Neural Relation Extraction with Selective Attention over Instances

Yankai Lin[1], **Shiqi Shen**[1], **Zhiyuan Liu**[1,2]*, **Huanbo Luan**[1], **Maosong Sun**[1,2]
[1] Department of Computer Science and Technology,
State Key Lab on Intelligent Technology and Systems,
National Lab for Information Science and Technology, Tsinghua University, Beijing, China
[2] Jiangsu Collaborative Innovation Center for Language Competence, Jiangsu, China

Abstract

Distant supervised relation extraction has been widely used to find novel relational facts from text. However, distant supervision inevitably accompanies with the wrong labelling problem, and these noisy data will substantially hurt the performance of relation extraction. To alleviate this issue, we propose a sentence-level attention-based model for relation extraction. In this model, we employ convolutional neural networks to embed the semantics of sentences. Afterwards, we build sentence-level attention over multiple instances, which is expected to dynamically reduce the weights of those noisy instances. Experimental results on real-world datasets show that, our model can make full use of all informative sentences and effectively reduce the influence of wrong labelled instances. Our model achieves significant and consistent improvements on relation extraction as compared with baselines. The source code of this paper can be obtained from `https://github.com/thunlp/NRE`.

1 Introduction

In recent years, various large-scale knowledge bases (KBs) such as Freebase (Bollacker et al., 2008), DBpedia (Auer et al., 2007) and YAGO (Suchanek et al., 2007) have been built and widely used in many natural language processing (NLP) tasks, including web search and question answering. These KBs mostly compose of relational facts with triple format, e.g., (*Microsoft*, `founder`, *Bill Gates*). Although existing KBs contain a massive amount of facts, they are still far from complete compared to the infinite real-world facts. To enrich KBs, many efforts have been invested in automatically finding unknown relational facts. Therefore, relation extraction (RE), the process of generating relational data from plain text, is a crucial task in NLP.

Most existing supervised RE systems require a large amount of labelled relation-specific training data, which is very time consuming and labor intensive. (Mintz et al., 2009) proposes distant supervision to automatically generate training data via aligning KBs and texts. They assume that if two entities have a relation in KBs, then all sentences that contain these two entities will express this relation. For example, (*Microsoft*, `founder`, *Bill Gates*) is a relational fact in KB. Distant supervision will regard all sentences that contain these two entities as active instances for relation `founder`. Although distant supervision is an effective strategy to automatically label training data, it always suffers from wrong labelling problem. For example, the sentence "*Bill Gates 's turn to philanthropy was linked to the antitrust problems Microsoft had in the U.S. and the European union.*" does not express the relation `founder` but will still be regarded as an active instance. Hence, (Riedel et al., 2010; Hoffmann et al., 2011; Surdeanu et al., 2012) adopt multi-instance learning to alleviate the wrong labelling problem. The main weakness of these conventional methods is that most features are explicitly derived from NLP tools such as POS tagging and the errors generated by NLP tools will propagate in these methods.

Some recent works (Socher et al., 2012; Zeng et al., 2014; dos Santos et al., 2015) attempt to use deep neural networks in relation classification without handcrafted features. These methods build classifier based on sentence-level annotated data, which cannot be applied in large-scale

[*] Corresponding author: Zhiyuan Liu (liuzy@tsinghua.edu.cn).

Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, pages 2124–2133,
Berlin, Germany, August 7-12, 2016. ©2016 Association for Computational Linguistics

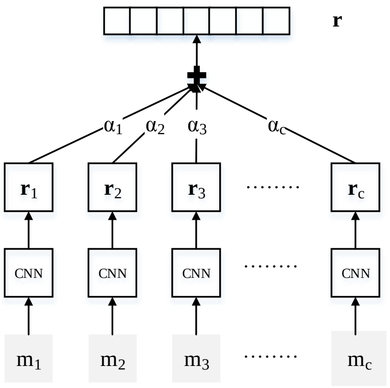

Figure 1: The architecture of sentence-level attention-based CNN, where m_i indicates the original sentence for an entity pair, α_i is the weight given by sentence-level attention.

KBs due to the lack of human-annotated training data. Therefore, (Zeng et al., 2015) incorporates multi-instance learning with neural network model, which can build relation extractor based on distant supervision data. Although the method achieves significant improvement in relation extraction, it is still far from satisfactory. The method assumes that at least one sentence that mentions these two entities will express their relation, and only selects the most likely sentence for each entity pair in training and prediction. It's apparent that the method will lose a large amount of rich information containing in neglected sentences.

