Distant Supervision for Relation Extraction with Linear Attenuation Simulation and Non-IID Relevance Embedding

Changsen Yuan,[1] Heyan Huang,[1,2*] Chong Feng,[1] Xiao Liu,[1] Xiaochi Wei[3]
[1]Department of Computer Science, Beijing Institute of Technology, China
[2]Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, China
[3]Baidu Inc. , China
{yuanchangsen, hhy63, fengchong, xiaoliu}@bit.edu.cn weixiaochi@baidu.com

Abstract

Distant supervision for relation extraction is an efficient method to reduce labor costs and has been widely used to seek novel relational facts in large corpora, which can be identified as a multi-instance multi-label problem. However, existing distant supervision methods suffer from selecting important words in the sentence and extracting valid sentences in the bag. Towards this end, we propose a novel approach to address these problems in this paper. Firstly, we propose a linear attenuation simulation to reflect the importance of words in the sentence with respect to the distances between entities and words. Secondly, we propose a non-independent and identically distributed (non-IID) relevance embedding to capture the relevance of sentences in the bag. Our method can not only capture complex information of words about hidden relations, but also express the mutual information of instances in the bag. Extensive experiments on a benchmark dataset have well-validated the effectiveness of the proposed method.

Introduction

Relation extraction, aiming to categorize semantic relations between entity pairs in plain texts, has been widely adopted in many natural language processing (NLP) tasks, such as question answering (Sadeghi, Divvala, and Farhadi 2015), text categorization (Huynh et al. 2011) and web search (Yan et al. 2009). Traditional supervised methods for relation extraction require a large amount of high-quality corpus for model training, which is extremely expensive and time-consuming. Additionally, these datasets are often restricted to certain domains. In recent years, distant supervision for relation extraction has been proposed to find abundant relational facts with large amount auto-generated labels. However, it has two major flaws in existing distant supervision methods.

Firstly, the existing approaches acquiescently assume that each word in the sentence has the same weight in relation extraction. This hypothesis is too strong and usually leads to wrong labels. The relationship between entities and words gradually decreases with the creasing of the distant between them. Therefore, words can not maintain the same weight in distant supervision. For example, (*South Korea*, *Seoul*, *Country*) is a relational fact in KB. Each word in the sentence "*many foreign investors say the investigation is emblematic of the political uncertainty they face in investing in South Korea, a concern that looms large as Washington and Seoul are negotiating a free trade agreement.*" is not always useful for "*Country*". Some invalid words exist in the long sentence. Moreover, (Mcdonald and Nivre 2007) showed that the accuracy of syntactic parsing decreases significantly with increasing sentence length. In the bag, we find that some instances are too long and contain some invalid words about target relation. And these invalid words are usually far away from entities. Long distance between entity and word indicates a weak correlation between them. Conversely, short distance between entity and word possesses strong correlation. These phenomena sometimes lead to wrong labels in distant supervision. Therefore, if we use the same weights about words in relation extraction, weights of words will not only affect the expression of sentences, but also have an important impact on the judgement of labels.

Secondly, distant supervision for relation extraction possesses an ideal hypothesis that all instances containing the same entity pairs express the same relation. However, this is far from reality, because there may exist multi relations between a specific entity pairs. For example, both the relation "*Born_in*" and "*Employ_by*" are valid between the entity pair "*Trump*" and "*the USA*". To solve this problem, the multi-instance learning (Hoffmann et al. 2011; Surdeanu et al. 2012) and sentence-level attention (Lin et al. 2016; Ji et al. 2017) have been proposed, but the above approaches also have flaws. In relation extraction, the multi-instance learning only selects the instance with the highest probability to be a valid candidate, so that a large amount of rich information is lost. And the sentence-level attention considers instances in the bag as independent and identically distributed (IID), therefore, the relevance of instances is ignored consequently. In contrast, these instances with the same entity pairs in the bag have more or less connections, which are important information of sentences. Toward this end, we assume that the relevance of sentences is able to selectively assign higher weights for valid sentences and lower weights for invalid sentences. For example, in Figure 1, sentence S1 expresses the relation "*Employ_by*" and sentence S2 expresses the relation "*Born_in*". But we can implicitly obtain the relation of "*Born_in*" between "*Trumph*" and "*the*

*Corresponding author

ID	Freebase-Relation	Instance	Relevance	Weight
S0	R = Born_in	Trump has been active in the USA to win the 2012 presidential to challenge Barack Obama.	0.61	0.26
S1		Trump is the 45th president of the USA.	0.59	0.25
S2		**Trump (born in the USA) is an Amercian businessman, reality television personality, politician.**	**1**	**0.42**
S3		On January 19, 2017, in Washington, DC, the USA President Trump and his wife Melania appeared at the Lincoln Memorial to attend the concert "Let America be great again."	0.17	0.07

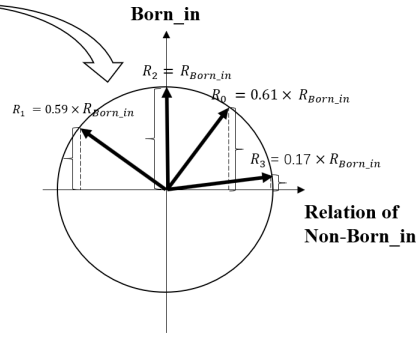


Figure 1: An example of non-IID relevance embedding in a bag. There are 4 sentences and 3-rd sentence is the best sentence to express relation of "*Born_in*". The right of Figure 1 represent the relevance of sentences. The vertical axis represents relation of "*Born_in*" with coordinating value from -1 to 1, while other directions represent other relations.

