]http://semanticweb.kaist.ac.kr

[5]http://www.donga.com/ (Korean) and http://chinese.donga.com/ (Chinese)

Systems	BLEU Score (Test set)	
	w/o HH-Conv.	w/ HH-Conv
uni-RNN	33.14	34.44
bi-RNN	35.31	36.66
Transformer	35.47	37.84

Table 4: Experimental results of SWRC corpus. The HH-Conv refers to Hangul-Hanja conversion function.

Systems	Domains	BLEU Score	
		w/o HH-c.	w/ HH-c
uni-RNN	Society	36.25	37.58
	All	39.84	40.70
bi-RNN	Society	39.08	40.00
	All	41.76	42.81
Transformer	Society	39.34	40.55
	All	44.70	44.88

Table 5: Experimental results of Dong-A corpus.

steps and checked the performance on the validation set after every 5,000 training steps. And we save the models every 5,000 training steps and evaluate the models using traditional machine translation evaluation metric.

5.3 Results

We used the BLEU score (Papineni et al., 2002) as our evaluation metric. Tables 4 and 5 show the experimental results for SWRC corpus and Dong-A corpus, respectively. All NMT models, trained with Korean sentences converted through Hangul-Hanja conversion as source sentences, improve the translation performance on all test sets in comparison to the NMT models for the original sentence pairs. The absolute BLEU improvement is about 1.57 on average for SWRC corpus and 0.93 on average for Dong-A corpus when applied the Hangul-Hanja conversion, respectively.

Our proposed method is to improve the translation performance of NMT models by converting only Sino-Korean words into corresponding Chinese characters in Korean sentences using the Hanjaro and sharing the source vocabulary and the target vocabulary.

In the work, we do not convert the entire Korean sentence into Chinese characters using a pre-specified dictionary and maximum matching mechanism as described in (Zhao et al., 2013). Unlike Chinese, which does not use inflectional morphemes, Korean belongs to an agglutinative language that tends to have a high rate of affixes or morphemes per word. Since some Korean syllables do not have corresponding Chinese characters, so converting all Korean syllables of Korean sentence into Chinese characters is an impossible mission. In fact, we built a bilingual dictionary for Korean and Chinese and used maximum matching mechanism to convert all the affixes and inflectional morphemes of Korean sentences into Chinese characters and trained an RNN based NMT model, but the performance was even lower.

In our implementation, we estimate that the main reason for improving performance is to make the distinction between homophones clearer by converting Sino-Korean words into Chinese characters. Many of the Korean vocabularies that employ the alphabetical writing system are homophones, which can confuse meaning or context. Especially, as mentioned in Section 3, 35% of Sino-Korean words are homophones. Therefore, it is possible to clarify the distinction between homophones by applying Hangul-Hanja conversion to Korean sentences, which leads to performance improvement in Korean-to-Chinese MT.

6 Analysis

6.1 Analysis on Sino-Korean word Conversion

In this subsection, we will analyze the conversion from Sino-Korean words to Chinese characters. To estimate how much Chinese characters converted from Sino-Korean words by Hangul-Hanja conversion function are included in the corresponding reference sentence, we propose *ratio of including the same Chinese character between the converted Korean sentence and Chinese sentence (reference sentence)* (ROIC):

$$ROIC = \frac{\sum_{w_i} f(w_i)}{|w|} \quad (1)$$

where $|w|$ is the number of Chinese words in converted Korean sentence, $f(w_i)$ is 1 if the Chinese word w_i of the converted Korean sentence is included in the corresponding Chinese sentence, and

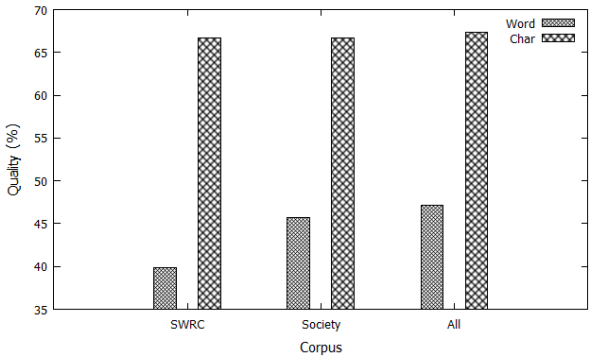


Figure 1: ROIC of each corpus. Word and Char denote the ROIC for Chinese word and the ROIC for Chinese character, respectively.

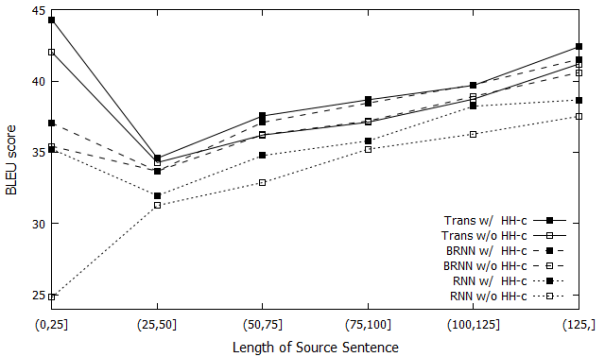


Figure 2: BLEU scores for the translation of sentences with different lengths.

0 otherwise. For example, in the second example of Table 1, because the five Chinese words such as 两国 (*two countries*), 领域 (*area*), 共同 (*common*), 利益 (*interests*), 确认 (*confirm*) are commonly observed between the converted Korean sentence and the reference sentence except for 广范 (*abroad*), so we say that the ROIC of the converted Korean sentence is $\frac{5}{6}$ (83.33%). We perform analysis of Sino-Korean word conversion in two separate ways: ROIC for Chinese word and ROIC for Chinese character.

Fig. 1 presents the ROIC of each corpus. It can be observed that for each corpus, more than 40% of the converted Chinese words or more than 65% of the converted Chinese characters are included in the reference sentence. So we can see that source vocabulary and target vocabulary share many words after converting Sino-Korean words into Chinese characters. Sharing source vocabulary and target vocabulary is especially useful for same alphabet languages, or for domains where professional terms are written in English (Zhang et al., 2018). Therefore, we set to share the source vocabulary and the target vocabulary of our NMT models, which leads to performance improvement.

6.2 Analysis of Translation Performance according to Different Sentence Lengths

Following Bahdanau et al. (2017), we group sentences of similar lengths together and compute BLEU scores, which are presented in Fig. 2. we conduct this analysis on Society corpus. It shows that our method leads to better translation performance for all the sentence lengths. Since we set the Maximum sentence length to 200 for the Society corpus, we also can see that the performance continues to improve when the length of the input sentence increases.

6.3 Analysis of Homophones Translation

In this subsection, we translate several sentences that contain two homophones and analyze how the Sino-Korean word conversion makes the distinction between homophones more apparent. We translated the sentences using the Transformer model trained with the Dong-A corpus. Table 6 presents the translation results of sentences with two homophones.

We can see that our NMT model clearly distinguishes between homophones for all examples, but the baseline model does not distinguish or translate homophones. For example, in the first example, the baseline model does not translate 유지* (*community leader*). In the second and third example, the baseline model translated them into the same words without distinguishing between the homophones. In the last example, 의사** (*wishes*) was improperly translated into 意向 (*intention*). Therefore, as mentioned in Section 5.3, these results indicate that our method helps distinguish homophones in Korean-to-Chinese machine translation.

7 Conclusion

This paper presents a simple novel method exploiting the shared vocabulary of a low-resource language pair for better machine translation. In detail, we convert Sino-Korean words in Korean sen-

Systems	Sentences
Korean	이 지역에 사는 유지*들이 이 마을을 유지**하고 관리해나가고 있다.
HH-Convert	이 地域에 사는 有志*들이 이 마을을 维持**하고 管理해나가고 있다.
Chinese	在这个区域生活的有志之士*在维护**和管理这个小区。
English	The community leaders* living in this area are maintaining** and managing this community.
Trans w/o HH-c	居住在该地区的维持**和管理村庄。
Trans w/ HH-c	居住在该地区的有志*们维持**这个村子，并进行管理。
Korean	이성* 간의 교제는 이성**에 따라 해야 한다.
HH-Convert	异性* 间의 交际는 理性**에 따라 해야 한다.
Chinese	异性*之间交往应该保持理性**。
English	A romantic relationship between the opposite sex* should be rational**.
Trans w/o HH-c	理性**间的交往应遵从理性**。
Trans w/ HH-c	异性*之间的交往应该根据理性**进行。
Korean	그는 천연자원*을 탐사하는 임무에 자원**했다.
HH-Convert	그는 天然资源*을 探查하는 任务에 自愿**했다.
Chinese	他自愿**参加勘探自然资源**的任务。
English	He volunteered** for the task of exploring natural resources*.
Trans w/o HH-c	他为探测天然资源**的任务提供了资源**。
Trans w/ HH-c	他自愿**担任探测天然资源*的任务。
Korean	의사*의 꿈은 포기했지만, 가족들은 그의 의사**를 존중해주었다.
HH-Convert	医师*의 꿈은 抛弃했지만, 家族들은 그의 意思**를 尊重해주었다.
Chinese	虽然放弃了医生*的梦想,但家人也尊重他的意愿**。
English	Although he gave up on his dream of becoming a doctor*, his family respected his wishes**.
Trans w/o HH-c	虽然医生*的梦想放弃了，但是家人却尊重了他的意向。
Trans w/ HH-c	虽然放弃了医生*的梦想，但家人却尊重了他的意愿**。

Table 6: Translation results of sentences with two homophones. The HH-Convert is Korean sentence converted by Hangul-Hanja conversion of the Hanjaro. Trans w/o HH-c and Trans w/ HH-c are the translation results of Transformer baseline model and Transformer using our method, respectively. The underline denotes homophone and the number of stars(*) distinguishes the meanings of the homophone in each example. In Chinese, English, and translation results, they denote words that are equivalent to the homophones in the sense of meaning.

tences into Chinese characters and then train machine translation model with the converted Korean sentences as source sentences.Our proposed improvement has been verified effective over RNN-based and latest Transformer NMT models. Besides, we regard that this is the first attempt which takes a linguistically motivated solution for low-resource translation using NMT models. Although this proposed method seems only suitable for the language pair of Korean and Chinese, it has enormous potential to work for any language pair which shares a considerable vocabulary from their shared history.

References

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser and Illia Polosukhin. 2017. Attention Is All You Need. *arXiv preprint arXiv:1706.03762*.

Benyamin Ahmadnia, Javier Serrano and Gholamreza Haffari. 2017. Persian-Spanish Low-Resource Statistical Machine Translation Through English as Pivot Language. *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP 2017)*, page 24–30.

Changhyun Kim, Young Kil Kim, Munpyo Hong, Young Ae Seo, Sung Il Yang and Sung-Kwon Choi. 2002. Verb Pattern Based Korean-Chinese Machine Transla-

tion System. *Proceedings of the First SIGHAN Workshop on Chinese Language Processing*, page 157–165.

Chenhui Chu, Toshiaki, Nakazawa, Daisuke, Kawahara and Sadao Kurohashi. 2013. Chinese-Japanese Machine Translation Exploiting Chinese Characters. *ACM Transactions on Asian Language Information Processing*, volume 12. page 1–25.

