DISTANT SUPERVISION RELATION EXTRACTION BASED ON MUTUAL INFORMATION AND MULTI-LEVEL ATTENTION

Y. Ye[*], *S. Jiang*[†], *S. Wang*[‡], *H. Li*[§]

Abstract: Distant supervision for relation extraction, an effective method to reduce labor costs, has been widely used to search for novel relational facts from text. However, distant supervision always suffers from incorrect labelling problems. Meanwhile, existing methods for noise reduction oftentimes ignore the commonalities in the instances. To alleviate this issue, we propose a distant supervision relation extraction model based on mutual information and multi-level attention. In our proposed method, we calculate mutual information based on the attention mechanism. Mutual information are used to build attention at both word and sentence levels, which is expected to dynamically reduce the influence of noisy instances. Extensive experiments using a benchmark dataset have validated the effectiveness of our proposed method.

Key words: *distant supervision, relation extraction, mutual information, multi-level attention*

Received: December 18, 2021
Revised and accepted: June 30, 2022

DOI: 10.14311/NNW.2022.32.010

1. Introduction

The purpose of relation extraction is to obtain semantic relations of entity pairs in plain text, which is an important subtask of natural language processing (NLP). Traditional supervised methods for relation extraction require a large amount of labeled training data for specific relations, which is extremely expensive and time-consuming. In recent years, various large-scale knowledge bases (KBs) such as Freebase [3], DBpedia [1] and YAGO [22] have been established, and they are widely used in many natural language processing (NLP) tasks [27, 5]. Based on

[*]Yuxin Ye; College of Computer Science and Technology, Jilin University, Changchun, China, E-mail: yeyx@jlu.edu.cn

[†]Song Jiang; College of Software, Jilin University, Changchun, China, E-mail: jiangsong20@mails.jlu.edu.cn

[‡]Shijia Wang; College of Computer Science and Technology, Jilin University, Changchun, China, E-mail: wangsj18@mails.jlu.edu.cn

[§]Huiying Li – Corresponding author; College of Computer Science and Technology, Jilin University, Changchun, China, E-mail: lihuiying@jlu.edu.cn

this fact, Mintz et al. [17] proposes distant supervision for relation extraction to automatically label data via aligning KBs and texts. Because distant supervision can reduce the human cost of labeling data, it is widely used in various fields [4, 6]. Distant supervision make the assumption that if two entities have a relation in KBs, then all instances that contain these two entities will express this relation. However, this is an ideal hypothesis since the relation within a specific entity pairs is not unique in actual situations [9]. This leads to the incorrect labelling problem in the distant supervision for relation extraction [15]. To solve this problem, many classic strategies have been proposed, such as multiple instance method [20], probability map method [28, 23] and attention mechanism [14], and so on. Many novel methods have also been proposed in the latest research. Shen et al. [21] proposed a novel edge-reasoning hybrid graph (ER-HG) model, which leverage five types of background information instead of a specific type of information in previous works to achieve reasonable interaction between different kinds of information and alleviate the effects of noise. Moreira et al. [18] proposed a distant supervision neural relation extractor (BERT-Side), which uses additional KBs information aligned to BERT and achieves better performance for relation extraction. He et al. [8] proposed the relation extraction model DSREFC, which integrates semantic features and syntactic features into the representation and uses attention mechanism to obtain bag representation. It can significantly improve the effect of relation extraction. Xiao et al. [26] extended the application scope of distant supervision to the document level. They proposed a novel pre-trained model for document-level relation extraction, which denoises the document-level distant supervision data via multiple pre-training tasks. However, there are two major flaws in the existing distant supervision methods.

Firstly, the existing approaches assume that each instance in the package is independent with each other and processes each instance separately [24, 7, 13, 2]. This processing method ignores the connection between instances. In fact, instances with the same relation must have certain commonalities, while sentences with different relations are different. If most instances in the package are correct, we can use the similarity between an instance and the whole to estimate the correctness of the label. We take the result of aligning the triples (Biden, Born_in, US) with the text as an example. The result are shown in Tab. I. Sentence 0, sentence 4 and sentence 5 are correctly labeled sentences, and sentence 1, sentence 2 and sentence 3 are incorrectly labeled. The commonality (born in) between the correctly labeled sentences, sentence 0 and sentence 4, can be clearly observed. (In fact, in most cases, the commonalities between sentences will be more obscure and abstract. The situation is simplified for the illustration.) The incorrectly labeled sentence 1 does not have any similar semantic information with other sentences in the package. This phenomenon shows that the correct instances are often similar, but the wrong instances are usually different. Mutual information is a representation of the commonality between sentences, and it can help us to find the correct instance in the package.