USA" from the S1. This phenomenon illustrates that there is the connection between two sentences. Therefore, non-independent and identically distributed (non-IID) are proposed to solve the relevance of instances and enhance valid sentences.

In this paper, we propose linear attenuation simulation and non-IID relevance embedding to increase valid instances and enhance the results of relation extraction. To address the first problem, we assume that the connection of entity and word changes with the distance between entity and word. This variation is linear attenuation. Linear attenuation simulation can reduce the weight of word with the increase of distance between entity and word. Thus, we use linear attenuation simulation to work out this problem.

To solve the next problem, we adopt non-IID relevance embedding to learn the relevance of instances. Non-IID relevance embedding builds non-IID representations via modeling each bag along with its corresponding neighbors. Concretely, we use the cosine similarity between two sentences (S_1, S_2) to represent the relevance of S_2 about S_1, where S_1 is the best sentence to express the relation. If the sentence has lower similarity with the best sentence which can perfectly perform relation in the bag, this sentence will be assigned to a low weight. Therefore, the non-IID relevance embedding can improve the weights of valid sentences and enhance the correct labels for relation extraction. The experimental results show that our method achieves significant and consistent improvements in relation extraction as compared with the state-of-the art methods.

The main contributions of this paper are summarized as follows:

- We propose a linear attenuation simulation to select useful words and alleviate the wrong labels which are caused by long distance between entity and word.
- To address the relevance of sentences, we develop innovative solutions that introduce non-IID relevance embedding to distant supervised relation extraction.
- In the experiments, results show that our model achieves better performance in distant supervised relation extraction.

Methodology

We propose a new model for relation extraction containing linear attenuation simulation and non-IID relevance embedding. Linear attenuation simulation not only can provide and remain important words, but also improve the representation of sentences in our model. Non-IID relevance embedding provides more information between each sentence in the bag, which is able to select valid instances and bring more relevant information. The overall structure of our proposed model is illustrated in Figure 2, our model consists of two main components: PCNNs Module and Attention Module. The PCNNs Module is used to extract features and compute the weights of words from a sentence in a bag. And the PCNNs Module is further comprised of *Vector Representation*, *Linear Attenuation Simulation*, and *Piecewise Convolution Neural Networks (PCNNs)*. The function of *Vector Representation* is to transform words into low-dimensional vectors. The function of *Linear Attenuation Simulation* is to assign weights to words. *PCNNs* is used to extract feature vector of the sentence. The Attention Module is used to compute the weights of all sentences in a bag, and feed the bag features into a softmax classifier. And the Attention Module is comprised of *Non-IID Relevance Embedding* and *Classifying*. We elaborate on these parts in following paragraphs.

Vector Representation

When using relation extraction, we require to translate each word to a low-dimensional vector. In this paper, we translate words into vectors by looking up the pre-trained word embeddings. In addition, position features (PFs) are used to specify entity pairs, which are also transformed into vectors by looking up the position embeddings.

Word Embeddings: Word embeddings are language modeling and feature learning techniques in NLP that map each word or phrase to a real-valued vector. They represent words between semantic and syntactic information. Given a sentence $\mathbf{X} = \{\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_k}\}$, where each word $\mathbf{w_i}$ is represented by a real-valued vector. Word representations are encoded by vectors in an embedding matrix. In this paper, we use the Skip-gram model (Mikolov et al. 2013) to train the word embeddings.