In this paper, we propose a sentence-level attention-based convolutional neural network (CNN) for distant supervised relation extraction. As illustrated in Fig. 1, we employ a CNN to embed the semantics of sentences. Afterwards, to utilize all informative sentences, we represent the relation as semantic composition of sentence embeddings. To address the wrong labelling problem, we build sentence-level attention over multiple instances, which is expected to dynamically reduce the weights of those noisy instances. Finally, we extract relation with the relation vector weighted by sentence-level attention. We evaluate our model on a real-world dataset in the task of relation extraction. The experimental results show that our model achieves significant and consistent improvements in relation extraction as compared with the state-of-the-art methods.

The contributions of this paper can be summarized as follows:

- As compared to existing neural relation extraction model, our model can make full use of all informative sentences of each entity pair.
- To address the wrong labelling problem in distant supervision, we propose selective attention to de-emphasize those noisy instances.
- In the experiments, we show that selective attention is beneficial to two kinds of CNN models in the task of relation extraction.

2 Related Work

Relation extraction is one of the most important tasks in NLP. Many efforts have been invested in relation extraction, especially in supervised relation extraction. Most of these methods need a great deal of annotated data, which is time consuming and labor intensive. To address this issue, (Mintz et al., 2009) aligns plain text with Freebase by distant supervision. However, distant supervision inevitably accompanies with the wrong labelling problem. To alleviate the wrong labelling problem, (Riedel et al., 2010) models distant supervision for relation extraction as a multi-instance single-label problem, and (Hoffmann et al., 2011; Surdeanu et al., 2012) adopt multi-instance multi-label learning in relation extraction. Multi-instance learning was originally proposed to address the issue of ambiguously-labelled training data when predicting the activity of drugs (Dietterich et al., 1997). Multi-instance learning considers the reliability of the labels for each instance. (Bunescu and Mooney, 2007) connects weak supervision with multi-instance learning and extends it to relation extraction. But all the feature-based methods depend strongly on the quality of the features generated by NLP tools, which will suffer from error propagation problem.

Recently, deep learning (Bengio, 2009) has been widely used for various areas, including computer vision, speech recognition and so on. It has also been successfully applied to different NLP tasks such as part-of-speech tagging (Collobert et al., 2011), sentiment analysis (dos Santos and Gatti, 2014), parsing (Socher et al., 2013), and machine translation (Sutskever et al., 2014). Due to the recent success in deep learning, many researchers have investigated the possibility of using neural networks to automatically learn features for relation extraction. (Socher et al., 2012) uses a recursive neural network in relation extraction. They parse the sentences first and then represent each node in the parsing tree as a vector. Moreover, (Zeng et al., 2014; dos Santos et al., 2015) adopt an end-to-end convolutional neural network for relation extraction. Besides, (Xie et al., 2016) attempts to incorporate the text information of entities for relation extraction.

Although these methods achieve great success, they still extract relations on sentence-level and suffer from a lack of sufficient training data. In addition, the multi-instance learning strategy of conventional methods cannot be easily applied in neural network models. Therefore, (Zeng et al., 2015) combines at-least-one multi-instance learning with neural network model to extract relations on distant supervision data. However, they assume that only one sentence is active for each entity pair. Hence, it will lose a large amount of rich information containing in those neglected sentences. Different from their methods, we propose sentence-level attention over multiple instances, which can utilize all informative sentences.

The attention-based models have attracted a lot of interests of researchers recently. The selectivity of attention-based models allows them to learn alignments between different modalities. It has been applied to various areas such as image classification (Mnih et al., 2014), speech recognition (Chorowski et al., 2014), image caption generation (Xu et al., 2015) and machine translation (Bahdanau et al., 2014). To the best of our knowledge, this is the first effort to adopt attention-based model in distant supervised relation extraction.

3 Methodology

Given a set of sentences $\{x_1, x_2, \cdots, x_n\}$ and two corresponding entities, our model measures the probability of each relation r. In this section, we will introduce our model in two main parts:

- **Sentence Encoder.** Given a sentence x and two target entities, a convolutional neutral network (CNN) is used to construct a distributed representation $\mathbf{x}$ of the sentence.