Chaoyu Guan, Yuhao Cheng and Hai Zhao. 2019. Semantic Role Labeling with Associated Memory Network. *2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, page 3361–3371.

Deng Cai, Hai Zhao, Zhisong Zhang, Yuan Xin, Yongjian Wu and Feiyue Huang. 2017. Fast and Accurate Neural Word Segmentation for Chinese. *ACL 2017*, page 608–615.

Deng Cai and Hai Zhao. 2016. Neural Machine Translation of Rare Words with Subword Units. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*, page 409–420

Dongdong Zhang, Mu Li, Chi-Ho Li and Ming Zhou. 2007. Phrase Reordering Model Integrating Syntactic Knowledge for SMT. *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Dong-il Kim, Zheng Cui, Jinji Li and Jong-Hyeok Lee. 2002. A Knowledge Based Approach to Identification of Serial Verb Construction in Chinese-to-Korean Machine Translation System. *Proceedings of the First SIGHAN Workshop on Chinese Language Processing*, volume 18, page 1–7.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. *arXiv preprint acXiv:1412.6980*.

Dzmitry Bahdanau, Kyunghyun Cho and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. *arXiv preprint arXiv:1409.0473*.

Fengshun Xiao, Jiangtong Li, Hai Zhao, Rui Wang and Kehai Chen. 2019. Lattice-based Transformer Encoder for Neural Machine Translation. *The 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart and Alexander M. Rush. 2018. OpenNMT: Neural Machine Translation Toolkit. *arXiv preprint arXiv:1805.11462*.

Hai Zhao and Chunyu Kit. 2008. Exploiting Unlabeled Text with Different Unsupervised Segmentation Criteria for Chinese Word Segmentation. *The 9th International Conference on Intelligent Text Processing and Computational Linguistics (CICLing-2008)*, volume 33. page 93–104.

Hai Zhao and Chunyu Kit. 2008. An Empirical Comparison of Goodness Measures for Unsupervised Chinese Word Segmentation with a Unified Framework. *The Third International Joint Conference on Natural Language Processing (IJCNLP-2008)*, volume 1. page 9–16.

Hai Zhao, Masao Utiyama, Eiichro Sumita, and Bao-Liang Lu. 2013. An Empirical Study on Word Segmentation for Chinese Machine Translation. *CICLing 2013*, page 248–263.

Hai Zhao, Tianjiao Yin and Jingyi Zhang. 2013. Vietnamese to Chinese Machine Translation via Chinese Character as Pivot. *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation (PACLIC 27)*, page 250–259.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. *In Proceedings of NIPS 2014*, page 3104–3112.

Jingyi Zhang, Masao Utiyama, Eiichro Sumita, Hai Zhao, Graham Neubig and Satoshi Nakamura. 2016. Learning local word reorderings for hierarchical phrase-based statistical machine translation. *Machine Translation*, volume 30. page 1–18.

Jinji Li, Dong-Il Kim and Jong-Hyeok Lee. 2008. Annotation Guidelines for Chinese-Korean Word Alignment. *In Proceedings of LREC*.

Jinji Li ,Jungi Kim, Dong-Il Kim and Jong-Hyeok Lee. 2009. Chinese Syntactic Reordering for Adequate Generation of Korean Verbal Phrases in Chinese-to-Korean SMT. *Proceedings of the Fourth Workshop on Statistical Machine Translation*, page 190–196.

Jin-Xia Huang and Key-Sun Choi. 2000. Using Bilingual Semantic Information in Chinese-Korean Word Alignment. *Proceedings of the 14th Pacific Asia Conference on Language, Information and Computation*, page 121–130.

Joon-Choul Shin and Cheol-Young Ock 2012. A Korean morphological analyzer using a pre-analyzed partial word-phrase dictionary. *KIISE:Software and Applications*, volume 39, page 415–424. [in Korean]

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A Method for Automatic Evaluation of Machine Translation. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, pages 311–318.

Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, page 1724–1734.

Lontu Zhang and Mamoru Komachi. 2018. machine translation of logographic language using sub-character level information. *In Proceedings of the Third Conference on Machine Translation*, page 17–25

Minh-Thang Luong and Christopher D Manning. 2015. Stanford neural machine translation systems for spoken language domains. *In In Proceedings of the International Workshop on Spoken Language Translation*

Minh-Thang Luong, Hieu Pham and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, page 1412–1421.

Nicola Ueffing and Hermann Ney. 2013. Using POS Information for SMT into Morphologically Rich Languages. *10th Conference of the European Chapter of the Association for Computational Linguistics*, page 347–354.

Nizar Habash and Jun Hu. 2009. Improving Arabic-Chinese Statistical Machine Translation using English as Pivot Language. *Proceedings of the Fourth Workshop on Statistical Machine Translation*, page 173–181.

Phuoc Nguyen, anh-dung Vo, Joon-Choul Shin, Phuoc Tran and Cheol-Young Ock. 2019. Korean-Vietnamese Neural Machine Translation System With Korean Morphological Analysis and Word Sense Disambiguation. *IEEE Access*, volume 7. page 32602–32614.

Rafael E. Banchs and Marta Ruiz Costa-jusaa. 2011. A Semantic Feature for Statistical Machine Translation. *Proceedings of the Fifth Workshop on Syntax, Semantics and Structure in Statistical Translation*, page 126–134.

Rico Sennrich and Barry Haddow. 2016. Linguistic Input Features Improve Neural Machine Translation. *Proceedings of the First Conference on Machine Translation*, volume 1, page 83–91.

Rico Sennrich, Barry Haddow and Alexandra Birch. 2016. Neural Machine Translation of Rare Words with Subword Units. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016)*.

Samira Tofighi Zahabi, Somayeh Bakhshaei and Shahram Khadivi. 2013. Using Context Vectors in Improving a Machine Translation System with Bridge Language. *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, volume 2. page 318–322.

Shaohui Kuang, Junhui Li, António Branco, Weihua Luo and Deyi Xiong. 2018. Attention Focusing for Neural Machine Translation by Bridging Source and Target Embeddings. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 1. page 1767–1776.

Shuo Zang, Hai Zhao, Chunyang Wu and Rui Wang. 2015. A Novel Word Reordering Method for Statistical Machine Translation. *The 2015 11th International Conference on Natural Computation (ICNC'15) and the 2015 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD'15)*, page 843–848.

Yingting Wu and Hai Zhao. 2018. Finding Better Subword Segmentation for Neural Machine Translation. *The Seventeenth China National Conference on Computational Linguistics*, Volume 11221, page 53—64.

Yingting Wu, Hai Zhao and Jia-Jun Tong. 2018. Multilingual Universal Dependency Parsing from Raw Text with Low Resource Language Enhancement. *Proceedings of CoNLL 2018*, page 74–80.

Yuanmei Lu, Toshiaki Nakazawa and Sadao Kurohashi. 2015. Korean-to-Chinese Word Translation using Chinese Character Knowledge. *Proceedings of MT Summit*, 15(1). page 256–269.

Zhisong Zhang, Rui Wang, Masao Utiyama, Eiichiro Sumita and Hai Zhao. 2018. Exploring Recombination for Efficient Decoding of Neural Machine Translation. *Proceedings of EMNLP 2018*, page 4785—4790.

Zuchao Li, Shexia He, Zhuosheng Zhang and Hai Zhao. 2018. Joint Learning for Universal Dependency Parsing. *Proceedings of CoNLL 2018*, page 65-73.

Zuchao Li, Jiaxun Cai, Shexia He and Hai Zhao. 2018. Seq2seq Dependency Parsing. *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*, page 3203–3214.

Zuchao Li, Shexia He, Hai Zhao, Yiqing Zhang, Zhuosheng Zhang, Xi Zhou and Xiang Zhou. 2019. Dependency or Span, End-to-End Uniform Semantic Role Labeling. *Proceedings of AAAI 2019*.

Adapting Neural Machine Translation for English-Vietnamese using Google Translate system for Back-translation

Nghia Luan Pham
Hai Phong University
Haiphong, Vietnam
luanpn@dhhp.edu.vn

Van Vinh Nguyen
University of Engineering and Technology
Vietnam National University
Hanoi, Vietnam
vinhnv@vnu.edu.vn

Abstract

Monolingual data have been demonstrated to be helpful in improving translation quality of both statistical machine translation (SMT) systems and neural machine translation (NMT) systems, especially in resource-poor language or domain adaptation tasks where parallel data are not rich enough. Google Translate is a well-known machine translation system. It has implemented the Google Neural Machine Translation (GNMT) over many language pairs and English-Vietnamese language pair is one of them.

In this paper, we propose a method to better leveraging monolingual data by exploiting the advantages of GNMT system. Our method for adapting a general neural machine translation system to a specific domain, by exploiting Back-translation technique using target-side monolingual data. This solution requires no changes to the model architecture from a standard NMT system. Experiment results show that our method can improve translation quality, results significantly outperforming strong baseline systems, our method improves translation quality in legal domain up to 13.65 BLEU points over the baseline system for English-Vietnamese pair language.

1 Introduction

Machine translation relies on the statistics of a large parallel corpus, datasets of paired sentences in both sides the source and target language. Monolingual data has been traditionally used to train language models which improved the fluency of statistical machine translation (Koehn2010). Neural machine translation (NMT) systems require a very large amount of training data to make generalizations, both on the source side and on the target side. This data typically comes in the form of a parallel corpus, in which each sentence in the source language is matched to a translation in the target language. Unlike parallel corpus, monolingual data are usually much easier to collect and more diverse and have been attractive resources for improving machine translation models since the 1990s when data-driven machine translation systems were first built. Adding monolingual data to NMT is important because sufficient parallel data is unavailable for all but a few popular language pairs and domains.

From the machine translation perspective, there are two main problems when translating English to Vietnamese: First, the own characteristics of an analytic language like Vietnamese make the translation harder. Second, the lack of Vietnamese-related resources as well as good linguistic processing tools for Vietnamese also affects to the translation quality. In the linguistic aspect, we might consider Vietnamese is a source-poor language, especially parallel corpus in many specific domains, for example, mechanical domain, legal domain, medical domain, etc.

Google Translate is a well-known machine translation system. It has implemented the Google Neural Machine Translation (GNMT) over many language pairs and English-Vietnamese language pair is one of them. The translation quality is good for the general domain of this language pair. So we want to leverage advantages of GNMT system *(resources, techniques,...)* to build a domain translation sys-

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 531-539, Hakodate, Japan, September 13-15, 2019

tem for this language pair, then we can improve the quality of translation by integrating more features of Vietnamese.

Language is very complicated and ambiguous. Many words have several meanings that change according to the context of the sentence. The accuracy of the machine translation depends on the topic thats being translated. If the content translated includes a lot of technical or specialized things, its unlikely that Google Translate will work. If the text includes jargon, slang and colloquial words this can be almost impossible for Google Translate to identify. If the tool is not trained to understand these linguistic irregularities, the translation will come out literal and (most likely) incorrect.

This paper presents a new method to adapt the general neural machine translation system to a different domain. Our experiments were conducted for the English-Vietnamese language pair in the direction from English to Vietnamese. We use domain-specific corpora comprising of two specific domains: legal domain and general domain. The data has been collected from documents, dictionaries and the IWSLT2015 workshop for the English-Vietnamese translation task.