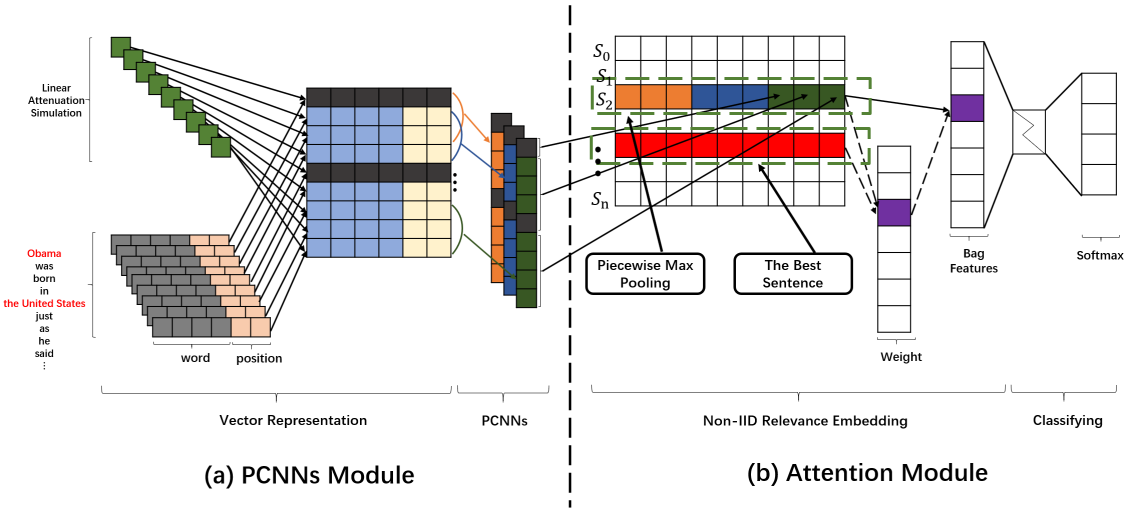


Figure 2: The architecture of model. The red segment is the best sentence which can express the relation of r.

Position Embeddings: In distant supervised relation extraction, we focus on assigning labels to entity pairs. Similar to (Zeng et al. 2014), we use position embeddings (PFs) to specify entity pairs. PFs are regarded as the combination of the relative distances from the current word to head entity and tail entity. For example, in the sentence "*Obama was born in the United States just as he has always said.*", the relative distances from "*he*" to head entity (*Obama*) and tail entity (*the United States*) are 7 and 3. Relative distances from "*in*" to head entity (*Obama*) and tail entity (*the United States*) are 4 and -1, respectively.

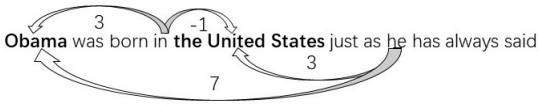


Figure 3: Position Embeddings

The position embedding matrices about entities are randomly initialized. Similar to the word embeddings, we transform the relative distances into real-valued vectors through looking up the position embedding matrices.

We assume that the size of word embedding is $d_w = 5$ and that the size of position embedding is $d_p = 1$. Finally, we combine the word embeddings and position embeddings of all words and transform it as a vector sequence $\mathbf{X} = \{\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_k}\}$, where k is the sentence length and $w_i \in R^d (d = d_w + d_p * 2)$.

Linear Attenuation Simulation

In relation extraction, words which close to the target entities often contain more information about relations. On the contrary, when some words have long relative distances, these words are regarded as less or useless information about relations.

Suppose there is a sentence $\mathbf{X}$ consisting of k words ($\mathbf{X} = \{\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_k}\}$), containing a head entity and a tail entity. To exploit the information of all words, our model represents the sentence $\mathbf{X}$ with a real-valued matrix when predicting relation r. It is straightforward that the sentence is made up all words, $\{\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_k}\}$. Each word contains different information which could decide relation of entity pairs. Then, the vector $\mathbf{X}$ is calculated as:

$$\mathbf{X} = \{\gamma_1 \mathbf{w_1}, \gamma_2 \mathbf{w_2}, \ldots, \gamma_i \mathbf{w_i}, \ldots, \gamma_k \mathbf{w_k}\} \quad (1)$$

where γ_i is the weights of each word. In general, we define γ_i in two ways.

Constant=1: Normally, we think that each word in the sentence has the same weight to express the information of relation. We hence set $\gamma_i = 1$. Then, the sentence vector $\mathbf{X}$:

$$\mathbf{X} = \{\mathbf{w_1}, \mathbf{w_2}, \ldots, \mathbf{w_k}\} \quad (2)$$

Constant≠1: However, with the increase of sentence length, the weight continues to decrease about the relation. Therefore, if we regard each word as the same weight, the unimportant and the low-weight words will be equally computed with the high-weight words during the training and testing.

Figure 4: Linear Attenuation Simulation

So, we use linear attenuation simulation to reduce the impact of words with low weight. Hence, γ_i is calculated as:

$$\gamma_i = \begin{cases} (1 - \frac{|d_{i1}|}{D}) + (1 - \frac{|d_{i2}|}{D}) & if\ d_{ij} \leq D \\ 0 & if\ d_{ij} > D \end{cases} \quad (3)$$

where d_{i1} is referred as the relative distance about head entity. d_{i2} is referred as the relative distance about tail entity. j is the number which is 1 or 2. D is referred as the threshold.

If the distance of some words about entities is greater than D, their weights will be regarded as 0. Weights of "*in*" about "*Obama*" and "*the United States*" are $1-\frac{3}{D}$ and $1-\frac{1}{D}$. Thus, the weight of "*in*" is $2-\frac{4}{D}$. Finally, we use the new $\mathbf{X}$ to accomplish the task of distant supervision.

PCNNs

In relation extraction, this model is employed to extract feature vectors of an instance.