- **Selective Attention over Instances**. When the distributed vector representations of all sentences are learnt, we use sentence-level attention to select the sentences which really express the corresponding relation.

3.1 Sentence Encoder

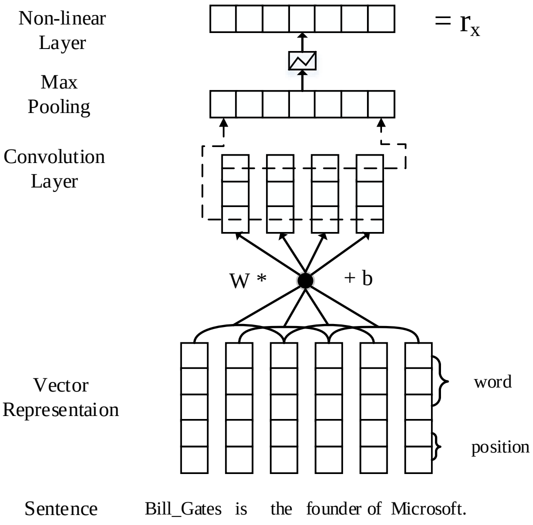

Figure 2: The architecture of CNN/PCNN used for sentence encoder.

As shown in Fig. 2, we transform the sentence x into its distributed representation $\mathbf{x}$ by a CNN. First, words in the sentence are transformed into dense real-valued feature vectors. Next, convolutional layer, max-pooling layer and non-linear transformation layer are used to construct a distributed representation of the sentence, i.e., $\mathbf{x}$.

3.1.1 Input Representation

The inputs of the CNN are raw words of the sentence x. We first transform words into low-dimensional vectors. Here, each input word is transformed into a vector via word embedding matrix. In addition, to specify the position of each entity pair, we also use position embeddings for all words in the sentence.

Word Embeddings. Word embeddings aim to transform words into distributed representations which capture syntactic and semantic meanings of the words. Given a sentence x consisting of m words $x = \{w_1, w_2, \cdots, w_m\}$, every word w_i is represented by a real-valued vector. Word representations are encoded by column vectors in an embedding matrix $\mathbf{V} \in \mathbb{R}^{d^a \times |V|}$where V is a fixed-sized vocabulary.

Position Embeddings. In the task of relation extraction, the words close to the target entities are usually informative to determine the relation between entities. Similar to (Zeng et al., 2014), we use position embeddings specified by entity pairs. It can help the CNN to keep track of how close each word is to head or tail entities. It is defined as the combination of the relative distances from the current word to head or tail entities. For example, in the sentence "Bill_Gates is the founder of Microsoft.", the relative distance from the word "founder" to head entity *Bill_Gates* is 3 and tail entity *Microsoft* is 2.

In the example shown in Fig. 2, it is assumed that the dimension d^a of the word embedding is 3 and the dimension d^b of the position embedding is 1. Finally, we concatenate the word embeddings and position embeddings of all words and denote it as a vector sequence $\mathbf{w} = \{\mathbf{w}_1, \mathbf{w}_2, \cdots, \mathbf{w}_m\}$, where $\mathbf{w}_i \in \mathbb{R}^d (d = d^a + d^b \times 2)$.

3.1.2 Convolution, Max-pooling and Non-linear Layers

In relation extraction, the main challenges are that the length of the sentences is variable and the important information can appear in any area of the sentences. Hence, we should utilize all local features and perform relation prediction globally. Here, we use a convolutional layer to merge all these features. The convolutional layer first extracts local features with a sliding window of length l over the sentence. In the example shown in Fig. 2, we assume that the length of the sliding window l is 3. Then, it combines all local features via a max-pooling operation to obtain a fixed-sized vector for the input sentence.

Here, convolution is defined as an operation between a vector sequence $\mathbf{w}$ and a convolution matrix $\mathbf{W} \in \mathbb{R}^{d^c \times (l \times d)}$, where d^c is the sentence embedding size. Let us define the vector $\mathbf{q}_i \in \mathbb{R}^{l \times d}$ as the concatenation of a sequence of w word embeddings within the i-th window:

$$\mathbf{q}_i = \mathbf{w}_{i-l+1:i} \quad (1 \le i \le m + l - 1). \tag{1}$$

Since the window may be outside of the sentence boundaries when it slides near the boundary, we set special padding tokens for the sentence. It means that we regard all out-of-range input vectors $\mathbf{w}_i (i < 1 \text{ or } i > m)$ as zero vector.