Secondly, word-level methods usually assign weight to the words in the instance according to some criteria such as distance, without considering the role of words in the semantic expression of sentences [31, 12, 19, 30]. As mentioned above, mutual information reflects the commonality between sentence, so mutual information can

Number	Sentence	Align label	True label
0	...[Biden] was born in [US]...	Born_in	Born_in
1	...[Biden] has said he love [US]...	Born_in	NA
2	...[Biden] was lived in [US] last year...	Born_in	PlaceOfLived
3	...[Biden] was the president of [US]...	Born_in	President
4	...[Biden] was born in Pennsylvania([US]) in 1942...	Born_in	Born_in
5	...[Biden] (born in [US]) is an American politician...	Born_in	Born_in

Tab. I *The result of alignment of triples* $(Biden, Born\_in, US)$ *and text.*

be applied to sentence-level noise reduction. Moreover, the semantics of sentences are composed of words, and the commonality between sentences is basically derived from the effective words in them. Therefore, mutual information can also play a positive role in distinguishing effective words in sentences.

Aiming at above problems, we propose a novel model for extracting relation based on mutual information and multi-level attention. Our method is based on the hypothesis that correct sentences are connected and often share some commonalities, while wrong sentences are usually irrelevant to others. Mutual information is a measure of the semantic similarity of instances in a package. It represents the degree of relevance between an instance and others in the package. We use mutual information to build attention at the word-level and sentence-level to reduce the impact of noise on the performance of relation extraction. The experimental results show that our model provides significant and consistent improvements in relation extraction, comparing with the state-of-the-art methods.

The main contributions of this paper are summarized as follows:

1. We introduce the concept of mutual information to distant supervised relation extraction. The mutual information of instances is constructed through the attention mechanism to reflect the semantic connection in the package.

2. Mutual information is used to establish multi-level attention. We use mutual information to build attention at the word level and sentence level, which is expected to dynamically reduce the influence of those noisy instances.

3. In order to ensure the accuracy of mutual information, we use dense network combined with PCNN as the coding layer. In the experiments, results show that our model achieves better performance in distant supervised relation extraction.

2. Methodology

We propose a new relation extraction model based on mutual information and multi-level attention, which can denoise the information in the package at the word-level and sentence-level. The semantics of a sentence consists of words, and the commonality between sentences comes from the valid words in the sentence. Therefore, comprehensively considering the influence of words and sentences on mutual information can help the model to better measure the similarity between

samples from multiple perspectives. The overall structure of the model is illustrated in Fig. 1. Our model consists of two main components: dense network module and attention module. In order to improve the computational efficiency of the model, we use dense network as an encoding layer in the model to convert sentences into features at different levels. Based on the similarity between samples, we adopt an attention mechanism to realize the dynamic change of the weight of words and sentences, thereby reducing the influence of noise on the prediction results of the model. We elaborate on these parts in following paragraphs.

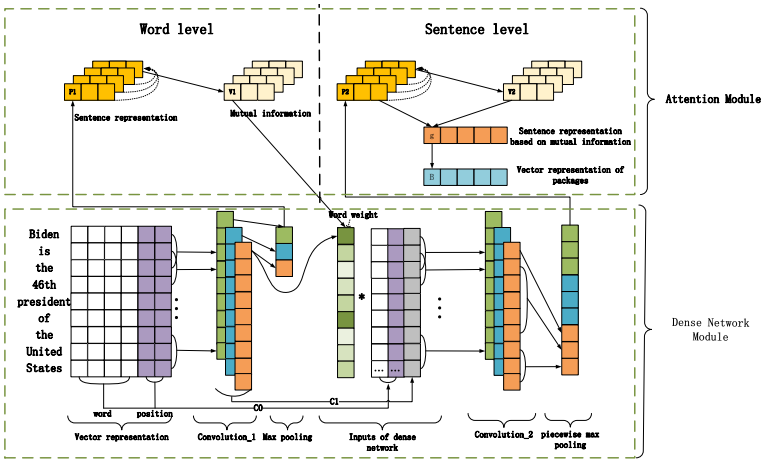

Fig. 1 *The architecture of distant supervision relation extraction model based on mutual information and multi-level attention.*

2.1 Coding layer based on dense network

In order to improve the computational efficiency of the model and the accuracy of mutual information, this paper introduces a dense network for model training [10]. Although the deep neural network performs well in the field of pictures, it is not suitable for processing text information [11]. Therefore, this paper uses a dense network with two layers of convolution as the coding layer of the model. Dense network connects each layer to every other layer in a feed-forward fashion. Each layer obtains additional inputs from all preceding layers and passes on its own feature-maps to all subsequent layers. This structure allows the information in the sentence to be used repeatedly, reducing the impact of noise on the coding result.