This paper is structured as follows. Section 2 summarizes the related works. Our method is described in Section 3. Section 4 presents the experiments and results. Analysis and discussions are presented in Section 5. Finally, conclusions and future works are presented in Section 6.

2 Related works

In statistical machine translation, the synthetic parallel corpus has been primarily proposed as a means to exploit monolingual data. By applying a self-training scheme, the pseudo parallel corpus was obtained by automatically translating the source-side monolingual data (Nicola Ueffing2007; Hua Wu and Zong2008). In a similar but reverse way, the target-side monolingual data were also employed to build the synthetic parallel corpus (Bertoldi and Federico2009; Patrik Lambert2011). The primary goal of these works was to adapt trained SMT models to other domains using relatively abundant in-domain monolingual data.

In (Bojar and Tamchyna2011a), synthetic parallel corpus by Back-translation has been applied successfully in phrase-based SMT. The method in this paper used back-translated data to optimize the translation model of a phrase-based SMT system and show improvements in the overall translation quality for 8 language pairs.

Recently, more research has been focusing on the use of monolingual data for NMT. Previous work combines NMT models with separately trained language models (Gülçehre et al.2015). In (Sennrich et al.2015), authors showed that target-side monolingual data can greatly enhance the decoder model. They do not propose any changes in the network architecture, but rather pair monolingual data with automatic Back-translations and treat it as additional training data. Contrary to this, (Zhang and Zong2016) exploit source-side monolingual data by employing the neural network to generate the synthetic large-scale parallel corpus and multi-task learning to predict the translation and the reordered source-side monolingual sentences simultaneously.

Similarly, recent studies have shown different approaches to exploiting monolingual data to improve NMT. In (Caglar Gulcehre and Bengio2015), authors presented two approaches to integrating a language model trained on monolingual data into the decoder of an NMT system. Similarly, (Domhan and Hieber2017) focus on improving the decoder with monolingual data. While these studies show improved overall translation quality, they require changing the underlying neural network architecture. In contrast, Back-translation allows one to generate a parallel corpus that, consecutively, can be used for training in a standard NMT implementation as presented by (Rico Sennrich and Birch016a), authors used 4.4M sentence pairs of authentic human-translated parallel data to train a baseline English to German NMT system that is later used to translate 3.6M German and 4.2M English target-side sentences. These are then mixed with the initial data to create human + synthetic parallel corpus which is then used to train new models.

In (Alina Karakanta and van Genabith2018), authors use back-translation data to improve MT for a resource-poor language, namely Belarusian (BE). They transliterate a resource-rich language (Russian, RU) into their resource-poor language (BE) and train a BE to EN system, which is then used to translate

monolingual BE data into EN. Finally, an EN to BE system is trained with that back-translation data.

Our method has some differences from the above methods. As described in the above, synthetic parallel data have been widely used to boost the performance of NMT. In this work, we further extend their application by training NMT with synthetic parallel data by using Google Translate system. Moreover, our method investigating Back-translation in Neural Machine Translation for the English-Vietnamese language pair in the legal domain.

3 Our method

In Machine Translation, translation quality depends on training data. Generally, machine translation systems are usually trained on a very large amount of parallel corpus. Currently, a high-quality parallel corpus is only available for a few popular language pairs. Furthermore, for each language pair, the size of specific domains corpora and the number of domains available are limited. The English-Vietnamese is resource-poor language pair thus parallel corpus of many domains in this pair is not available or only a small amount of this data. However, monolingual data for these domains are always available, so we want to leverage a very large amount of this helpful monolingual data for our domain adaptation task in neural machine translation for English-Vietnamese pair.

The main idea in this paper, that is leveraging domain monolingual data in the target language for domain adaptation task by using Back-translation technique and Google Translate system. In this section, we present an overview of the NMT system which is used in our experiments and the next we describe our main idea in detail.

3.1 Neural Machine Translation

Given a source sentence $x = (x_1, ..., x_m)$ and its corresponding target sentence $y = (y_1, ..., y_n)$, the NMT aims to model the conditional probability $p(y|x)$ with a single large neural network. To parameterize the conditional distribution, recent studies on NMT employ the encoder-decoder architecture (Kalchbrenner and Blunsom2013; Kyunghyun Cho and Bengio014b; Ilya Sutskever and Le2014). Thereafter, the attention mechanism (Dzmitry Bahdanau and Bengio2014; Minh-Thang Luong and Manning2015b) has been introduced and successfully addressed the quality degradation of NMT when dealing with long input sentences (Kyunghyun Cho and Bengio014a).

In this study, we use the attentional NMT architecture proposed by (Dzmitry Bahdanau and Bengio2014). In their work, the encoder, which is a bidirectional recurrent neural network, reads the source sentence and generates a sequence of source representations $h = (h_1, ..., h_m)$. The decoder, which is another recurrent neural network, produces the target sentence one symbol at a time. The log conditional probability thus can be decomposed as follows:

$$\log p(y|x) = \sum_{i=1}^{n} \log p(y_t|y_{<t}, x) \quad (1)$$

where $y_{<t} = (y_1, ..., y_{t-1})$. As described in Equation 2, the conditional distribution of $p(y_t|y_{<t}, x)$ is modeled as a function of the previously predicted output y_{t-1}, the hidden state of the decoder s_t, and the context vector c_t.

$$p(y_t|y_{<t}, x) \propto \exp\{g(y_{t-1}, s_t, c_t)\} \quad (2)$$

The context vector c_t is used to determine the relevant part of the source sentence to predict y_t. It is computed as the weighted sum of source representations $h_1, ..., h_m$. Each weight α_{ti} for h_i implies the probability of the target symbol y_t being aligned to the source symbol x_i:

$$c_t = \sum_{i=1}^{m} \alpha_{ti} h_i \quad (3)$$

Given a sentence-aligned parallel corpus of size N, the entire parameter θ of the NMT model is jointly trained to maximize the conditional probabilities of all sentence pairs $\{(x^n, y^n)\}_{n=1}^{N}$:

$$\theta^* = arg\max_{\theta} \sum_{n=1}^{N} \log p(y^n|x^n) \quad (4)$$

where θ^* is the optimal parameter.

3.2 Back-translation using Google's Neural Machine Translation

In recent years, machine translation has grown in sophistication and accessibility beyond what we imaged. Currently, there are a number of online translation services ranging in ability, such as Google Translate[1], Bing Microsoft Translator[2], Babylon Translator[3], Facebook Machine Translation, etc. The Google Translate service is one of the most used machine services because of its convenience.

The Google Translate is launched in 2006 as a statistical machine translation, Google Translate has improved dramatically since its creation. Most significantly in 2017, Google moved away from Phrase-Based Machine Translation and was replaced by Neural Machine Translation (GNMT) (Johnson et al.2017). According to Googles own tests, the accuracy of the translation depends on the languages translated. Many languages have even low accurate because of their complexity and differences.

The Back-translation techniques, the first trains an intermediate system on the parallel data which is used to translate the target monolingual data into the source language. The result is a parallel corpus where the source side is synthetic machine translation output while the target is text written by humans. The synthetic parallel corpus is then simply added to the parallel corpus available to train a final system that will translate from the source to the target language. Although simple, this method has been shown to be helpful for phrase-based translation (Bojar and Tamchyna2011b), NMT (Rico Sennrich and Birch2016) as well as unsupervised MT (Guillaume Lample and Ranzato2018). Although here we focus on adapting English to Vietnamese and investigate, experiment on legal domain data. However, this method can be also applied to many other different domains for this language pair.

To take advantages of the Google Translate and helpfulness of domain monolingual data, we use the back-translation technique combine with the Google Translate to synthesize parallel corpus for training our translation system. Our method is described in detail in Figure 1.

[1]https://translate.google.com

[2]https://www.bing.com/translator

[3]https://translation.babylon-software.com/

In Figure 1, our method includes 3 stages, with details as follows:

- **Stage 1:** In this stage, we use Google Translate to translate domain monolingual data in Vietnamese *(target language side)*. The output of this stage is a translation in English *(source language side)*. This technique is called Back-translation. In this case, using the high-quality model to back-translate domain-specific monolingual target data, and then building a new model with this synthetic training data, might be useful for domain adaptation.

- **Stage 2:** In this stage, at first we synthesize parallel corpus by combine input domain monolingual data with output translation in stage 1, because input monolingual data in the legal domain, therefore we consider this synthetic parallel corpus is also in the legal domain. Next, we mix synthetic parallel corpus with an original parallel corpus which is provided by the IWSLT2015[4] workshop *(this corpus in general domain)*, this is the most interesting scenario which allows us to trace the changes in quality with increases in synthetic-to-original parallel data ratio.

- **Stage 3:** With the parallel corpus mixed in stage 2, we conduct training NMT systems from English to Vietnamese and evaluate translation quality in the legal domain and general domain.

4 Experiments setup

In this section, we describe the data sets used in our experiments, data preprocessing, the training and evaluation in detail.

4.1 Datasets and Preprocessing

Datasets We experiment on the data sets of the English-Vietnamese language pair. All experiments, we consider two different domains that are legal domain and general domain. The summary of the parallel and monolingual data is presented in Table 1.

[4]http://workshop2015.iwslt.org/

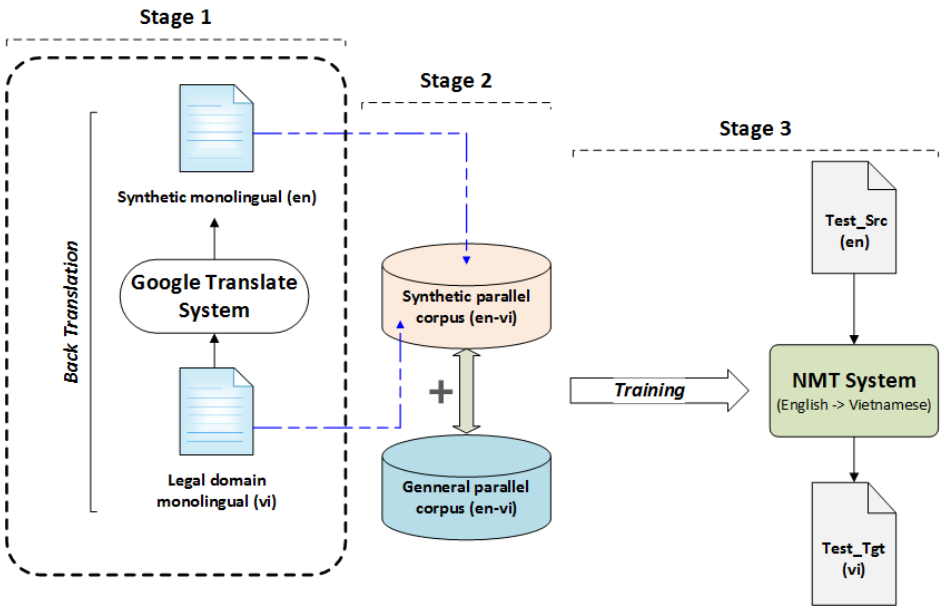


Figure 1: An illustration for our method, includes 3 stages: 1) Back-translation legal domain monolingual text by using Google Translate system; 2) synthesize parallel data from synthetic monolingual and legal domain monolingual in stage 1, and 3) combine synthetic parallel corpus with general parallel corpus for training NMT system

- For training baseline systems, we use the English-Vietnamese parallel corpus which is provided by IWSLT2015 *(133k sentence pairs)*, this corpus was used as general domain training data and tst2012/tst2013 data sets were selected as validation *(val)* and test data respectively.
- For creating the source side data *(English)*, we use 100k sentences in legal domain in target side *(Vietnamese)*.
- To evaluation, we use 500 sentence pairs in legal domain and 1,246 sentence pairs in general domain *(tst2013 data set)*.

