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.

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-

<table>
<thead>
<tr>
<th>Language</th>
<th>Word</th>
</tr>
</thead>
<tbody>
<tr>
<td>JA-character</td>
<td>風景</td>
</tr>
<tr>
<td>JA-ideograph</td>
<td>回几亜之日京</td>
</tr>
</tbody>
</table>
| JA-stroke     | 回田払乙日顷回手
            | 一曰曰一、
            | 〇回回回々手一— … |
| ZH-character | 風景 |
| ZH-ideograph  | 回几亜之日京 |
| ZH-stroke     | 回田払乙日顷
            | 〇回回回々手一— … |
| EN          | landscape |

Table 1: Examples of decomposition of a Japanese word “風景” and Chinese word “风景,” both meaning “landscape” in English.
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 ‘㇐’, ‘㇑’, ‘㇓’, and ‘㇝’) which compose hundreds of ideographs, and more than 90,000 different characters. Table 2 shows examples of how strokes and ideographs compose different characters.

<table>
<thead>
<tr>
<th>Character</th>
<th>Semantic ideograph</th>
<th>Phonetic ideograph</th>
<th>Pinyin</th>
</tr>
</thead>
<tbody>
<tr>
<td>驰 run</td>
<td>马 horse</td>
<td>也</td>
<td>chí</td>
</tr>
<tr>
<td>池 pool</td>
<td>水( chí) water</td>
<td>也</td>
<td>chí</td>
</tr>
<tr>
<td>施 impose</td>
<td>方 direction</td>
<td>也</td>
<td>shī</td>
</tr>
<tr>
<td>弩 loosen</td>
<td>弓 bow</td>
<td>也</td>
<td>chí</td>
</tr>
<tr>
<td>地 land</td>
<td>土 soil</td>
<td>也</td>
<td>dì</td>
</tr>
<tr>
<td>驱 drive</td>
<td>马 horse</td>
<td>区</td>
<td>qū</td>
</tr>
</tbody>
</table>

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:

$$PE_{(pos, 2i)} = \sin(pos/10000^{2i/d_{model}})$$
$$PE_{(pos, 2i+1)} = \cos(pos/10000^{2i/d_{model}})$$

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.”

$$Attention(Q, K, V) = softmax(QK^T/\sqrt{d_k})V$$
$$MultiHead(Q, K, V) = Concat(h_1, ..., h_i)W^o$$
$$h_i = Attention(QW_i^Q, KW_i^K, VW_i^V)$$
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.

\[
F F N(x) = \max(0, xW_1 + b_1)W_2 + b_2
\] (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 ⇒ L1 encoder ⇒ L2 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, a 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 strokes3.

---

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.

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
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.

<table>
<thead>
<tr>
<th>Granularity</th>
<th>JA–ZH</th>
<th>ZH–JA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Character</td>
<td>24.18 (29.60)</td>
<td>29.79 (40.00)</td>
</tr>
<tr>
<td>Ideograph</td>
<td></td>
<td></td>
</tr>
<tr>
<td>w/ IDCs</td>
<td>25.76*</td>
<td>32.61*</td>
</tr>
<tr>
<td>w/o IDCs</td>
<td>25.14* (32.00)</td>
<td>32.17* (42.60)</td>
</tr>
<tr>
<td>Stroke</td>
<td></td>
<td></td>
</tr>
<tr>
<td>w/ IDCs</td>
<td>26.39*</td>
<td>32.99*</td>
</tr>
<tr>
<td>w/o IDCs</td>
<td>24.75* (32.10)</td>
<td>30.59* (42.20)</td>
</tr>
</tbody>
</table>

Table 4: BLEU scores with different token sharing rates on test set.

<table>
<thead>
<tr>
<th>r</th>
<th>JA–ZH</th>
<th>ZH–JA</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>19.72</td>
<td>25.23</td>
</tr>
<tr>
<td>0.7</td>
<td>23.60</td>
<td>28.32</td>
</tr>
<tr>
<td>0.9</td>
<td>23.04</td>
<td>28.84</td>
</tr>
</tbody>
</table>

**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.

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.
Table 5: Translation examples from three UNMT models in six translation directions.

<table>
<thead>
<tr>
<th>Type</th>
<th>Sentence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference–JA</td>
<td>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。</td>
</tr>
<tr>
<td>Reference–ZH</td>
<td>图 3 所示的是 2 个关于判断“会”是否是固有表达的例句。</td>
</tr>
<tr>
<td>Character–JA</td>
<td>図 3 に示すような 2 つの判断については「会」が 固有 表現で ある か 否 か を 判定する例文を示す。</td>
</tr>
<tr>
<td>Character–ZH</td>
<td>图 3 中显示了判断“会”是固有名词还是有 2 个例句。</td>
</tr>
<tr>
<td>Ideograph–JA</td>
<td>図 3 に示すように 2 つの判断については「会」が 固有 表現で ある か 否 か を 判定する例文を示す。</td>
</tr>
<tr>
<td>Ideograph–ZH</td>
<td>图 3 中显示了判断“会”是固有名词的 2 个例句。</td>
</tr>
<tr>
<td>Stroke–JA</td>
<td>図 3 に示すのは、2 つの判断については「会」が 固有 表現で ある か 否 か を 判定する例文を示す。</td>
</tr>
<tr>
<td>Stroke–ZH</td>
<td>图 3 中显示了判断“会”是固有表达的 2 个例句。</td>
</tr>
</tbody>
</table>

English: Figure 3 showed 2 example sentences of judging whether “会” is an inherent expression.

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 “名詞” (“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