For the characteristics of text data, we use the ordinary max-pooling as the first pooling layer. In the second pooling layer, we use the method of piecewise max-pooling (PCNNs) [31]. This method divides any sentence into three paragraphs $\{c_1, c_2, c_3\}$ according to the position of the entity. Then the pooled results of each segment will be spliced, and finally the sentence vector representation is output by

the filter. Through PCNNs, the key information of each sentence can be captured quickly.

As shown in Eq. (1), we input the sentence vector representation x into the layer 1 convolutional neural network, and get the layer 1 convolution output c^1. The input of the second layer of convolutional network is the stitching of two vectors $[c^1, c^0]$, and the output information is c^2.

$$c^k = \begin{cases} x, & k = 0 \\ H_1([c^0]), & k = 1 \\ H_2([c^1, c^0]), & k = 2. \end{cases} \tag{1}$$

In order to prevent the loss of the key information of the input sentence, here H_1 only performs convolutional operation, and convolution and dropout operations are included in H_2.

We perform pooling operations on c^1 and c^2, and the results are expressed as p^1 and p^2 respectively, which are the basis for obtaining mutual information in the next step.

2.2 Acquisition of mutual information

We gather the commonalities of instance in the same relation sample to form mutual information. The specific acquisition process is shown in Fig. 2.

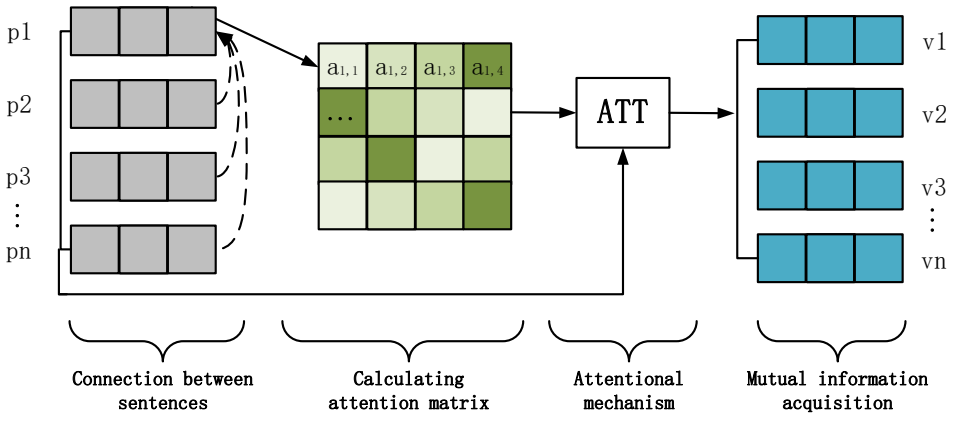

Fig. 2 *The process of acquiring mutual information.*

Assuming there are n sentences in the package, the encoded sentences vector is expressed as $p = p_1, \ldots, p_n, 1 \leq i \leq n$. We obtain the commonalities between sentences through the attention mechanism.

Distant supervision relation extraction generally deals with long texts, in which there is a lot of noise. If we directly use the normalized text vector to calculate the similarity between sentences, the invalid information in the text will affect the

model accuracy. Similar to the processing methods of Wang et al. [25] and Yang et al. [29], here we only use the product s of the sentence vector as a rough estimate of similarity.

Firstly, we correlate each sentence with each other, and get the correlation value $s_{i,j}$, which is a rough estimate of the correlation between the i-th and j-th sentences in the package. The correlation value s represents the similarity of the two sentences. In order to prevent the duplication of instance's own information, when $i = j$, we set the correlation value s to zero. s is defined as follows:

$$s_{i,j} = \begin{cases} 0, & i = j \\ p_i^\top \cdot p_j, & otherwise. \end{cases} \tag{2}$$

Then calculate the attention matrix a based on s:

$$a_{i,j} = \frac{\exp(s_{i,j})}{\sum_{k=1}^{n} \exp(s_{i,k})}, \tag{3}$$

$a_{i,j}$ is the weight of commonality between the i-th sentence and j-th sentences in the package.