Preprocessing Each training corpus is tokenized using the tokenization script in Moses (Koehn et al.2007) for English. For cleaning, we only applied the script *clean-n-corpus.perl* in Moses to remove lines in the parallel data containing more than 80 tokens.

In Vietnamese, a word boundary is not white space. White spaces are used to separate syllables in Vietnamese, not words. A Vietnamese word consist of one or more syllables. We use vnTokenizer (Phuong et al.2013) for word segmentation. However, we only used for separation marks such as dots, commas and other special symbols.

4.2 Settings

We have trained a Neural Machine Translation system by using the OpenNMT[5] toolkit (Klein et al.2018) with the seq2seq architecture of (Sutskever et al.2014), this is a state-of-the-art open-source neural machine translation system, started in December 2016 by the Harvard NLP group and SYSTRAN. This architecture is formed by an encoder, which converts the source sentence into a sequence of numerical vectors, and a decoder, which predicts the target sentence based on the encoded source sentence. In our NMT models is trained with the default model, which consists of a 2-layer Long Short-Term Memory (LSTM) network (Luong et al.2015) with 500 hidden units on both the encoder/decoder and the general attention type of (Minh-Thang Luong and Manning2015a).

For translation evaluation, we use standard BLEU score metric (Bi-Lingual Evaluation Understudy) (Kishore Papineni and Zhu2002) that is currently one of the most popular methods of automatic ma-

[5]http://opennmt.net/

Data Sets		Language	
		English	**Vietnamese**
Training	Sentences	**133316**	
	Average Length	16.62	16.68
	Words	1952307	1918524
	Vocabulary	40568	28414
Val	Sentences	**1553**	
	Average Length	16.21	16.97
	Words	13263	12963
	Vocabulary	2230	1986
General_test	Sentences	**1246**	
	Average Length	16.15	15.96
	Words	18013	16989
	Vocabulary	2708	2769
Legal_test	Sentences	**500**	
	Average Length	15.21	15.48
	Words	7605	7740
	Vocabulary	1530	1429

Table 1: The Summary statistics of data sets: English-Vietnamese

chine translation evaluation. The translated output of the test set is compared with different manually translated references of the same set.

4.3 Experiments and Results

In our experiments, we train NMT models with parallel corpus composed of: (1) synthetic data only; (2) IWSLT 2015 parallel corpus only; and (3) a mixture of parallel corpus and synthetic data. We trained 5 NMT systems and evaluated the quality of translation on the general domain data and the legal domain data. We also compare the translation quality of our systems with Google Translate, Our systems are described as follows:

- ***The system are built using IWSLT2015 data only:*** This baseline system is trained on general domain data which is provided by IWSLT2015 workshop. Training data *(133k sentences pairs)* and tst2012 data set were selected as validation *(val)*, we call this system is ***Baseline***.
- ***The system are built using synthetic data only:*** Such systems represent the case where no parallel data is available but monolingual data can be translated via an existing MT system and provided as a training corpus to a new NMT system. This case we use 100k sentences in Vietnamese in the legal domain and use Google Translate system for Back-translation. The synthetic parallel data is used for training NMT system and tst2012 data set were selected as validation *(val)*, this system is called ***Synthetic***.
- ***The system are built using mixture of parallel corpus and synthetic data:*** This is the most interesting scenario which allows us to trace the changes in quality with increases in synthetic-to-orginal data ratio. we train 2 NMT systems, the first system is trained on IWSLT2015 data *(133k sentences pairs)* + Synthetic *(50k sentences pairs)* and second system is trained on IWSLT2015 *(133k sentences pairs)* + Synthetic *(100k sentences pairs)*, and tst2012 data set were selected as validation *(val)*, these systems is called ***Baseline_Syn50*** and ***Baseline_Syn100*** respectively.

Our NMT systems are evaluated in the general domain and legal domain. We also compare translation quality with Google Translate on the same test domain data set. Experiment results are shown by the bleu score as table 2 and table 3.

As the results in table 2 and table 3, the Baseline NMT system achieved 25.43 BLEU score in general domain but reduced to 19.23 in the legal domain. After applying Back-translation, the results are im-

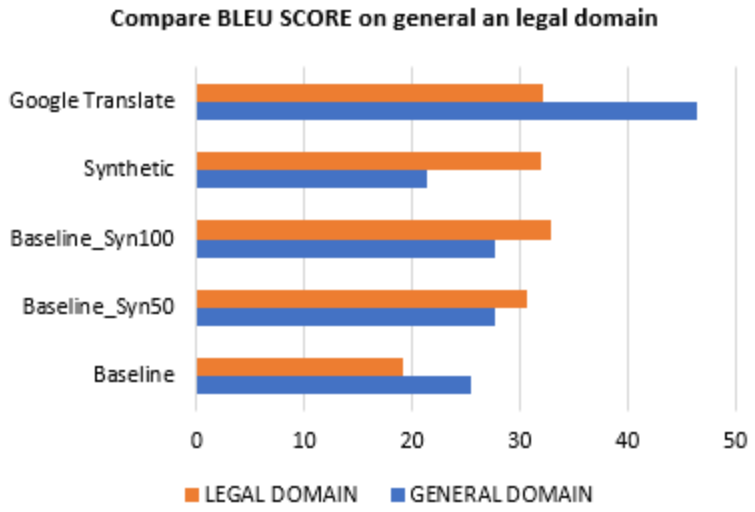


Figure 2: Comparison of translation quality when translating in the legal domain and general domain.

SYSTEM	BLEU SCORE
Baseline	**25.43**
Baseline_Syn50	**27.74**
Baseline_Syn100	**27.68**
Synthetic	**21.42**
Google Translate	**46.47**

Table 2: The experiment results of our systems in the general domain

SYSTEM	BLEU SCORE
Baseline	**19.23**
Baseline_Syn50	**30.61**
Baseline_Syn100	**32.88**
Synthetic	**31.98**
Google Translate	**32.05**

Table 3: The experiment results of our systems in the legal domain

proved, significantly outperforming strong baseline systems, our method improves translation quality in legal domain up to 13.65 BLEU points over baseline system and 2.25 BLEU points over baseline system in general domain.

In Figure 2 is shown the comparison of translation quality when translating in the legal domain and general domain. In general domain, Google Translate's bleu score is 46.47 points, the baseline system is 25.43 points and bleu score of our systems are higher than the baseline system, reaching 27.68; 27.74 points respectively. In the legal domain, Google Translate's bleu score is 32.05 points, the baseline system is 19.23 points and bleu score of our systems are higher than the baseline system, reaching 31.98, 32.61 and 32.88 points respectively. Thus, Back-translation uses Google Translate for English - Vietnamese language pair in the legal domain can improve the translation quality of the English - Vietnamese translation system.

5 Analysis and discussions

The Back-translation technique enables the use of synthetic parallel data, obtained by automatically translating cheap and in many cases available information in the target language into the source language. The synthetic parallel data generated in this way is combined with parallel texts and used to improve the quality of NMT systems. This method is simple and it has been also shown to be helpful for machine translation.

We have experimented with different synthetic data rates and observed effects on translation results. However, we have not investigated to answer issues for adapting the legal domain in NMT of English-Vietnamese language pair such as:

- Does back-translation direction matter?
- How much monolingual back-translation data is necessary to see a significant impact in MT quality?

- Which sentences are worth back translating and which can be skipped?

Overall, we are becoming smarter in selecting incremental synthetic data in NMT that helps improve both: performance of the systems and translation accuracy.

6 Conlustion

In this work, we presented a simple but effective method to adapt general neural machine translation systems into the legal domain for English-Vietnamese language pairs. We empirically showed that the quality of the NMT system is selected for Back-translation for synthetic parallel corpus generation very significant *(here we selected Google Translate for leverage advantages of this translation system)*, and neural machine translation performance can be improved by iterative back-translation in a parallel resource-poor language like Vietnamese. Our method improved translation quality by BLEU score up to 13.65 points, results significantly outperforming strong baseline systems on the general domain and legal domain.

In future work, we also want to explore the effect of adding synthetic parallel data to other resource-poor domains of English - Vietnamese language pair. We will investigate the true merits and limits of Back-translation.

Acknowledgments

This work is funded by the project: Building a machine translation system to support translation of documents between Vietnamese and Japanese to help managers and businesses in Hanoi approach Japanese market, under grant number TC.02-2016-03.

References

Alina Karakanta, J. D. and van Genabith, J. (2018). Neural machine translation for low resource languages without parallel corpora. Machine Translation, 32, 23pp.

Bertoldi, N. and Federico, M. (2009). Domain adaptation for statistical machine translation with monolingual resources. In Proceedings of the fourth workshop on statistical machine translation. Association for Computational Linguistics, pages 182189.

Bojar, O. and Tamchyna, A. (2011a). Improving translation model by monolingual data. In Proceedings of the Sixth Workshop on Statistical Machine Translation, WMT@EMNLP 2011, pages 330336.

Bojar, O. and Tamchyna, A. (2011b). Improving translation model by monolingual data. In Workshop on Statistical Machine Translation.

Caglar Gulcehre, Orhan Firat, K. X. K. C. L. B. H.-C. L. F. B. H. S. and Bengio, Y. (2015). On using monolingual corpora in neural machine translation. CoRR, abs/1503.03535.

Domhan, T. and Hieber, F. (2017). Using target side monolingual data for neural machine translation through multi-task learning. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 15001505.

Dzmitry Bahdanau, K. C. and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

Guillaume Lample, Alexis Conneau, L. D. and Ranzato, M. (2018). Unsupervised machine translation using monolingual corpus only. In International Conference on Learning Representations (ICLR).

Gülçehre, Ç., Firat, O., Xu, K., Cho, K., Barrault, L., Lin, H., Bougares, F., Schwenk, H., and Bengio, Y. (2015). On using monolingual corpora in neural machine translation. *CoRR*, abs/1503.03535.

Hua Wu, H. W. and Zong, C. (2008). Domain adaptation for statistical machine translation with domain dictionary and monolingual corpora. In Proceedings of the 22nd International Conference on Computational Linguistics-Volume 1. Association for Computational Linguistics, pages 9931000.

Ilya Sutskever, O. V. and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In Advances in neural information processing systems. pages 31043112.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Kalchbrenner, N. and Blunsom, P. (2013). Recurrent continuous translation models. In EMNLP. volume 3, page 413.

Kishore Papineni, Salim Roukos, T. W. and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL) pp. 311-318.

Klein, G., Kim, Y., Deng, Y., Nguyen, V., Senellart, J., and Rush, A. (2018). OpenNMT: Neural machine translation toolkit. In *Proceedings of the 13th Conference of the Association for Machine Translation in the*

Americas (Volume 1: Research Papers), pages 177–184, Boston, MA. Association for Machine Translation in the Americas.