CNN: Convolution neural networks is a typical neural networks. Convolution is an operation between the weight matrix $\mathbf{A}$, and the input matrix $\mathbf{B}$. $\mathbf{A}$ is regarded as the filter for the convolution. For example, we assume that $\mathbf{A} = (a_{ij})_{m\times n}$ and $\mathbf{B} = (b_{ij})_{m\times n}$, then $\mathbf{C} = \mathbf{A} \otimes \mathbf{B} = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}b_{ij}$ is defined as convolution, where $\mathbf{C}$ is convolution, m is the length of filter ($m = 3$) and $n = d_w + d_p * 2$. We consider $\mathbf{S}$ to be a sequence $\{\mathbf{q_1}, \mathbf{q_2}, \cdots, \mathbf{q_s}\}$. Normally, let $\mathbf{Q_{i:j}}$ refer to the concatenation of $\mathbf{q_i}$ to $\mathbf{q_j}$. Thus, the convolution operation between the matrix of sentence, $\mathbf{Q}$, and the matrix of weight, $\mathbf{W}$, results in another vector.

$$\mathbf{c_j} = \mathbf{A} \otimes \mathbf{Q_{i:j}} \tag{4}$$

where $j = i + m - 1$.

Piecewise Max-pooling: PCNNs (Zeng et al. 2015), a variation of CNN, adopts piecewise max-pooling in relation extraction to extract features. This method can obtain the structural information. Each convolution, $\mathbf{c_j}$, is divided into three parts $\mathbf{c_j} = \{\mathbf{c_{j1}}, \mathbf{c_{j2}}, \mathbf{c_{j3}}\}$ by head entity and tail entity. Then, the max-pooling procedure is performed in three parts separately. Next, we can concatenate all vectors $\mathbf{p_j} = [p_{j1}, p_{j2}, p_{j3}]$, which $p_{jh} = max\,(\mathbf{c_{jh}})\,(h = 1, 2, 3)$. Finally, we compute the feature vectors by a non-linear function at the output.

Non-IID Relevance Embedding

Given a bag $\mathbf{B} = \{\mathbf{s_0}, \mathbf{s_1}, \cdots, \mathbf{s_n}\}$, if we assume the predefined semantic relation is r, we can select the best sentence, $\mathbf{s_i}$, which can better perform the r than the rest of sentences in the bag via multi-instance learning (MIL). And we consider that sentences in the bag can express r and are non-IID. Traditionally, these sentences are often viewed as independent, which inevitably leads to loss of information for distant supervision. To incorporate the non-IID, we compute similarity of remaining sentences with $\mathbf{s_i}$. There is a sentence, $\mathbf{s_j}$, in the bag. If $\mathbf{s_j}$ has a high similarity with $\mathbf{s_i}$, $\mathbf{s_j}$ could have a high weight in the bag. Higher similarity is, higher weight is in the bag. As shown in Figure 5, the bag has 4 sentences, and $\mathbf{s_1}$ is the best performance of r by MIL. The weight of $\mathbf{s_2}$ about r can be computed by $\alpha_{1,2}$. In other words, $\alpha_{1,2}$ can select the weight of $\mathbf{s_2}$ about the relation of r. Hence, the weights of sentences about r is calculated as:

$$\alpha_{i,j} = \frac{e_{i,j}}{\sum_k e_{i,k}} \tag{5}$$

where $\alpha_{i,j}$ is the weight of each sentence and $e_{i,j}$ is the similarity of sentence about the r. $e_{i,j}$ is calculated as:

$$e_{i,j} = \frac{\mathbf{s_i} \cdot \mathbf{s_j}}{||\mathbf{s_i}|| \times ||\mathbf{s_j}||} \tag{6}$$

where $\mathbf{s_i}$ is the best sentence of r, and $\mathbf{s_j}$ is sentence in the bag. The set vector $\mathbf{B}$ is calculated as a weighted sum of these sentence vectors:

$$\mathbf{B} = \sum_j \alpha_{i,j}\mathbf{s_j} \tag{7}$$

	s_0	s_1	s_2	s_3
s_0	$\alpha_{0,0}$	$\alpha_{0,1}$	$\alpha_{0,2}$	$\alpha_{0,3}$
s_1	**$\alpha_{1,0}$**	**$\alpha_{1,1}$**	**$\alpha_{1,2}$**	**$\alpha_{1,3}$**
s_2	$\alpha_{2,0}$	$\alpha_{2,1}$	$\alpha_{2,2}$	$\alpha_{2,3}$
s_3	$\alpha_{0,3}$	$\alpha_{1,3}$	$\alpha_{2,3}$	$\alpha_{3,3}$

Figure 5: Non-IID Relevance Embedding. The part of bold font is the weights of sentences in a bag.

Classifying

In this section, we use softmax to get the conditional probability, as:

$$p(r|B, \theta) = \frac{exp(o_r)}{\sum_{k=1} exp(o_k)} \tag{8}$$

where r is the representation of relation r, and o is the final output, which is defined as:

$$\mathbf{o} = \mathbf{MB} + \mathbf{D} \tag{9}$$

where M is the matrix of relations and D is a bias vector. We define the objection function using cross-entropy(Shore and Johnson 1980) as:

$$J(\theta) = \sum_{i=1}^{n} \log\left(p\left(r_i|B_i, \theta\right)\right) \tag{10}$$

where n is the number of sentences and θ indicates all parameters of our model. In this paper, we combine dropout to prevent overfitting.