Hence, the i-th filter of convolutional layer is computed as:

$$\mathbf{p}_i = [\mathbf{W}\mathbf{q} + \mathbf{b}]_i \tag{2}$$

where $\mathbf{b}$ is bias vector. And the i-th element of the vector $\mathbf{x} \in \mathbb{R}^{d^c}$ as follows:

$$[\mathbf{x}]_i = \max(\mathbf{p}_i), \tag{3}$$

Further, PCNN (Zeng et al., 2015), which is a variation of CNN, adopts piecewise max pooling in relation extraction. Each convolutional filter $\mathbf{p}_i$ is divided into three segments $(\mathbf{p}_{i1}, \mathbf{p}_{i2}, \mathbf{p}_{i3})$ by head and tail entities. And the max pooling procedure is performed in three segments separately, which is defined as:

$$[\mathbf{x}]_{ij} = \max(\mathbf{p}_{ij}), \tag{4}$$

And $[\mathbf{x}]_i$ is set as the concatenation of $[\mathbf{x}]_{ij}$.

Finally, we apply a non-linear function at the output, such as the hyperbolic tangent.

3.2 Selective Attention over Instances

Suppose there is a set S contains n sentences for entity pair $(head, tail)$, i.e., $S = \{x_1, x_2, \cdots, x_n\}$.

To exploit the information of all sentences, our model represents the set S with a real-valued vector $\mathbf{s}$ when predicting relation r. It is straightforward that the representation of the set S depends on all sentences' representations $\mathbf{x}_1, \mathbf{x}_2, \cdots, \mathbf{x}_n$. Each sentence representation $\mathbf{x}_i$ contains information about whether entity pair $(head, tail)$ contains relation r for input sentence x_i.

The set vector $\mathbf{s}$ is, then, computed as a weighted sum of these sentence vector $\mathbf{x}_i$:

$$\mathbf{s} = \sum_i \alpha_i \mathbf{x}_i, \tag{5}$$

where α_i is the weight of each sentence vector $\mathbf{x}_i$. In this paper, we define α_i in two ways:

Average: We assume that all sentences in the set X have the same contribution to the representation of the set. It means the embedding of the set S is the average of all the sentence vectors:

$$\mathbf{s} = \sum_i \frac{1}{n} \mathbf{x}_i, \tag{6}$$

It's a naive baseline of our selective attention.

Selective Attention: However, the wrong labelling problem inevitably occurs. Thus, if we regard each sentence equally, the wrong labelling sentences will bring in massive of noise during training and testing. Hence, we use a selective attention to de-emphasize the noisy sentence. Hence, α_i is further defined as:

$$\alpha_i = \frac{\exp(e_i)}{\sum_k \exp(e_k)}, \tag{7}$$

where e_i is referred as a query-based function which scores how well the input sentence x_i and the predict relation r matches. We select the bilinear form which achieves best performance in different alternatives:

$$e_i = \mathbf{x}_i \mathbf{A} \mathbf{r}, \tag{8}$$

where $\mathbf{A}$ is a weighted diagonal matrix, and $\mathbf{r}$ is the query vector associated with relation r which indicates the representation of relation r.

Finally, we define the conditional probability $p(r|S, \theta)$ through a softmax layer as follows:

$$p(r|S, \theta) = \frac{\exp(o_r)}{\sum_{k=1}^{n_r} \exp(o_k)}, \tag{9}$$

where n_r is the total number of relations and $\mathbf{o}$ is the final output of the neural network which corresponds to the scores associated to all relation types, which is defined as follows:

$$\mathbf{o} = \mathbf{M}\mathbf{s} + \mathbf{d}, \tag{10}$$

where $\mathbf{d} \in \mathbb{R}^{n_r}$ is a bias vector and $\mathbf{M}$ is the representation matrix of relations.

(Zeng et al., 2015) follows the assumption that at least one mention of the entity pair will reflect their relation, and only uses the sentence with the highest probability in each set for training. Hence, the method which they adopted for multi-instance learning can be regarded as a special case as our selective attention when the weight of the sentence with the highest probability is set to 1 and others to 0.