Koehn, P. (2010). Statistical machine translation. Cambridge University Press.

Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., and Herbst, E. (2007). Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Kyunghyun Cho, Bart Van Merrienboer, C. G. D. B. F. B.-H. S. and Bengio, Y. (2014b). Learning phrase representations using rnn encoder-decoder for statistical machine translation. arXiv preprint arXiv:1406.1078.

Kyunghyun Cho, Bart van Merrienboer, D. B. and Bengio, Y. (2014a). On the properties of neural machine translation: Encoder-decoder approaches. In Eighth Workshop on Syntax, Semantics and Structure in Statistical Translation (SSST8).

Luong, M., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. *CoRR*, abs/1508.04025.

Minh-Thang Luong, H. P. and Manning, C. D. (2015a). Effective approaches to attention-based neural machine translation. In Proc of EMNLP.

Minh-Thang Luong, H. P. and Manning, C. D. (2015b). Effective approaches to attentionbased neural machine translation. arXiv preprint arXiv:1508.04025.

Nicola Ueffing, Gholamreza Haffari, A. S. (2007). Transductive learning for statistical machine translation. In Annual Meeting-Association for Computational Linguistics. volume 45, page 25.

Patrik Lambert, Holger Schwenk, C. S. a. S. A.-R. (2011). Investigations on translation model adaptation using monolingual data. In Proceedings of the Sixth Workshop on Statistical Machine Translation. Association for Computational Linguistics, pages 284293.

Phuong, L.-H., Nguyen, H., Roussanaly, A., and Ho, T. (2013). A hybrid approach to word segmentation of vietnamese texts.

Rico Sennrich, B. H. and Birch, A. (2016). Improving neural machine translation models with monolingual data. Conference of the Association for Computational Linguistics (ACL).

Rico Sennrich, B. H. and Birch, A. (2016a). Improving neural machine translation models with monolingual data. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 8696.

Sennrich, R., Haddow, B., and Birch, A. (2015). Improving neural machine translation models with monolingual data. *CoRR*, abs/1511.06709.

Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proc. NIPS*, Montreal, CA.

Zhang, J. and Zong, C. (2016). Exploiting source-side monolingual data in neural machine translation. pages 1535–1545.

Re-unifying Floating Numeral Quantifiers and Secondary Predicates in Japanese*

Hideaki Yamashita
Yokohama City University
hy_aka_sange@hotmail.com

Abstract

Miyagawa 1989 unified the treatment of floating numeral quantifiers and secondary predicates in Japanese as *adverbial adjuncts* adjoined to vP/VP not forming a constituent with its host DP it is associated with. In so doing, he showed how analyzing such phenomena instantiates trace, contributing to the linguistic theory at that time. The aim of this paper is to scrutinize such unification building on (i) the conditions that apply to the ellipsis of arguments and adjuncts and (ii) the conditions that apply to the multiple long-distance scrambling. It shows that, while the unification of floating numeral quantifiers and secondary predicates in Japanese is kept intact, it needs to be re-unified; they *can* be *adnominal adjuncts* which is adjoined to its host DP forming a base-generated single constituent. In so doing, it shows that how analyzing such phenomena instantiates free application of merge, contributing to the current linguistic theory.

1 Introduction

2019 marks the 30th anniversary of Miyagawa 1989, one of the most influential book on Japanese generative syntax which showed how the study on syntax of Japanese can contribute to the linguistic theory. One significant finding Miyagawa provided in that seminal work is about the common properties floating numeral quantifiers (FNQs) and secondary predicates (2ndPs) exhibit. He unified the treatment of FNQs and 2ndPs in Japanese as *adverbial adjuncts* adjoined to vP/VP not forming a constituent with its host DP it is associated with, having the structure depicted in (1) (which is adjusted to the minimalist framework (Chomsky 1995, et. seq.)).[1]

(1) a. [$_{vP/VP}$ [$_{DP}$ host DP] [$_{ClP}$ FNQ] V/v]
a'. [$_{vP/VP}$ [$_{DP}$ sake-o] [$_{ClP}$ 3-bon] V/v]
b. [$_{vP/VP}$ [$_{DP}$ host DP] [$_{AdvP}$ 2ndP] V/v]
b'. [$_{vP/VP}$ [$_{DP}$ sake-o] [$_{AdvP}$ hiya-de] V/v]

In so doing, he showed how analyzing such phenomena instantiates trace, contributing to the linguistic theory at that time, especially the (trace) theory of movement.

The aim of this paper is to scrutinize such unification building on (i) the conditions that apply to the ellipsis of arguments and adjuncts and (ii) the conditions that apply to multiple long-distance scrambling. It shows that, while the unification of FNQs and 2ndPs in Japanese is kept intact (supporting Miyagawa's insight), it needs to be re-unified (departing from Miyagawa's analysis); they *can* be *adnominal adjuncts* which is adjoined to its host DP forming a base-generated single constituent, having the structure depicted in (2).

(2) a. [$_{DP}$ [$_{DP}$ host DP] [$_{ClP}$ FNQ]]
a'. [$_{DP}$ [$_{DP}$ sake-o] [$_{ClP}$ 2-hon]]
b. [$_{DP}$ [$_{DP}$ host DP] [$_{AdvP}$ 2ndP]]
b'. [$_{DP}$ [$_{DP}$ sake-o] [$_{AdvP}$ hiya-de]]

Put it differently, I will argue for the Single Constituent (SinC) analysis (3) and argue against the Independent Constituent (InC) analysis (4).

* This work merged my works on floating numeral quantifiers (Yamashita 2015, 2016) and secondary predicates (Yamashita 2019), which is then developed by incorporating the two "species" simultaneously. I'm indebted to those people, especially Hisatsugu Kitahara, Masao Ochi, Yuta Sakamoto, Takashi Toyoshima, and Asako Uchibori, who I had fruitful discussions with, and gave me insightful comments. Needless to say, all the inadequacies are my own.

[1] I will be agnostic about the 'label' of FNQs and 2ndPs and use 'ClP' (= Classifiers Phrase) and 'AdvP' (Adverbial Phrase) merely for expository purpose, without making any theoretical commitment.

Pacific Asia Conference on Language, Information and Computation (PACLIC 33), pages 540-548, Hakodate, Japan, September 13-15, 2019

(3) SinC (Single Constituent) analysis
(Re-unification of FNQs and 2ndPs):
FNQs and 2ndPs can form a base-generated single constituent with its host DPs; they can be adjoined to DP ((2)).

(4) InC (Independent Constituent) analysis
(Unification of FNQs and 2ndPs):
FNQs and 2ndPs do not form a base-generated single constituent with its host DPs; they are adjoined to vP/VP ((1)).

In so doing, I will show how analyzing such phenomena instantiates free application of merge, aiming to contribute to the current linguistic theory, especially the theory of merge developed in Chomsky 2013, 2015, and Chomsky et. al. 2019 in which free application of merge is elaborated.

The organization of this paper is as follows. In Section 2, I will show that re-unification is called for, building on the conditions that apply to the ellipsis of arguments and adjuncts. In Section 3, I will show that re-unification is called for, building on the conditions that apply to the multiple long-distance scrambling. In Section 4, I will discuss a theoretical issues regarding the theory of merge, building on new set of data involving multiple long-distance scrambling. Section 5 is a conclusion.

2 Ellipsis – Argument/Adjunct Asymmetry on Argument Ellipsis

In this section, building on the paradigm involving *argument/adjunct asymmetry on argument ellipsis* (AE), I show that, contrary to the classic InC analyses that treat FNQs and 2ndPs in Japanese as adverbial adjuncts adjoined to vP/VP not forming a constituent with its host DP it is associated with, such adjuncts *can* be *adnominal adjuncts* which is adjoined to its host DP forming a base-generated single constituent, and the host DP as a result functions as a lower segment of DP, as in (2) above.

One of the prominent features of Japanese syntax is a frequent use of null arguments (see Oku 1998, Shinohara 2006, Saito 2007, Takahashi 2008a, et. seq., Yamashita 2014, et. seq., Funakoshi 2016, Sakamoto 2016, et. seq., a.o.). For the sake of exposition, let us assume that (i) such null arguments result from AE, an ellipsis operation (involving LF-copying), which exhibits the so-called *argument/adjunct asymmetry on AE*, as summarized in (5)a and (5)b, and (ii) AE is subject to the basic assumption about the potential target of ellipsis operation in (5)c (Oku 1998, Shinohara 2006, Saito 2007, Yamashita 2014, et. seq., Sakamoto 2016, et. seq., a.o.).

(5) a. Null arguments are derived through AE, which is an LF-copying operation.[2]
b. AE is applicable to only arguments, but not applicable to adjuncts.[3]
c. Any segment can be the target of syntactic operation (e.g., ellipsis).

With this in mind, let us look at the ellipsis paradigm involving (i) FNQs and its host DPs (6) (taken from Yamashita 2015, 2016) and (ii) 2ndPs and its host DPs (7) (taken from Yamashita 2019), which shows exactly the same behaviors.[4, 5, 6]

(6) [Mari-wahaha-ni iPad-o 2-dai
M.-TOP mom-DAT iPad-ACC 2-CL
katta].
bought
'Mari bought 2 iPads for (her) mother.'
a. [Ken-mo haha-ni iPad-o 2-dai
K.-also mom-DAT iPad-ACC 2-CL
katta].
bought
'Ken also bought 2 iPads for (his) mother.'
b. [Ken-mo haha-ni ~~iPad-o~~ 2-dai katta].
c. *[Ken-mo haha-ni iPad-o ~~2-dai~~ katta].

[2] See Shinohara 2006, Saito 2007, Yamashita 2014, Sakamoto 2017, 2019, a.o., for arguments against PF-deletion and pro analyses of AE.
[3] See Oku 1998, Saito 2007 for more discussion on the impossibility of adjunct ellipsis. See also Funakoshi 2016 for the legitimate cases of adjunct ellipsis.
[4] All the Japanese examples are transcribed in the *Hepburn (Hebon)* system Romanization. The translations in single quotes are intended to give the (rough) structure of the examples and are not meant to be the correct English translations. Translations and glosses are provided minimally, only when necessary.
[5] I use the double strike-through (~~*XP*~~) to indicate ellipsis.
[6] Both (i) FNQs and its host subject DPs and (ii) 2ndPs and its host subject DPs exhibit the same properties, but due to the space limitation, I can only provide here examples involving object DPs.

d. [Ken-mo haha-ni ~~iPad-o~~ ~~2-dai~~ katta].
e. *[Ken-mo ~~haha-ni~~ iPad-o ~~2-dai~~ katta].

(6)a is the sentence without any ellipsis applied. (6)b shows that it is possible to delete the host DP *iPad-o* alone, excluding the FNQ *2-dai*; this can be achieved under both the SinC ((3)) and InC analysis ((4)) since it involves AE of a host DP which is selected directly by the predicate; *[$_{DP}$ ~~[$_{DP}$ host DP]~~ [$_{ClP}$ FNQ]]* under (3), and *[$_{VP}$ ~~[$_{DP}$ host DP]~~ [$_{ClP}$ FNQ] V]* under (4). (6)c shows that it is not possible to delete the FNQ *2-dai* alone, excluding the host DP *iPad-o*. This fact indicates that FNQ in Japanese behaves like adjunct. Assuming this is on the right track, the deviance of (6)c can be captured under both the SinC ((3)) and InC analysis ((4)) by the assumption (5)b.