Experiments

Our experiments are intended to show that our model can capture high weight words and take full advantage of informative sentences for distant supervised relation extraction. In the experiments, we first introduce the dataset and evaluation metrics used. Next, we determine some parameters of our model by cross-validation. Finally, we evaluate the effects of linear attenuation simulation and non-IID relevance embedding, and we also compare our method to some classical methods.

Dataset and Evaluation Metrics

We evaluate our model on the New York Times (NYT)[1] corpus which is developed by (Riedel, Yao, and McCallum 2010) and has also been used by (Hoffmann et al. 2011; Surdeanu et al. 2012; Lin et al. 2016). This dataset was generated by aligning Freebase relations. The sentences from 2005 to 2006 are used for training, and the sentences from 2007 are used for testing.

Following the previous work (Lin et al. 2016; Ji et al. 2017), we evaluate our method in the held-out evaluation. It evaluates our model by comparing the relation facts discovered from the test articles with those in Freebase. In the experiments, we assume that the NYT has the similar data structure every year. So, the held-out evaluation provides an approximate measure of precision without consuming human evaluation. We report both the precision/recall curves and Precision@N (P@N) in our experiments.

Experimental Settings

Word Embedding: In this paper, we employ the Skip-gram model[2] (Mikolov et al. 2013) to train the word embeddings on the NYT corpus. The vector representations of words which learned by word2vec models have been shown to carry semantic meanings and are useful in NLP tasks.

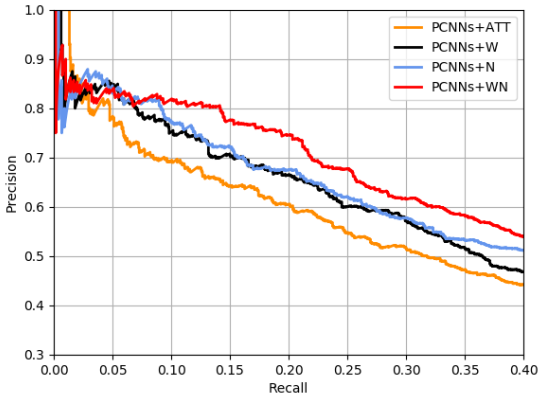
Parameters Setting: In this section, we study the influence of one parameter on our model: the threshold value D is defined in Equation (3). We tune our models using three-fold validation on the training set. We use a grid search to determine the optional parameter: $D \in \{30, 40, 50, 60, 70, 80\}$. For other parameters, we follow the settings used in (Lin et al. 2016). For training, we set the iteration number over all the training data as 14. Table 1 shows all parameters used in the experiments.

Setting	Number
Window size	3
Feature maps	230
Word dimension	50
Position dimension	5
Batch size	160
Learning rate	0.01
Dropout probability	0.5
Threshold	60

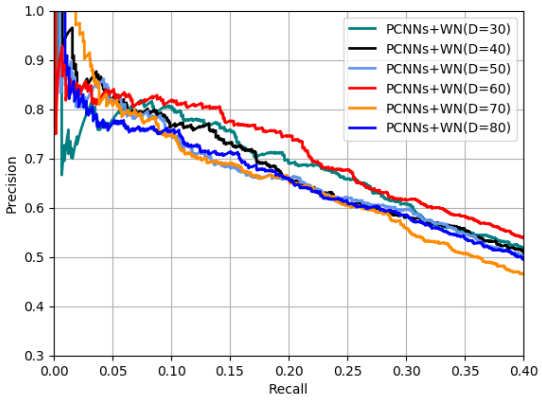
Table 1: Parameters Setting

Effect of Linear Attenuation Simulation and Non-IID Relevance Embedding

To prove the influence about linear attenuation simulation and non-IID relevance embedding, we compared with different methods by held-out evaluation. We select *PCNNs+ATT* as our baseline. *PCNNs* represents *CNN* with piecewise max-pooling, and *ATT* represents sentence-level attention. *PCNNs+ATT* has better performance than other methods in distant supervision. In order to demonstrate the validity of our method, we carried out some experiments. *PCNNs+W* represents linear attenuation simulation with *PCNNs*. *PCNNs+N* represents non-IID relevance embedding with *PCNNs*. *PCNNs+WN* represents linear attenuation simulation and non-IID relevance embedding with *PCNNs*. To determine the threshold value, D, we select the different values in the experiments, $D \in \{30, 40, 50, 60, 70, 80\}$. Experimental results are in Figure 6(b).



(a) Comparison of baseline and our approach.



(b) Comparison of different values (D).

Figure 6: Effect of Linear Attenuation Simulation and Non-IID Relevance Embedding.

Figure 6 shows that when linear attenuation simulation and non-IID relevance embedding is used in *PCNNs*, our method has achieved good results in relation extraction. Figure 6(a) shows that when linear attenuation simulation or non-IID relevance embedding is used alone in *PCNNs*, they all perform better than *PCNNs+ATT*. And *PCNNs+WN* achieves the highest precision compared to other methods. These results indicate that linear attenuation simulation can selectively assign different weights to words, and alleviate wrong labels for relation extraction. Moreover, we also notice that non-IID relevance embedding can capture the rel-

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

[2] http://code.google.com/p/word2vec/

P@N(%)	PCNNs+MIL	PCNNs+ATT	PCNNs+W	PCNNs+N	PCNNs+WN
P@100	72.3	76.2	**83.0**	81.0	**83.0**
P@200	69.7	73.1	77.0	79.5	**82.0**
P@300	64.1	67.4	72.0	76.7	**80.3**
Average	68.7	72.2	77.0	79.1	**81.8**

Table 2: P@N for relation extraction

evance of sentences, and enhance the correct labels. Figure 6(b) shows that when D is 60, our method can get the best performance. Hence, linear attenuation simulation and non-IID relevance embedding are important factors in distant supervision.

