Explicit Contextual Semantics for Text Comprehension

Zhuosheng Zhang\textsuperscript{1,2,3,*}, Yuwei Wu\textsuperscript{1,2,3,4,*}, Zuchao Li\textsuperscript{1,2,3}, Hai Zhao\textsuperscript{1,2,3,†}

\textsuperscript{1}Department of Computer Science and Engineering, Shanghai Jiao Tong University
\textsuperscript{2}Key Laboratory of Shanghai Education Commission for Intelligent Interaction and Cognitive Engineering, Shanghai Jiao Tong University, Shanghai, China
\textsuperscript{3}MoE Key Lab of Artificial Intelligence, AI Institute, Shanghai Jiao Tong University, Shanghai, China
\textsuperscript{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, \textit{machine reading comprehension} (MRC) and \textit{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-
Harvard was a founding member of the Association of American Universities in 1900. James Bryant Conant led the university through the Great Depression and World War II and began to reform the curriculum and liberalize admissions after the war. The undergraduate college became coeducational after its 1977 merger with Radcliffe College.

**Question**

What was the name of the leader through the Great Depression and World War II?

**Answer**

James Bryant Conant

---

Besides, 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...
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.

### Table 1: A machine reading comprehension example.

<table>
<thead>
<tr>
<th>Passage</th>
<th>Question</th>
<th>Answer</th>
</tr>
</thead>
<tbody>
<tr>
<td>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.</td>
<td>What changes the mineral content of a rock?</td>
<td>heat and pressure.</td>
</tr>
</tbody>
</table>

### Table 2: A textual entailment example.

<table>
<thead>
<tr>
<th>Premise</th>
<th>Hypo.</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>A man parasails in the choppy water.</td>
<td>The man is competing in a competition.</td>
<td>Neutral</td>
</tr>
<tr>
<td>The man parasailed in the calm water.</td>
<td>The man parasailed.</td>
<td>Contradiction</td>
</tr>
<tr>
<td>The water was choppy as the man parasailed.</td>
<td>The water was choppy as the man parasailed.</td>
<td>Entailment</td>
</tr>
</tbody>
</table>

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,

\[
[\text{ARG}_0 \text{ Charlie}] \ [V \text{ sold}] \ [\text{ARG}_1 \text{ a book}] \ [\text{ARG}_2 \text{ to Sherry}] \ [AM-\text{\_\_\_\_}_\text{last week}].
\]

where \text{ARG}_0\text{ represents the seller (agent)}, \text{ARG}_1\text{ represents the thing sold (theme)}, \text{ARG}_2\text{ represents the buyer (recipient), AM-\_\_\_\_}_\text{last week} is an adjunct indicating the timing of the action and \text{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.
Word Representation  The word representation of our SRL model is the concatenation of two vectors: an ELMo embedding $e^{(l)}$ and predicate indicator embedding (PIE) $e^{(p)}$. 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 $e^{(p)}$ to mark whether a word is a predicate when predicting and labeling the arguments. The final word representation is given by $e = e^{(l)} \oplus e^{(p)}$, 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 $S = \{e_1, e_2, \cdots, e_n\}$ as input, the hidden state $h = \{h_1, h_2, \cdots, h_n\}$ 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,

$$y^* = \arg\max_{\tilde{y} \in C} s(x, \tilde{y})$$

where $C$ 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% on the test set. At test time, we perform Viterbi decoding to enforce valid spans using BIO constraints. 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 labels.
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 task6. 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.