The crucial paradigm to the present work is the contrast between the legitimate (6)d and the illegitimate (6)e. Let us first consider (6)d, which deletes both the FNQ and its host DP (direct object (DO)). Under the SinC approach, this is readily allowed because it involves a run-of-the-mill AE (~~[$_{DP}$ [$_{DP}$ host DP] [$_{ClP}$ FNQ]]~~). Next, consider (6)e, which deletes the FNQ and the indirect object (IO) which is not associated with the FNQ. Note first that the deviance of (6)e in contrast to (6)d suggests that the legitimate ellipsis of FNQ does not involve the Principle of Minimal Compliance effect (Richards 1998); the possible AE of IO will not save the otherwise impossible adjunct ellipsis of FNQ in (6)e. Hence, this suggests that it is not the possible AE of DO that saves the otherwise impossible adjunct ellipsis of FNQ in (6)d. In addition, the deviance of (6)e also suggests that AE is not applicable to the "derived" constituent (aka surprising constituent). Even if "oblique movement" (Takano 2002; see also Sohn 1994) were available to form the otherwise independent constituents into the single constituent, the resulting constituent is not subject to AE. Given that the FNQ, being an adjunct, cannot undergo ellipsis (6)c, then the legitimate ellipsis in (6)d constitutes evidence for the SinC analysis where the FNQ and its host DP form a base-generated single constituent. And the contrast between the legitimate (6)d and the illegitimate (6)e constitute evidence against the InC analysis where the FNQ and its host DP do not form a base-generated single constituent, but are instead generated as independent constituents.

(7) [Mari-wahaha-ni sake-o hiya-de M.-TOP mom-DAT sake-ACC cold-DE furumatta]-ga,
served-but
'Mari served sake cold for (her) mom, but …'
a. [Ken-wa haha-ni sake-o hiya-de K.-TOP mom-DAT sake-ACC cold-DE furumawanakatta].
served.not
'Ken did not serve sake cold for (his) mom.'
b. [Ken-wa haha-ni ~~sake-o~~ hiya-de furumawanakatta].
c. *[Ken-wa haha-ni sake-o ~~hiya-de~~ furumawanakatta].
d. [Ken-wa haha-ni ~~sake-o~~ ~~hiya-de~~ furumawanakatta].
e. *[Ken-wa ~~haha-ni~~ sake-o ~~hiya-de~~ furumawanakatta].

(7)a is the sentence without any ellipsis applied. (7)b shows that it is possible to delete the host DP alone, excluding the 2ndP; this can be achieved under both the SinC ((3)) and InC analysis ((4)) since it involves AE of a host DP which is selected directly by the predicate; *[$_{DP}$ ~~[$_{DP}$ host DP]~~ [$_{AdvP}$ 2ndP]]* under (3), and *[$_{VP}$ ~~[$_{DP}$ host DP]~~ [$_{AdvP}$ 2ndP] V]* under (4). (7)c shows that it is not possible to delete the 2ndP alone, excluding the host DP. This fact indicates that 2ndP in Japanese behaves like adjunct. Assuming this is on the right track, the deviance of (5)c can be captured under both the SinC ((3)) and InC analysis ((4)) by the assumption (5)b.

The crucial paradigm to the present work is the contrast between the legitimate (7)d and the illegitimate (7)e. Let us first consider (7)d, which deletes both the 2ndP and its host DP. Under the SinC approach, this is readily allowed because it involves a run-of-the-mill AE (~~[$_{DP}$ [$_{DP}$ host DP] [$_{AdvP}$ 2ndP]]~~). Next, consider (7)e, which deletes the 2ndP and the argument which is not associated with the ADV. Note first that the deviance of (7)e in contrast to (7)d suggests that the legitimate ellipsis of 2ndP does not involve the Principle of Minimal Compliance effect (Richards 1998); the

possible AE will not save the otherwise impossible adjunct ellipsis of 2ndP in (7)e. Hence, this suggests that it is not the possible AE of host DP per se that saves the otherwise impossible adjunct ellipsis of 2ndP in (7)d. In addition, the deviance of (7)e also suggests that AE is not applicable to the "derived" constituent. Even if "oblique movement" (Takano 2002; see also Sohn 1994) were available to form the otherwise independent constituents into the single constituent, the resulting constituent is not subject to AE. Given that the 2ndP, being an adjunct, cannot undergo ellipsis ((7)c), then the legitimate ellipsis in (7)d constitutes evidence for the SinC analysis where the 2ndP and its host DP form a base-generated single constituent. And the contrast between the legitimate (7)d and the illegitimate (7)e constitute evidence against the InC analysis where the 2ndP and its host DP do not form a base-generated single constituent, but are instead generated as independent constituents.

In sum, both (i) FNQs and its host DPs (6) and (ii) 2ndPs and its host DPs (7) show exactly the same behaviors with respect to the ellipsis paradigm, i.e., *argument/adjunct asymmetry on AE*. And I have shown that while the SinC analysis of these adjuncts (i.e., FNQs and 2ndPs) and its host DPs ((3)) can capture the structural difference between the adjuncts and its host DP (DO) and the adjuncts and the non-host DP (IO) (accounting for the ellipsis paradigm), the InC analysis ((4)) cannot (failing to account for the ellipsis paradigm). But what is crucial is that FNQs and 2ndPs should be treated uniformly (supporting Miyagawa's (1989) insight), but it needs to be re-unified (departing from Miyagawa's analysis); they *can* be *adnominal adjuncts* which is attached to its host DP by adjoining to it and forming a constituent with its host argument DP, having the structure depicted in (2) above.

3 Scrambling – Ban on Split Multiple Long-distance Scrambling

In this section, building on the paradigm involving what Yamashita (2015, 2016) referred as *the ban on split multiple long-distance scrambling* (BSML) in Japanese, I show that, contrary to the classic InC analyses that treat FNQs and 2ndPs in Japanese as adverbial adjuncts adjoined to vP/VP not forming a constituent with its host DP it is associated with, such adjuncts *can* be *adnominal adjuncts* which is adjoined to its host DP forming a base-generated single constituent, and the host DP as a result functions as a lower segment of DP, as in (2) above.

In addition to the frequent use of null arguments, one of the prominent features of Japanese syntax is that scrambling, especially, long-distance scrambling (LDS) is allowed (Saito 1985, Sakai 1994, Saito and Fukui 1998, a.o.). For the sake of exposition, let us assume that (i) such LDS results from overt upward/leftward movement, an optional dislocation operation, summarized in (8), which is evidenced by (9)–(12), and (ii) scrambling is subject to the basic assumption about the potential target of movement operation in (8)c.

(8) a. LDS is unbounded; i.e., super-LDS is possible.
b. LDS can apply multiply.
c. Any segment can be the target of syntactic operation (e.g., scrambling).

In (9), *Aya-ni* or *sake-o* which is first generated in the most embedded clause (CP1) undergoes LDS to the top of CP2 (which is immediately above CP1).

(9) a. [$_{CP3}$ Yui-ga [$_{CP2}$ Aya-ni$_j$ Ken-wa
Y.-NOM Aya-DAT K.-TOP
[$_{CP1}$ Mari-ga naze t$_j$ sake-o
M.-NOM why sake-ACC
(2-hon/hiya-de) furumatta-to] omotta-ka]
2-CL/cold-DE served-C thought-Q
shiritagatteiru-yo].
wants.to.know-SFP
'[Yui wants to know [Q Ken thought [that Mari served Aya (two) sake(s) (cold) why]]].'
b. [$_{CP3}$ Yui-ga [$_{CP2}$ sake-o$_i$ Ken-wa
[$_{CP1}$ Mari-ga naze Aya-ni t$_i$
(2-hon/hiya-de) furumatta-to] omotta-ka]
shiritagatteiru-yo].

In addition, it is not impossible to continue LDS to the higher clause(s) (Sakai 1994). As can be seen in (10), *Aya-ni* or *sake-o* can undergo subsequent LDS (hereafter, super-LDS) to the top of CP3, which end up in crossing two CP boundaries from its base-generated position.

(10) a. [$_{CP3}$ Aya-ni$_j$ Yui-ga [$_{CP2}$ t$_j$ Ken-wa [$_{CP1}$ Mari-ga naze t$_j$ sake-o (2-hon/hiya-de) furumatta-to] omotta-ka] shiritagatteiru-yo].

b. [$_{CP3}$ Sake-o$_i$ Yui-ga [$_{CP2}$ t$_i$ Ken-wa [$_{CP1}$ Mari-ga naze Aya-ni t$_i$ (2-hon/hiya-de) furumatta-to] omotta-ka] shiritagatteiru-yo].

Furthermore, the number of phrases that can undergo LDS is in principle unlimited (Saito and Fukui 1998:444, fn.8), and two or more phrases can undergo LDS; that is, multiple (super-)LDS is possible, as in (11) and (12).

(11) a. [$_{CP3}$ Yui-ga [$_{CP2}$ Aya-ni$_j$ sake-o$_i$ Ken-wa [$_{CP1}$ Mari-ga naze t$_j$ t$_i$ (2-hon/hiya-de) furumatta-to] omotta-ka] shiritagatteiru-yo].

b. [$_{CP3}$ Yui-ga [$_{CP2}$ sake-o$_i$ Aya-ni$_j$ Ken-wa [$_{CP1}$ Mari-ga naze t$_j$ t$_i$ (2-hon/hiya-de) furumatta-to] omotta-ka] shiritagatteiru-yo].

(12) a. [$_{CP3}$ Aya-ni$_j$ sake-o$_i$ Yui-ga [$_{CP2}$ t$_j$ t$_i$ Ken-wa [$_{CP1}$ Mari-ga naze t$_j$ t$_i$ (2-hon/hiya-de) furumatta-to] omotta-ka] shiritagatteiru-yo].

b. [$_{CP3}$ Sake-o$_i$ Aya-ni$_j$ Yui-ga [$_{CP2}$ t$_i$ t$_j$ Ken-wa [$_{CP1}$ Mari-ga naze t$_j$ t$_i$ (2-hon/hiya-de) furumatta-to] omotta-ka] shiritagatteiru-yo].

Thus, LDS in Japanese can in principle be unbounded, crossing more than two or more clausal boundaries (yielding (super-)LDS) and can apply multiply, moving two or more phrases (yielding multiple (super-)LDS).

An interesting quirk about multiple LDS is that there is a curious constraint exemplified in (13), which is referred as *the ban on split multiple LDS (BSML)* (Yamashita 2015, 2016).

(13) a. *[$_{CP3}$ Aya-ni$_j$ Yui-ga [$_{CP2}$ t$_j$ sake-o$_i$ Ken-wa [$_{CP1}$ Mari-ga naze t$_j$ t$_i$ (2-hon/hiya-de) furumatta-to] omotta-ka] shiritagatteiru-yo].

b. *[$_{CP3}$ Sake-o$_i$ Yui-ga [$_{CP2}$ Aya-ni$_j$ t$_i$ Ken-wa [$_{CP1}$ Mari-ga naze t$_j$ t$_i$ (2-hon/hiya-de) furumatta-to] omotta-ka] shiritagatteiru-yo].