Comparison with Traditional Approaches

Held-out Evaluation: To evaluate the proposed method, we select the following seven traditional methods for comparison.

- **Mintz** (Mintz et al. 2009) proposed a traditional distant supervision model.
- **MultiR** (Hoffmann et al. 2011) proposed a probabilistic graphical model with multi-instance learning.
- **MIML** (Surdeanu et al. 2012) proposed a multi-instance and multi-label model.
- **PCNNs+MIL** (Zeng et al. 2015) proposed piecewise convolutional neural networks (PCNNs) with multi-instance learning.
- **PCNNs+ATT** (Lin et al. 2016) proposed a selective attention over instances with PCNNs and CNNs.
- **APCNNs+D** (Ji et al. 2017) proposed background information of entities by an attention layer to help relation classification.
- **SEE-TRANS** (He et al. 2018) proposed syntax-aware entity embedding with PCNNs+ATT.
- **PCNNs+WN** is our method with PCNNs.

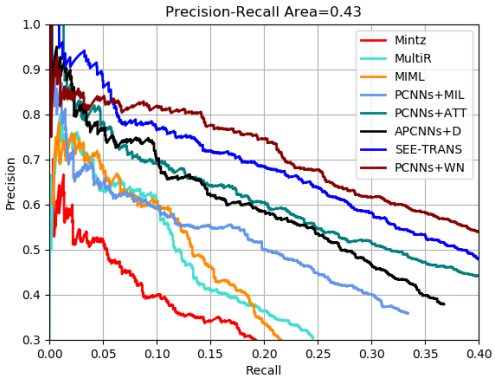


Figure 7: Performance comparison among different methods

Figure 7 shows that the precision-recall curves for each method. We can observe that: (1) *PCNNs+WN* achieves higher precision. *PCNNs+WN* enhance the mean average precision to approximately 43%. When the recall is greater than 0.07, performance of our method drops out quickly. The results demonstrate that our method is an effective way to distant supervised relation extraction and *PCNNs+WN* can alleviate the error propagation. (2) The precision of our method has declined when recall is less than 0.07. Because linear attenuation simulation reduces some words in long sentences. Maybe these words have effects on certain performance of relations. But in the experiments, our method has better performance than other methods and improves the overall effect of relation extraction. These results demonstrate that our method possesses important effects for distant supervision.

P@N Evaluation: In this section, we report the P@100, P@200, P@300 and the average of them for *PCNNs+MIL*, *PCNNs+ATT*, *PCNNs+W*, *PCNNs+W*, and *PCNNs+WN*.

Table 2 shows that: (1) *PCNNs+WN* achieves the best performance in all test settings. *PCNNs+WN* outperforms *PCNNs+ATT* over 9.6% in the average. It demonstrates the validity of linear attenuation simulation and non-IID relevance embedding for distant supervision. (2) For both *PCNNs+W* and *PCNNs+N*, the results of these methods are better than *PCNNs+ATT*. Because linear attenuation simulation can alleviate words of low weight and non-IID relevance embedding can capture valid information of each sentence about relation in a bag.

Case Study

Figure 8 shows an example of *PCNNs+WN* from the testing data. The entity-relation tuple is (*Fort-Dix*, *New-Jersey*, *contains*). There are 6 sentences containing the entity pair. The 4-th sentence, being the part of bold font, is the best sentence to express "*contains*". Our model not only can capture relation of sentences, but also can analyze correlations between 4-th sentence and each sentence in this bag. Relevance represents the correlation between 4-th sentence and each sentence in a bag. Hence, our model assigns high weights to valid sentences for our task. We argue that linear attenuation simulation and non-IID relevance embedding can enhance the performance in distant supervision. We can clearly distinguish valid sentences and invalid sentences. Therefore, linear attenuation simulation and non-IID relevance embedding can provide more information of sentences and alleviate wrong labels.