3.3 Optimization and Implementation Details

Here we introduce the learning and optimization details of our model. We define the objective function using cross-entropy at the set level as follows:

$$J(\theta) = \sum_{i=1}^{s} \log p(r_i|S_i, \theta), \tag{11}$$

where s indicates the number of sentence sets and θ indicates all parameters of our model. To solve the optimization problem, we adopt stochastic gradient descent (SGD) to minimize the objective function. For learning, we iterate by randomly selecting a mini-batch from the training set until converge.

In the implementation, we employ dropout (Srivastava et al., 2014) on the output layer to prevent overfitting. The dropout layer is defined as an element-wise multiplication with a a vector $\mathbf{h}$ of Bernoulli random variables with probability p. Then equation (10) is rewritten as:

$$\mathbf{o} = \mathbf{M}(\mathbf{s} \circ \mathbf{h}) + \mathbf{d}. \tag{12}$$

In the test phase, the learnt set representations are scaled by p, i.e., $\hat{\mathbf{s}}_i = p\mathbf{s}_i$. And the scaled set vector $\hat{\mathbf{r}}_i$ is finally used to predict relations.

4 Experiments

Our experiments are intended to demonstrate that our neural models with sentence-level selective attention can alleviate the wrong labelling problem and take full advantage of informative sentences for distant supervised relation extraction. To this end, we first introduce the dataset and evaluation metrics used in the experiments. Next, we use cross-validation to determine the parameters of our model. And then we evaluate the effects of our selective attention and show its performance on the data with different set size. Finally, we compare the performance of our method to several state-of-the-art feature-based methods.

4.1 Dataset and Evaluation Metrics

We evaluate our model on a widely used dataset[1] which is developed by (Riedel et al., 2010) and has also been used by (Hoffmann et al., 2011; Surdeanu et al., 2012). This dataset was generated by aligning Freebase relations with the New York Times corpus (NYT). Entity mentions are found using the Stanford named entity tagger (Finkel et al., 2005), and are further matched to the names of Freebase entities. The Freebase relations are divided into two parts, one for training and one for testing. It aligns the the sentences from the corpus of the years 2005-2006 and regards them as training instances. And the testing instances are the aligned sentences from 2007. There are 53 possible relationships including a special relation NA which indicates there is no relation between head and tail entities. The training data contains 522,611 sentences, 281,270 entity pairs and 18,252 relational facts. The testing set contains 172,448 sentences, 96,678 entity pairs and 1,950 relational facts.

Similar to previous work (Mintz et al., 2009), we evaluate our model in the held-out evaluation. It evaluates our model by comparing the relation

[1] `http://iesl.cs.umass.edu/riedel/ecml/`

facts discovered from the test articles with those in Freebase. It assumes that the testing systems have similar performances in relation facts inside and outside Freebase. Hence, the held-out evaluation provides an approximate measure of precision without time consumed human evaluation. We report both the aggregate curves precision/recall curves and Precision@N (P@N) in our experiments.

4.2 Experimental Settings

4.2.1 Word Embeddings

In this paper, we use the word2vec tool [2] to train the word embeddings on NYT corpus. We keep the words which appear more than 100 times in the corpus as vocabulary. Besides, we concatenate the words of an entity when it has multiple words.

4.2.2 Parameter Settings

Following previous work, we tune our models using three-fold validation on the training set. We use a grid search to determine the optimal parameters and select learning rate λ for SGD among $\{0.1, 0.01, 0.001, 0.0001\}$, the sliding window size $l \in \{1, 2, 3, \cdots, 8\}$, the sentence embedding size $n \in \{50, 60, \cdots, 300\}$, and the batch size B among $\{40, 160, 640, 1280\}$. For other parameters, since they have little effect on the results, we follow the settings used in (Zeng et al., 2014). For training, we set the iteration number over all the training data as 25. In Table 1 we show all parameters used in the experiments.

Table 1: Parameter settings

Window size l	3
Sentence embedding size d^c	230
Word dimension d^a	50
Position dimension d^b	5
Batch size B	160
Learning rate λ	0.01
Dropout probability p	0.5

4.3 Effect of Sentence-level Selective Attention

To demonstrate the effects of the sentence-level selective attention, we empirically compare different methods through held-out evaluation. We select the CNN model proposed in (Zeng et al., 2014) and the PCNN model proposed in (Zeng et al., 2015) as our sentence encoders and implement them by ourselves which achieve comparable results as the authors reported. And we compare the performance of the two different kinds of CNN with sentence-level attention (ATT) , its naive version (AVE) which represents each sentence set as the average vector of sentences inside the set and the at-least-one multi-instance learning (ONE) used in (Zeng et al., 2015).