As first observed on independent ground by Sakai (1994) and Sohn (1994), the otherwise possible multiple LDS/super-LDS becomes impossible when scrambled phrases end up in different landing sites: *(super-)LDSed phrases cannot be "split" apart, in the sense that they cannot end up in different landing sites; rather, they need to be in the same landing site*. To put it differently, *multiple (super-)LDSed phrases must be "adjacent" to each other*. Recall the legitimate cases of multiple (super-)LDS in (11) and (12), where multiple (super-)LDSed phrases are not split apart and kept adjacent.[7] If not, it becomes ungrammatical. Descriptively speaking, the BSML emerges when two clause-mate phrases which are not a syntactic constituent, such as *Aya-ni* (IO of CP1) and *sake-o* (DO of CP1), are not adjacent to each other at their landing sites. The state-of-affairs can be schematically represented as in (14).

(14) a. Multiple LDS
OK [$_{CP3}$ … [$_{CP2}$ *XP YP* … [$_{CP1}$ … t_{XP} t_{YP} …]]]

b. Multiple super-LDS
OK [[$_{CP3}$ *XP YP* … [$_{CP2}$ t_{XP} t_{YP} … [$_{CP1}$ … t_{XP} t_{YP} …]]]

c. Split multiple LDS
* [$_{CP3}$ *XP* … [$_{CP2}$ t_{XP} *YP* … [$_{CP1}$ … t_{XP} t_{YP} …]]]

With this in mind, let us look at the scrambling paradigm involving (i) FNQs and its host DPs (15) (based on Yamashita 2016: (14)) and (ii) 2ndPs and its host DPs (16), which shows exactly the same behaviors. That a multiple LDS ((15)a–b and (16)a–b) and a multiple super-LDS ((15)c–d and (16)c–d) is possible is not surprising. But what is interesting is that the BSML is inapplicable to a case involving both FNQs and its host DPs and 2ndPs and its host DPs; crucially, even what seems like an instance of split multiple LDS involving FNQs or 2ndPs and its host DPs is grammatical as in (15)e–f and (16)e–f.

[7] This holds true for (i) the otherwise possible multiple super-LDS involving argument and adjunct (e.g. *naze*) (a.k.a. the Free Ride effect) becomes impossible when they are split apart (Sohn 1994) and (ii) the otherwise possible multiple super-LDS involving two or more adjuncts.

(15) a. [$_{CP3}$ Yui-ga [$_{CP2}$ sake-o$_i$ 2-hon$_h$ Ken-wa [$_{CP1}$ Mari-ga naze t_i t_h (hiya-de) furumatta-to] omotta-ka] shiritagatteiru-yo].

b. [$_{CP3}$ Yui-ga [$_{CP2}$ 2-hon$_h$ sake-o$_i$ Ken-wa [$_{CP1}$ Mari-ga naze t_i t_h (hiya-de) furumatta-to] omotta-ka] shiritagatteiru-yo].

c. [$_{CP3}$ sake-o$_i$ 2-hon$_h$ Yui-ga [$_{CP2}$ Ken-wa [$_{CP1}$ Mari-ga naze t_i t_h (hiya-de) furumatta-to] omotta-ka] shiritagatteiru-yo].

d. [$_{CP3}$ 2-hon$_h$ sake-o$_i$ Yui-ga [$_{CP2}$ Ken-wa [$_{CP1}$ Mari-ga naze t_i t_h (hiya-de) furumatta-to] omotta-ka] shiritagatteiru-yo].

e. OK [$_{CP3}$ sake-o$_i$ Yui-ga [$_{CP2}$ 2-hon$_h$ Ken-wa [$_{CP1}$ Mari-ga naze t_i t_h (hiya-de) furumatta-to] omotta-ka] shiritagatteiru-yo].

f. OK [$_{CP3}$ 2-hon$_h$ Yui-ga [$_{CP2}$ sake-o$_i$ Ken-wa [$_{CP1}$ Mari-ga naze t_i t_h (hiya-de) furumatta-to] omotta-ka] shiritagatteiru-yo].

(16) a. [$_{CP3}$ Yui-ga [$_{CP2}$ sake-o$_i$ hiya-de$_g$ Ken-wa [$_{CP1}$ Mari-ga naze t_i t_g (hiya-de) furumatta-to] omotta-ka] shiritagatteiru-yo].

b. [$_{CP3}$ Yui-ga [$_{CP2}$ hiya-de$_g$ sake-o$_i$ Ken-wa [$_{CP1}$ Mari-ga naze t_i t_g (hiya-de) furumatta-to] omotta-ka] shiritagatteiru-yo].

c. [$_{CP3}$ sake-o$_i$ hiya-de$_g$ Yui-ga [$_{CP2}$ Ken-wa [$_{CP1}$ Mari-ga naze t_i t_g (hiya-de) furumatta-to] omotta-ka] shiritagatteiru-yo].

d. [$_{CP3}$ hiya-de$_g$ sake-o$_i$ Yui-ga [$_{CP2}$ Ken-wa [$_{CP1}$ Mari-ga naze t_i t_g (hiya-de) furumatta-to] omotta-ka] shiritagatteiru-yo].

e. OK [$_{CP3}$ sake-o$_i$ Yui-ga [$_{CP2}$ hiya-de$_g$ Ken-wa [$_{CP1}$ Mari-ga naze t_i t_g (hiya-de) furumatta-to] omotta-ka] shiritagatteiru-yo].

f. OK [$_{CP3}$ hiya-de$_g$ Yui-ga [$_{CP2}$ sake-o$_i$ Ken-wa [$_{CP1}$ Mari-ga naze t_i t_g (hiya-de) furumatta-to] omotta-ka] shiritagatteiru-yo].

The SinC approach offers a straightforward explanation why the so-called BSML is not at work in these cases. Under this analysis, what we see in e–f examples in (15) and (16) is not the same kind of split multiple LDS taking place in (13) which is ruled out by BSML; what is taking place here is *scrambling out of scrambled phrase*, having the derivation depicted in (17), which, in terms of derivation, is just like scrambling of DP out of scrambled CP (18), which is readily possible (Saito 1985, a.o.).

(17) $[_{ZP}\ YP \ldots [_{XP} \ldots t_{YP} \ldots] \ldots t_{XP} \ldots]$

(18) a. [$_{CP3}$ Aya-ni$_j$ Yui-ga [$_{CP2}$ [$_{CP1}$ Mari-ga naze t_j sake-o (2-hon/hiya-de) ageta-to]$_k$ Ken-wa t_k omotta-ka] shiritagatteiru-yo].

b. [$_{CP3}$ Sake-o$_i$ Yui-ga [$_{CP2}$ [$_{CP1}$ Mari-ga naze Aya-ni t_i (2-hon/hiya-de) ageta-to]$_k$ Ken-wa t_k omotta-ka] shiritagatteiru-yo].

Thus, whatever mechanisms that is responsible for BSML is not applicable for multiple application of LDS involved in (15)e–f, (16)e–f, and (18) since these are not "split" multiple LDS in terms of derivational procedure.

Under the InC approach, on the other hand, the multiple application of LDS involved in (15)e–f, (16)e–f, and (18) is the same kind of multiple application of LDS involving IO and DO in (13); i.e., it is an instance of "split" multiple LDS involving BSML. Thus, there is no way to tease apart the difference, failing to account for the contrast.

In sum, both (i) FNQs and its host DPs (15) and (ii) 2ndPs and its host DPs (16) show exactly the same behaviors with respect to the scrambling paradigm, i.e., *BSML*. While the SinC analysis can capture the difference between multiple LDS of FNQ or 2ndP and its host DP and multiple LDS of IO and DO (accounting for the absence/presence of BSML paradigm), the InC analysis cannot (failing to account for the absence/presence of BSML paradigm). But what is crucial is that FNQs and 2ndPs should be treated uniformly (supporting Miyagawa's (1989) insight), but it needs to be re-unified (departing from Miyagawa's analysis); they *can* be *adnominal adjuncts* which is adjoined to its host DP forming a base-generated single constituent, having the structure depicted in (2) above.

4 Theoretical Issue: On the Free Application of Merge

In this section, building on yet another paradigm of multiple long-distance scrambling (i.e., BSML) involving both FNQs and 2ndPs, I will discuss theoretical issues regarding merge and claim that the paradigm in question instantiates the free application of merge developed in Chomsky 2013, 2015, and Chomsky et. al. 2019.

Before discussing the crucial paradigm, let us first note that (as implicitly hinted in some of the examples in above) (i) FNQs and 2ndPs can co-occur and (ii) the order among FNQs, 2ndPs, and the host DPs are flexible as in (19).

(19) a. sake-o 2-hon hiya-de
b. sake-o hiya-de 2-hon
c. 2-hon sake-o hiya-de
d. hiya-de sake-o 2-hon
e. 2-hon hiya-de sake-o
f. hiya-de 2-hon sake-o

Now I propose that, extending the already proposed structure of FNQs and 2ndPs in (2), that the flexible word order in (19) *can* be generated by free application of merge proposed and developed in Chomsky 2013, 2015, and Chomsky et. al. 2019. More specifically, FNQs and 2ndP can either be left-adjoined and right-adjoined via External Pair Merge with its host DP as depicted in (20).

(20) a. $[_{DP}[_{DP}$ sake-o$]$ $[_{ClP}$ 2-hon$]$ $[_{AdvP}$ hiya-de$]]$
b. $[_{DP}[_{DP}$ sake-o$]$ $[_{AdvP}$ hiya-de$]$ $[_{ClP}$ 2-hon$]]$
c. $[_{DP}[_{ClP}$ 2-hon$]$ $[_{DP}$ sake-o$]$ $[_{AdvP}$ hiya-de$]]$
d. $[_{DP}[_{AdvP}$ hiya-de$][_{DP}$ sake-o$]$ $[_{ClP}$ 2-hon$]]$
e. $[_{DP}[_{ClP}$ 2-hon$]$ $[_{AdvP}$ hiya-de$]$ $[_{DP}$ sake-o$]]$
f. $[_{DP}[_{AdvP}$ hiya-de$]$ $[_{ClP}$ 2-hon$]$ $[_{DP}$ sake-o$]]$

With this in mind, let us look at the scrambling paradigm involving both FNQs and 2ndPs co-occur, which is schematically illustrated below, by focusing on cases in (19)a=(20)a and (19)b=(20)b.[8]

(21) a. Super-LDS of FNQ and LDS of 2ndP
OK $[_{CP3}$ *FNQ …* $[_{CP2}$ t_{FNQ} *2ndP …* $[_{CP1}$ *… OBJ* t_{FNQ} t_{2ndP} *…]]]*
b. Super-LDS of 2ndP and LDS of FNQ
OK $[_{CP3}$ *2ndP …* $[_{CP2}$ *FNQ* t_{2ndP} *…* $[_{CP1}$ *… OBJ* t_{FNQ} t_{2ndP} *…]]]*

What is of interest is that, these FNQs and 2ndPs need not be adjacent and can undergo (super-)LDS ending up in non-adjacent positions yielding split multiple LDS on the surface; yet, the BSML is not at work here.