<table>
<thead>
<tr>
<th>Model</th>
<th>Accuracy (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep Gated Attn. BiLSTM</td>
<td>85.5</td>
</tr>
<tr>
<td>Gumbel TreeLSTM</td>
<td>86.0</td>
</tr>
<tr>
<td>Residual stacked</td>
<td>86.0</td>
</tr>
<tr>
<td>Distance-based SAN</td>
<td>86.3</td>
</tr>
<tr>
<td>BCN + CoVe + Char</td>
<td>88.1</td>
</tr>
<tr>
<td>DIIN</td>
<td>88.0</td>
</tr>
<tr>
<td>DR-BiLSTM</td>
<td>88.5</td>
</tr>
<tr>
<td>CAFE</td>
<td>88.5</td>
</tr>
<tr>
<td>MAN</td>
<td>88.3</td>
</tr>
<tr>
<td>KIM</td>
<td>88.6</td>
</tr>
<tr>
<td>DMAN</td>
<td>88.8</td>
</tr>
<tr>
<td>ESIM + TreeLSTM</td>
<td>88.6</td>
</tr>
<tr>
<td>ESIM + ELMo</td>
<td>88.7</td>
</tr>
<tr>
<td>DCRCN</td>
<td>88.9</td>
</tr>
<tr>
<td>LM-Transformer</td>
<td>89.9</td>
</tr>
<tr>
<td>MT-DNN†</td>
<td>91.1</td>
</tr>
<tr>
<td>Baseline (ELMo)</td>
<td>88.4</td>
</tr>
<tr>
<td>+ SRL</td>
<td>89.1</td>
</tr>
<tr>
<td>Baseline (BERTBASE)</td>
<td>89.2</td>
</tr>
<tr>
<td>+ SRL</td>
<td>89.6</td>
</tr>
<tr>
<td>Baseline (BERTLARGE)</td>
<td>90.4</td>
</tr>
<tr>
<td>+ SRL</td>
<td>91.3</td>
</tr>
</tbody>
</table>

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
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. 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 BERTLARGE model also achieves a new state-of-the-art, even outperforms all the ensemble models in the leaderboard. 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. 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,
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 †.

<table>
<thead>
<tr>
<th>Model</th>
<th>EM</th>
<th>F1</th>
<th>RERR</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Published</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MQAN_{single-task}</td>
<td>-</td>
<td>75.5</td>
<td>-</td>
</tr>
<tr>
<td>MQAN_{multi-task}</td>
<td>-</td>
<td>74.3</td>
<td>-</td>
</tr>
<tr>
<td>BiDAF+ELMo</td>
<td>-</td>
<td>85.6</td>
<td>-</td>
</tr>
<tr>
<td>R.M. Reader</td>
<td>78.9</td>
<td>86.3</td>
<td>-</td>
</tr>
<tr>
<td>BERT_{BASE}</td>
<td>80.8</td>
<td>88.5</td>
<td>-</td>
</tr>
<tr>
<td>BERT_{LARGE}†</td>
<td>84.1</td>
<td>90.9</td>
<td>-</td>
</tr>
</tbody>
</table>

*Our implementation*

<table>
<thead>
<tr>
<th></th>
<th>EM</th>
<th>F1</th>
<th>RERR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline (ELMo)</td>
<td>77.5</td>
<td>85.2</td>
<td>-</td>
</tr>
<tr>
<td>+SRL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline (BERT_{BASE})</td>
<td>81.3</td>
<td>88.5</td>
<td>-</td>
</tr>
<tr>
<td>+SRL</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline (BERT_{LARGE})</td>
<td>84.2</td>
<td>90.9</td>
<td>-</td>
</tr>
<tr>
<td>+SRL</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

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 hyperparameter 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.

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.

Table 6: Comparison with different NLP tags.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dev</th>
<th>Test</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>88.89</td>
<td>88.65</td>
</tr>
<tr>
<td><strong>Word + SRL</strong></td>
<td><strong>89.11</strong></td>
<td><strong>89.09</strong></td>
</tr>
<tr>
<td>Word + POS</td>
<td>88.90</td>
<td>88.68</td>
</tr>
<tr>
<td>Word + NE</td>
<td>89.14</td>
<td>88.51</td>
</tr>
</tbody>
</table>
References


Robin Jia and Percy Liang. Adversarial examples for evaluating reading comprehension systems.


Chen Shi, Shujie Liu, Shuo Ren, Shi Feng, Mu Li, Ming Zhou, Xu Sun, and Houfeng Wang.