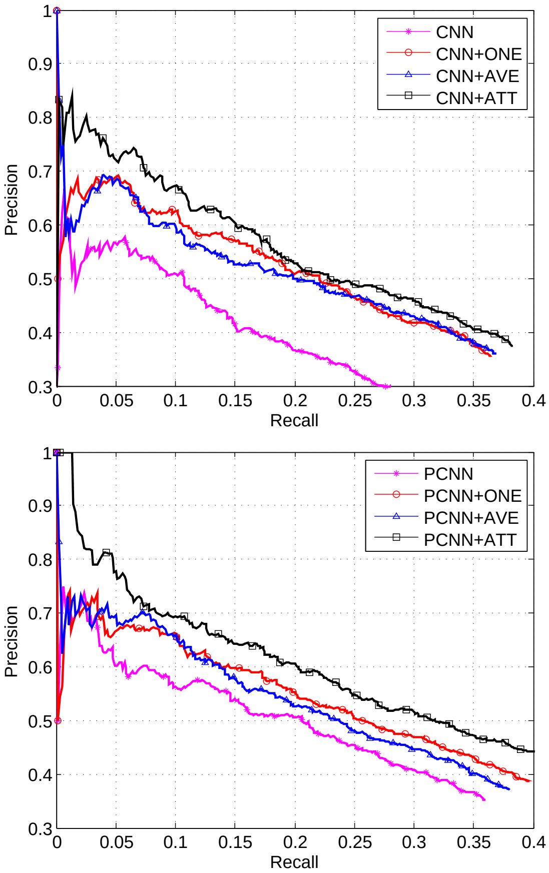

Figure 3: Top: Aggregate precion/recall curves of CNN, CNN+ONE, CNN+AVE, CNN+ATT. Bottom: Aggregate precion/recall curves of PCNN, PCNN+ONE, PCNN+AVE, PCNN+ATT

From Fig. 3, we have the following observation: (1) For both CNN and PCNN, the ONE method brings better performance as compared to CNN/PCNN. The reason is that the original distant supervision training data contains a lot of noise and the noisy data will damage the performance of relation extraction. (2) For both CNN and PCNN, the AVE method is useful for relation extraction as compared to CNN/PCNN. It indicates that considering more sentences is beneficial to relation extraction since the noise can be reduced by mutual complementation of information. (3) For both

[2] `https://code.google.com/p/word2vec/`

CNN and PCNN, the AVE method has a similar performance compared to the ONE method. It indicates that, although the AVE method brings in information of more sentences, since it regards each sentence equally, it also brings in the noise from the wrong labelling sentences which may hurt the performance of relation extraction. (4) For both CNN and PCNN, the ATT method achieves the highest precision over the entire range of recall compared to other methods including the AVE method. It indicates that the proposed selective attention is beneficial. It can effectively filter out meaningless sentences and alleviate the wrong labelling problem in distant supervised relation extraction.

4.4 Effect of Sentence Number

In the original testing data set, there are 74,857 entity pairs that correspond to only one sentence, nearly 3/4 over all entity pairs. Since the superiority of our selective attention lies in the entity pairs containing multiple sentences, we compare the performance of CNN/PCNN+ONE, CNN/PCNN+AVE and CNN/PCNN+ATT on the entity pairs which have more than one sentence. And then we examine these three methods in three test settings:

- **One**: For each testing entity pair, we randomly select one sentence and use this sentence to predict relation.
- **Two**: For each testing entity pair, we randomly select two sentences and proceed relation extraction.
- **All**: We use all sentences of each entity pair for relation extraction.

Note that, we use all the sentences in training. We will report the P@100, P@200, P@300 and the mean of them for each model in held-out evaluation.