Related Work

Relation extraction is one of the most important tasks in NLP. Many methods have been proposed in relation extrac-

Tuple	Instance	Relevance	PCNNs+WN
/location/location/contains (Fort_Dix, New_Jersey)	the action yesterday by the base closure and realignment commission would combine the efforts of Fort_ Dix , mcguire air force base and naval air engineering station lakehurst under a central commander at mcg uire , the largest of the three installations in central New_Jersey .	0.57	0.13
	these three national guard soldiers , all sergeants , all deployed with the first battalion , 69th infantry , of the new york national guard , were part of a wave of roughly 700 guardsmen who landed at Fort_Dix in New_Jersey in early september for several days of reorientation before dispersing across the region .	0.69	0.16
	he received a commission from the army dental corps , as a captain , where he was stationed in orleans , france and Fort_Dix , New_Jersey .	0.62	0.15
	that kind of immediate action occurred in 1976 after four cases of swine influenza were detected at Fort_Dix , a military base in New_Jersey .	**1**	**0.23**
	New_Jersey bases , including picatinny arsenal and Fort_Dix , gained jobs through the transfer of units from elsewhere .	0.73	0.17
	the list also urged closing or realigning several small bases in New_Jersey , but adding more than 1,800 p ositions to the work force at four bases , including picatinny arsenal and Fort_Dix .	0.65	0.1

Figure 8: Some examples of PCNNs+WN in the NYT

tion, such as bootstrapping, unsupervised relation discovery and supervised classification. Supervised methods are the classical approaches to deal with the relation extraction and perform good expression (Bunescu and Mooney 2005; Zhang and Zhou 2006; Zelenko, Aone, and Richardella 2003). However, these approaches heavily depend on high quality training data.

Recently, deep learning has been widely used to automatically extract relation. It is the most representative progress in deep neural networks to cope with relation extraction, such as convolutional neural network (CNN) (Zeng et al. 2014; Santos, Xiang, and Zhou 2015), recurrent neural networks (RNN) (Cho et al. 2014; Liu et al. 2015), long short-term memory network (LSTM) (Miwa and Bansal 2016; Yan et al. 2015; Sundermeyer, Schluter, and Ney 2012) and attention-based bidirectional LSTM (Zhou et al. 2016). In general, relation extraction need mass high quality training data, which could spend much time and energy. To figure out this issue, (Mintz et al. 2009) used distant supervision to automatically produce training data via aligning KBs and texts. They assume that if two entities have a relation in KBs, all sentences which contain these two entities will express the same relation. Distant supervision is an effective method to automatically label datasets, but it often suffers from incorrect information. To alleviate this issue, some researchers regarded relation classification as a multi-instance multi-label learning problem (Riedel, Yao, and McCallum 2010; Hoffmann et al. 2011; Sundermeyer, Schluter, and Ney 2012). The term 'multi-instance learning' was proposed to predict the drug activity (Dietterich, Lathrop, and Lozano-Pérez 1997). In multi-instance learning, the uncertainty sentences can be regarded as the label of bag. Thus, the focus of multi-instance learning is to discriminate the label of bag. However, multi-instance learning is difficult to apply in neural network models. (Zeng et al. 2015) proposed at-least-one multi-instance learning and piecewise convolutional neural networks(*PCNNs+MIL*) to extract the relations in distant supervision. But *PCNNs+MIL* ignores a lot of useful information. To capture the informative sentences and reduce the influence of wrong labelled sentences, a sentence-level attention mechanism over multiple instances was proposed (Lin et al. 2016; Ji et al. 2017; Liu et al. 2017). To exploit impact between syntax information and relation extraction, (He et al. 2018) proposed to learn syntax-aware entity embedding for relation extraction. Learning from non-IID data is a recent topic (Cao 2014; Shi et al. 2017; Pang et al. 2017) to address the intrinsic data complexities, with preliminary work reported such as for clustering (Wang et al. 2011). However, the non-IID in distant supervision is seldom exploited.

Traditional methods assume that each word of the sentence is regarded as the same weight and each sentence are independent in a bag. Actually, each word could not have the same weight in the sentence and each sentence are not independent in a bag. To address these issues, we propose a novel model which can capture informative words and sentences.

Conclusion

In this paper, we exploit linear attenuation simulation and non-IID relevance embedding with piecewise convolutional neural networks (PCNNs) for distant supervised relation extraction. We apply the linear attenuation simulation to capture the words of high weights in the sentence, and then we use the non-IID relevance embedding to extract connections about surrounding sentences in the bag. We conduct experiments on a widely used benchmark dataset. The experiments show that proposed method has better performance than comparable methods. These results demonstrate that our approach can effectively deal with the task of relation extraction.

In the future, we will explore the following directions:

- Our method not only can be used in distant supervised relation extraction, but also can be used in other fields, such as event detection and question answering.
- Reinforcement learning (RL) is one of the effective methods for NLP task. In the future, we can combine our method with reinforcement learning for distant supervision.

Acknowledgments

We would like to thank Yuxiang Zhou, Rihai Su, Qian Liu and Luyang Liu for their insightful comments and suggestions. We also very appreciate the comments from anonymous reviewers which will help further improve

our work. This work is supported by National Key R&D Plan(No.2017YFB0803302), National Natural Science Foundation of China (No.61751201) and Research Foundation of Beijing Municipal Science & Technology Commission (Grant No. Z181100008918002).

References

Bunescu, R. C., and Mooney, R. J. 2005. Subsequence kernels for relation extraction. In *Proceedings of NIPS*, 171–178.

Cao, L. 2014. Non-iidness learning in behavioral and social data. *The Computer Journal* 57(9):1358–1370.

Cho, K.; Van Merrienboer, B.; Gulcehre, C.; Bahdanau, D.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *Computer Science*.