These possible cases raises a potential problem to the SinC analysis which assigns the structure in (20), which simply adjoins FNQ and 2ndP to the host DP. This is so because, under this structure, two adjuncts do not form a base-generated single constituent. What can form a base-generated single constituent (and which in turn can be the target of syntactic operations based on segment; recall (5)c) with the structure (20) is depicted in (22) and (23) respectively, where the box indicates the possible constituency.

(22) a. $[_{DP}\boxed{[_{DP}$ sake-o$]}$ $[_{ClP}$ 2-hon$]$ $[_{AdvP}$ hiya-de$]]$
b $[_{DP}\boxed{[_{DP}$ sake-o$]$ $[_{ClP}$ 2-hon$]}$ $[_{AdvP}$ hiya-de$]]$
c. $\boxed{[_{DP}[_{DP}$ sake-o$]$ $[_{ClP}$ 2-hon$]$ $[_{AdvP}$ hiya-de$]]}$

(23) a. $[_{DP}\boxed{[_{DP}$ sake-o$]}$ $[_{AdvP}$ hiya-de$]$ $[_{ClP}$ 2-hon$]]$
b. $[_{DP}\boxed{[_{DP}$ sake-o$]$ $[_{AdvP}$ hiya-de$]}$ $[_{ClP}$ 2-hon$]]$
c. $\boxed{[_{DP}[_{DP}$ sake-o$]$ $[_{AdvP}$ hiya-de$]$ $[_{ClP}$ 2-hon$]]}$

The point is that if (22) and (23) are the only available structure, FNQs and 2ndPs must undergo the split multiple LDS just like that of IOs and DOs in (13), which end up in BSML.

But the problem is only apparent, and I argue that the possible cases can be handled properly under the SinC analysis. Then what kind of structure is formed for cases involving FNQs, 2ndPs, and the host DPs to account for the paradigm under discussion? I propose, based on the free application of merge advocated in Chomsky 2013, 2015, and Chomsky et. al. 2019, that in addition to the "normal" case where both FNQs and 2ndPs are adjoined (either leftward and/or rightward) to its host DP (20)/ (22)/(23), (24) is possible where FNQ and 2ndP are adjoined to each other first, and then this amalgam as a whole is adjoined to the DP. (25) and (26) depict

[8] Due to space limitation, I can only provide the schematics. I also note here that paradigm remain the same for c–f examples.

what can form a base-generated single constituent (and which in turn can be the target of syntactic operations based on segment; recall (5)c), where the box indicates the base-generated constituency.

(24) a. [DP [DP host DP][[ClP FNQ] [AdvP 2ndP]]]
a'. [DP [DP sake-o] [[ClP 2-hon] [AdvP hiya-de]]]
b. [DP [DP host DP][[AdvP 2ndP] [ClP FNQ]]]
b'. [DP [DP sake-o] [[AdvP hiya-de] [ClP 2-hon]]]

(25) a. [DP [DP sake-o] [[ClP 2-hon] [AdvP hiya-de]]]
b. [DP [DP sake-o] [[ClP 2-hon] [AdvP hiya-de]]]
c. [DP [DP sake-o] [[ClP 2-hon] [AdvP hiya-de]]]

(26) a. [DP [DP sake-o] [[AdvP hiya-de] [ClP 2-hon]]]
b. [DP [DP sake-o] [[AdvP hiya-de] [ClP 2-hon]]]
c. [DP [DP sake-o] [[AdvP hiya-de] [ClP 2-hon]]]

What is crucial is that with the structure (24), two adjuncts form a base-generated single constituent (as in (25)b and (26)b), and hence may be the target of syntactic operations. This makes it possible to allow the derivation where "FNQ+2ndP" or "2ndP+FNQ" first undergoes LDS as a single constituent, and then one of them is scrambled out undergoing super-LDS as depicted in (27).

(27) a. Super-LDS of FNQ and LDS of 2ndP
OK [CP3 *FNQ* ... [CP2 t_{FNQ} *2ndP* ...
[CP1 ... *OBJ* t_{FNQ} t_{2ndP} ...]]]
b. Super-LDS of 2ndP and LDS of FNQ
OK [CP3 *2ndP* ... [CP2 *FNQ* t_{2ndP} ...
[CP1 ...*OBJ* t_{FNQ} t_{2ndP} ...]]]

Thus, despite the surface, (21) is not an instance of the BSML.

To sum up, the absence of BSML with what seems like a split multiple LDS involving FNQ and 2ndP illustrated in (21) is accounted for under the SinC analysis since it involves the derivation depicted in (27) which shows that it is irrelevant to the BSML, which in turn is made available by the free application of merge, especially External Pair Merge (i.e., adjunction in the classical terminology), developed in Chomsky 2013, 2015, and Chomsky et. al. 2019, where selection (in a broad sense) can be irrelevant to the application of merge.

5 Conclusion

To conclude, building on the evidence involving (i) the argument/adjunct asymmetry on argument ellipsis and (ii) the ban on split multiple long-distance scrambling, I argued for the Single Constituent (SinC) analysis for the floating numeral quantifiers (FNQs) and the secondary predicates (2ndPs) in Japanese; these adjuncts in Japanese *can* be *adnominal adjuncts* which is adjoined to its host DP it is associated with, forming a base-generated single constituent, and the host DP as a result functions as a lower segment of DP, as depicted in (2) above. I also discussed cases where both FNQ and 2ndP co-occur with its host DP and its theoretical implication, and showed that the proposed analysis shed lights to the recent theory of merge allowing its free application developed in Chomsky 2013, 2015, and Chomsky et. al. 2019.

Last but not the least, although the SinC analysis developed in this paper counters with the classic analyses that treats FNQs and 2ndPs as adverbial adjuncts which are externally pair merged with vP/VP and not with DP – the Independent Constituent (InC) analysis (Miyagawa 1989, Koizumi 1994, a.o.)) –, to the extent that this analysis is on the right track, it re-unified the treatment of FNQs and 2ndPs in Japanese as in (3), reproduced here as (28), updating the unification first pursued in Miyagawa 1989 (4), reproduced here as (29).

(28) SinC (Single Constituent) analysis
(Re-unification of FNQs and 2ndPs):
FNQs and 2ndPs can form a base-generated single constituent with its host DPs; they can be adjoined to DP ((2)).

(29) InC (Independent Constituent) analysis
(Unification of FNQs and 2ndPs):
FNQs and 2ndPs do not form a base-generated single constituent with its host DPs; they are adjoined to vP/VP ((1)).

Thus, although the exact analysis is different and essentially contradictory, it nonetheless provides support for Miyagawa's insight that FNQs and 2ndPs in Japanese are of the same syntactic species and hence these two elements calls for a unification.

References

Chomsky, Noam. 1995. *The Minimalist Program*. Cambridge, MA: MIT Press.

Chomsky, Noam. 2013. Problems of projection. *Lingua* 130: 33–49.

Chomsky, Noam. 2015. Problems of projection: Extensions. In *Structures, Strategies and Beyond – Studies in Honor of Adriana Belletti*, eds. Elisa Di Domenico, Cornelia Hamann, and Simona Matteini, 3–16. Amsterdam/Philadelphia: John Benjamins.

Chomsky, Noam, Ángel J. Gallego, and Dennis Ott. 2019. Generative grammar and the faculty of language: insights, questions, and challenges. To appear in *Generative Syntax: Questions, Crossroads, and Challenges. Special issue of Catalan Journal of Linguistics*, eds. Ángel J. Gallego, and Dennis Ott. [lingbuzz/003507]

Funakoshi, Kenshi. 2016. Verb-stranding VP ellipsis in Japanese. *Journal of East Asian Linguistics* 25: 113–142.

Koizumi, Masatoshi. 1994. Secondary predicates. *Journal of East Asian Linguistics* 3: 25–79.

Miyagawa, Shigeru. 1989. *Structure and Case Marking in Japanese*. San Diego: Academic Press.

Oku, Satoshi. 1998. A theory of selection and reconstruction in the Minimalist Program. Doctoral dissertation, University of Connecticut, Storrs.

Richards, Norvin. 1998. The Principle of Minimal Compliance. *Linguistic Inquiry* 29: 599–629.

Saito, Mamoru. 1985. Some asymmetries in Japanese and their theoretical implications. Doctoral dissertation, MIT.

Saito, Mamoru. 2007. Notes on East Asian argument ellipsis. *Language Research* 43: 203–227.

Saito, Mamoru and Naoki Fukui. 1998. Order in phrase structure and movement. Linguistic Inquiry 29: 439–474.

Sakai, Hiromu. 1994. Derivational economy in long distance scrambling. In *Formal Approaches to Japanese Linguistics 1: Proceedings of the First Conference on Formal Approaches to Japanese Linguistics*, eds. Hiroyuki Ura and Masatoshi Koizumi, 295–314. Cambridge: MITWPL.

Sakamoto, Yuta. 2016. Phases and argument ellipsis in Japanese. *Journal of East Asian Linguistics* 25: 243–274.

Sakamoto, Yuta. 2017. Escape from silent syntax. Doctoral dissertation, University of Connecticut, Storrs.

Sakamoto, Yuta. 2019. Overtly empty but covertly complex. *Linguistic Inquiry* 50: 105–136.

Shinohara, Michie. 2006. On some differences between the major deletion phenomena and Japanese argument ellipsis. Ms., Nanzan University.

Sohn, Keun-Won. 1994. Adjunction to arguments, Free Ride, and a Minimalist Program. In *Formal Approaches to Japanese Linguistics 1: Proceedings of the First Conference on Formal Approaches to Japanese Linguistics*, eds. Hiroyuki Ura and Masatoshi Koizumi, 315–334. Cambridge, MA: MITWPL.

Takahashi, Daiko. 2008a. Noun phrase ellipsis. In *Handbook of Japanese Linguistics*, eds. Shigeru Miyagawa and Mamoru Saito, 395–423. Oxford: Oxford University Press.

Takahashi, Daiko. 2008b. Quantificational null objects and argument ellipsis. *Linguistic Inquiry* 39, 307–326.

Takahashi, Daiko. 2014. Argument ellipsis, anti-agreement, and scrambling. In *Japanese Syntax in Comparative Perspective*, ed. Mamoru Saito, 88–116. Oxford: Oxford University Press.

Takano, Yuji. 2002. Surprising constituent. *Journal of East Asian Linguistics* 11, 243–301.

Yamashita, Hideaki. 2014. On the nature of Condition on Extraction out of Argument Ellipsis Site. Poster presented at *the 7th Formal Approaches to Japanese Linguistics*, NINJAL and International Christian University.

Yamashita, Hideaki. 2015. Reconsidering the constituency of floating numeral quantifiers in Japanese. In *Handbook of the 151st Meeting of the Linguistic Society of Japan*, 312–317. Kyoto: Linguistic Society of Japan.

Yamashita, Hideaki. 2016. On the constituency and structure of floating numeral quantifiers in Japanese. In *Proceedings of FAJL 8: Formal Approaches to Japanese Linguistics*, eds. Ayaka Sugawara, Shintaro Hayashi, and Satoshi Ito, 209–220. Cambridge, MA: MITWPL.

Yamashita, Hideaki. 2019. On the ellipsis of subject- and object-oriented adverbs in Japanese. In *Handbook of the 158th Meeting of the Linguistic Society of Japan*, 190–196. Kyoto: Linguistic Society of Japan.