Table 2 shows the P@N for compared models in three test settings. From the table, we can see that: (1) For both CNN and PCNN, the ATT method achieves the best performance in all test settings. It demonstrates the effectiveness of sentence-level selective attention for multi-instance learning. (2) For both CNN and PCNN, the AVE method is comparable to the ATT method in the One test setting. However, when the number of testing sentences per entity pair grows, the performance of the AVE methods has almost no improvement. It even drops gradually in P@100, P@200 as the sentence number increases. The reason is that, since we regard each sentence equally, the noise contained in the sentences that do not express any relation will have negative influence in the performance of relation extraction. (3) CNN+AVE and CNN+ATT have 5% to 8% improvements compared to CNN+ONE in the ONE test setting. Since each entity pair has only one sentence in this test setting, the only difference of these methods is from training. Hence, it shows that utilizing all sentences will bring in more information although it may also bring in some extra noises. (4) For both CNN and PCNN, the ATT method outperforms other two baselines over 5% and 9% in the Two and All test settings. It indicates that by taking more useful information into account, the relational facts which CNN+ATT ranks higher are more reliable and beneficial to relation extraction.

4.5 Comparison with Feature-based Approaches

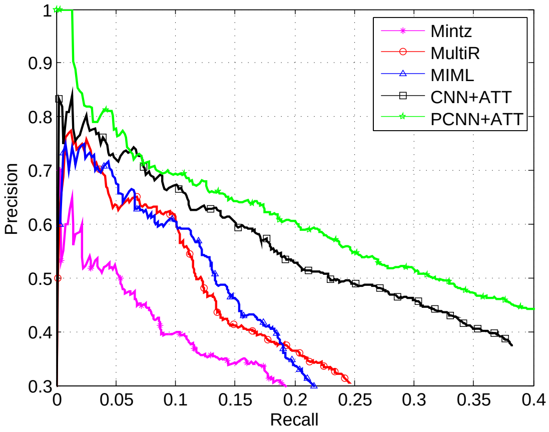

Figure 4: Performance comparison of proposed model and traditional methods

To evaluate the proposed method, we select the following three feature-based methods for comparison through held-out evaluation:

Mintz (Mintz et al., 2009) is a traditional distant supervised model.

MultiR (Hoffmann et al., 2011) proposes a probabilistic, graphical model of multi-instance learning which handles overlapping relations.

MIML (Surdeanu et al., 2012) jointly models both multiple instances and multiple relations.

We implement them with the source codes released by the authors.

Table 2: P@N for relation extraction in the entity pairs with different number of sentences

Test Settings	One				Two				All			
P@N(%)	100	200	300	Mean	100	200	300	Mean	100	200	300	Mean
CNN+ONE	68.3	60.7	53.8	60.9	70.3	62.7	55.8	62.9	67.3	64.7	58.1	63.4
+AVE	75.2	67.2	58.8	67.1	68.3	63.2	60.5	64.0	64.4	60.2	60.1	60.4
+ATT	**76.2**	65.2	**60.8**	67.4	76.2	65.7	62.1	68.0	**76.2**	68.6	59.8	68.2
PCNN+ONE	73.3	64.8	56.8	65.0	70.3	67.2	63.1	66.9	72.3	69.7	64.1	68.7
+AVE	71.3	63.7	57.8	64.3	73.3	65.2	62.1	66.9	73.3	66.7	62.8	67.6
+ATT	73.3	**69.2**	**60.8**	**67.8**	**77.2**	**71.6**	**66.1**	**71.6**	**76.2**	**73.1**	**67.4**	**72.2**

Fig. 4 shows the precision/recall curves for each method. We can observe that: (1) CNN/PCNN+ATT significantly outperforms all feature-based methods over the entire range of recall. When the recall is greater than 0.1, the performance of feature-based method drop out quickly. In contrast, our model has a reasonable precision until the recall approximately reaches 0.3. It demonstrates that the human-designed feature cannot concisely express the semantic meaning of the sentences, and the inevitable error brought by NLP tools will hurt the performance of relation extraction. In contrast, CNN/PCNN+ATT which learns the representation of each sentences automatically can express each sentence well. (2) PCNN+ATT performs much better as compared to CNN+ATT over the entire range of recall. It means that the selective attention considers the global information of all sentences except the information inside each sentence. Hence, the performance of our model can be further improved if we have a better sentence encoder.

4.6 Case Study

Table 3 shows two examples of selective attention from the testing data. For each relation, we show the corresponding sentences with highest and lowest attention weight respectively. And we highlight the entity pairs with bold formatting.