Dietterich, T. G.; Lathrop, R. H.; and Lozano-Pérez, T. 1997. Solving the multiple instance problem with axis-parallel rectangles. *Artif. Intell.* 89(1-2):31–71.

He, Z.; Chen, W.; Li, Z.; Zhang, M.; Zhang, W.; and Zhang, M. 2018. SEE: syntax-aware entity embedding for neural relation extraction. In *Proceedings of AAAI*.

Hoffmann, R.; Zhang, C.; Ling, X.; Zettlemoyer, L.; and Weld, D. S. 2011. Knowledge-based weak supervision for information extraction of overlapping relations. In *Proceedings of ACL*, 541–550.

Huynh, D.; Tran, D.; Ma, W.; and Sharma, D. 2011. A new term ranking method based on relation extraction and graph model for text classification. In *Proceedings of ACSC*, 145–152.

Ji, G.; Liu, K.; He, S.; and Zhao, J. 2017. Distant supervision for relation extraction with sentence-level attention and entity descriptions. In *Proceedings of AAAI*, 3060–3066.

Lin, Y.; Shen, S.; Liu, Z.; Luan, H.; and Sun, M. 2016. Neural relation extraction with selective attention over instances. In *Proceedings of ACL*, 2124–2133.

Liu, Y.; Wei, F.; Li, S.; Ji, H.; Zhou, M.; and Wang, H. 2015. A dependency-based neural network for relation classification. In *Proceedings of ACL*, 285–290.

Liu, T.; Wang, K.; Chang, B.; and Sui, Z. 2017. A soft-label method for noise-tolerant distantly supervised relation extraction. In *Proceedings of EMNLP*, 1790–1795.

Mcdonald, R. T., and Nivre, J. 2007. Characterizing the errors of data-driven dependency parsing models. In *Proceedings of EMNLP-CoNLL*, 122–131.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013. Efficient estimation of word representations in vector space. *CoRR* abs/1301.3781.

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

Miwa, M., and Bansal, M. 2016. End-to-end relation extraction using lstms on sequences and tree structures. In *Proceedings of ACL*, 1105–1116.

Pang, G.; Cao, L.; Chen, L.; and Liu, H. 2017. Learning homophily couplings from non-iid data for joint feature selection and noise-resilient outlier detection. In *Proceedings of IJCAI*, 2585–2591.

Riedel, S.; Yao, L.; and McCallum, A. 2010. Modeling relations and their mentions without labeled text. In *Proceedings of ECML/PKDD*, 148–163.

Sadeghi, F.; Divvala, S. K.; and Farhadi, A. 2015. Viske: Visual knowledge extraction and question answering by visual verification of relation phrases. In *Proceedings of CVPR*, 1456–1464.

Santos, C. N. D.; Xiang, B.; and Zhou, B. 2015. Classifying relations by ranking with convolutional neural networks. *Computer Science* 86(86):132–137.

Shi, Y.; Li, W.; Gao, Y.; Cao, L.; and Shen, D. 2017. Beyond IID: learning to combine non-iid metrics for vision tasks. In *Proceedings of AAAI*, 1524–1531.

Shore, J. E., and Johnson, R. W. 1980. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *Information Theory IEEE Transactions on* 26(1):26–37.

Sundermeyer, M.; Schluter, R.; and Ney, H. 2012. Lstm neural networks for language modeling. In *Proceedings of INTERSPEECH*, 601–608.

Surdeanu, M.; Tibshirani, J.; Nallapati, R.; and Manning, C. D. 2012. Multi-instance multi-label learning for relation extraction. In *Proceedings of EMNLP-CoNLL*, 455–465.

Wang, C.; Cao, L.; Wang, M.; Li, J.; Wei, W.; and Ou, Y. 2011. Coupled nominal similarity in unsupervised learning. In *Proceedings of CIKM*, 973–978.

Yan, Y.; Okazaki, N.; Matsuo, Y.; Yang, Z.; and Ishizuka, M. 2009. Unsupervised relation extraction by mining wikipedia texts using information from the web. In *Proceedings of ACL/IJCNLP*, 1021–1029.

Yan, X.; Mou, L.; Li, G.; Chen, Y.; Peng, H.; and Jin, Z. 2015. Classifying relations via long short term memory networks along shortest dependency path. *Computer Science* 42(1):56–61.

Zelenko, D.; Aone, C.; and Richardella, A. 2003. Kernel methods for relation extraction. *Journal of Machine Learning Research* 3(3):1083–1106.

Zeng, D.; Liu, K.; Lai, S.; Zhou, G.; and Zhao, J. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING*, 2335–2344.

Zeng, D.; Liu, K.; Chen, Y.; and Zhao, J. 2015. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Proceedings of EMNLP*, 1753–1762.

Zhang, M. L., and Zhou, Z. H. 2006. Adapting rbf neural networks to multi-instance learning. *Neural Proceedings Letters* 23(1):1–26.

Zhou, P.; Shi, W.; Tian, J.; Qi, Z.; Li, B.; Hao, H.; and Xu, B. 2016. Attention-based bidirectional long short-term memory networks for relation classification. In *Proceedings of ACL*, 207–212.