From the table we find that: The former example is related to the relation `employer_of`. The sentence with low attention weight does not express the relation between two entities, while the high one shows that *Mel Karmazin* is the chief executive of *Sirius Satellite Radio*. The later example is related to the relation `place_of_birth`. The sentence with low attention weight expresses where *Ernst Haefliger* is died in, while the high one expresses where he is born in.

Table 3: Some examples of selective attention in NYT corpus

Relation	employer_of
Low	When Howard Stern was preparing to take his talk show to **Sirius Satellite Radio**, following his former boss, **Mel Karmazin**, Mr. Hollander argued that ...
High	**Mel Karmazin**, the chief executive of **Sirius Satellite Radio**, made a lot of phone calls ...
Relation	place_of_birth
Low	**Ernst Haefliger**, a Swiss tenor who ... roles , died on Saturday in **Davos**, Switzerland, where he maintained a second home.
High	**Ernst Haefliger** was born in **Davos** on July 6, 1919, and studied at the Wettinger Seminary ...

5 Conclusion and Future Works

In this paper, we develop CNN with sentence-level selective attention. Our model can make full use of all informative sentences and alleviate the wrong labelling problem for distant supervised relation extraction. In experiments, we evaluate our model on relation extraction task. The experimental results show that our model significantly and consistently outperforms state-of-the-art feature-based methods and neural network methods.

In the future, we will explore the following directions:

- Our model incorporates multi-instance learning with neural network via instance-level selective attention. It can be used in not only distant supervised relation extraction but also other multi-instance learning tasks. We will explore our model in other area such as text categorization.

- CNN is one of the effective neural networks for neural relation extraction. Researchers also propose many other neural network models for relation extraction. In the future, we will incorporate our instance-level selective attention technique with those models for relation extraction.

Acknowledgments

This work is supported by the 973 Program (No. 2014CB340501), the National Natural Science Foundation of China (NSFC No. 61572273, 61303075) and the Tsinghua University Initiative Scientific Research Program (20151080406).

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. *Dbpedia: A nucleus for a web of open data*. Springer.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Yoshua Bengio. 2009. Learning deep architectures for ai. *Foundations and trends® in Machine Learning*, 2(1):1–127.

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 KDD*, pages 1247–1250.

Razvan Bunescu and Raymond Mooney. 2007. Learning to extract relations from the web using minimal supervision. In *Proceedings of ACL*, volume 45, page 576.

Jan Chorowski, Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. End-to-end continuous speech recognition using attention-based recurrent nn: first results. *arXiv preprint arXiv:1412.1602*.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *JMLR*, 12:2493–2537.

Thomas G Dietterich, Richard H Lathrop, and Tomás Lozano-Pérez. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artificial intelligence*, 89(1):31–71.

Cıcero Nogueira dos Santos and Maıra Gatti. 2014. Deep convolutional neural networks for sentiment analysis of short texts. In *Proceedings of COLING*.

Cıcero Nogueira dos Santos, Bing Xiang, and Bowen Zhou. 2015. Classifying relations by ranking with convolutional neural networks. In *Proceedings of ACL*, volume 1, pages 626–634.

Jenny Rose Finkel, Trond Grenager, and Christopher Manning. 2005. Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of ACL*, pages 363–370. Association for Computational Linguistics.

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 ACL-HLT*, pages 541–550.

Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of ACL-IJCNLP*, pages 1003–1011.

Volodymyr Mnih, Nicolas Heess, Alex Graves, et al. 2014. Recurrent models of visual attention. In *Proceedings of NIPS*, pages 2204–2212.

Sebastian Riedel, Limin Yao, and Andrew McCallum. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of ECML-PKDD*, pages 148–163.

Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proceedings of EMNLP-CoNLL*, pages 1201–1211.

Richard Socher, John Bauer, Christopher D Manning, and Andrew Y Ng. 2013. Parsing with compositional vector grammars. In *Proceedings of ACL*. Citeseer.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *JMLR*, 15(1):1929–1958.

Fabian M Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of WWW*, pages 697–706. ACM.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of EMNLP*, pages 455–465.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of NIPS*, pages 3104–3112.

Ruobing Xie, Zhiyuan Liu, Jia Jia, Huanbo Luan, and Maosong Sun. 2016. Representation learning of knowledge graphs with entity descriptions.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Aaron Courville, Ruslan Salakhutdinov, Richard Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. *Proceedings of ICML.*

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of 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 EMNLP*.