Finally, the mutual information v can be obtained by using the attention mechanism:

$$v_i = \sum_{j=1}^{n} a_{i,j} \cdot p_j^k, \tag{4}$$

v_i is the sum of information similar to p_j in the packet.(The superscript of p_j^k indicates that it comes from the k-th layer of the model, and its subscript indicates that it is the representation of the j-th sentence in the packet. The rest of the variables are the same.) In the process of collecting mutual information, key information is easier to be discovered and identified through the interaction and comparison between sentences.

Our model uses a two-layer dense network, so we can get two sentence information vectors p^1 and p^2. They come from the first and second layers of the model, respectively. By Eq. (1)–(4), p^1 and p^2 will be transformed into two corresponding mutual information v^1 and v^2. v^1 comes from the lower layer of the model and is used to improve the anti-noise performance of the model at the word level. v^2 comes from the upper level of the model, and we use it to improve the model's anti-noise performance at the sentence level.

2.3 Sentence-level attention based on mutual information

The traditional attention mechanism judges the matching score between the instance and the relation based on the information in the package. Although this method finally uses all the sentence information to form the representation of the package, the sentences are independent of each other, and the connection between the instances is ignored. Mutual information is an abstraction of commonality in instances, and is an effective way to solve this problem. We attach the mutual information collected from the package to the single sentence representation to obtain the new sentence representation g, which is defined as:

$$g_i = [p_i^2 + v_i^2, p_i^2 \cdot v_i^2]. \tag{5}$$

Among them, $p_i^2 + v_i^2$ is a supplement to the original sentence representation p_i^2 (They come from the second layer of the model). The scoring function can perform relation matching score on sentence i based on the information of all sentences in the package, which avoids the situation of incomplete information in a single sentence, and also emphasizes important information in the sentence. And we refer to the work of [25] to use $p_i^2 \cdot v_i^2$ as the emphasis vector and splice it after the sentence vector representation to further emphasize the mutual information.

We use the attention mechanism to obtain the matching score e_i between sentence p_i^2 and the prediction relation r based on g_i, which has rich semantic information:

$$e_i = g_i W r, \tag{6}$$

where W is the diagonal matrix that measures the score, and r is the query vector for predicting the relation.

The weight β_i of each sentence is defined as:

$$\beta_i = \frac{\exp(e_i)}{\sum_{k=1}^{n} \exp(e_k)}. \tag{7}$$

Finally, we get the packet vector representation b:

$$b = \sum_{i=1}^{n} \beta_i g_i, \tag{8}$$

where B is the final vector representation of the entire package.

2.4 Word-level attention based on mutual information

In distant supervision relation extraction, the model will not only be affected by incorrectly labeled data, but also by noise words in sentences. Not all words in the sentence are effective for relation prediction. The commonality of sentences comes from the words that make up them, so mutual information can help us distinguish the validity of words. In our method, words determine their weight in sentences based on mutual information.

When calculating the correlation between mutual information and words, the influence of noise words is also considered here. Refer to the calculation method of mutual information, here we use the product of the mutual information v^1 and the convolution output of each word in the sentence as the correlation value. But the value should be mapped to the interval [0,1], and the associated values do not affect each other. Here, the sigmoid function is used for normalization, and the weight γ_j of the j-th word in the i-th sentence in the packet is defined as:

$$\gamma_j = sigmoid((v_i^1)^\top c_j^1). \tag{9}$$

In the middle layer of the dense network, the obtained mutual information is associated with the output information of the convolutional layer to obtain the

weight of each word in the sentence. The second layer convolution output c^2 of the model is defined as:

$$c_j^2 = H_2(\gamma_j[c_j^1, c_j^0]), \tag{10}$$

where H_2 is the convolution operation and $[c_j^1, c_j^0]$ represents the splicing vector of c_j^1 and c_j^0.

2.5 Objective function

After enriching the sentence representation, we directly apply it to the original loss function. First use softmax to get conditional probability distributions about different relations.

$$p(r|b,\theta) = \frac{\exp(o_r)}{\Sigma_{k=1}^{n_r} \exp(o_k)}. \tag{11}$$

Among them, r is the representation of a certain relation, n_r is the number of relation types, and o_r is the relation score finally output by the neural network.

The definition of o is as follows:

$$o = Mb + d, \tag{12}$$

where M is the representative matrix of the relation, and d is the bias vector.

Finally, we use cross entropy to define the objective function:

$$J(\theta) = \Sigma_{i=1}^{T} \log(p(r_i|b_i,\theta)). \tag{13}$$

Randomly select a fixed batch of data from the training set for training each time. T is the number of sentences in each batch, and θ is all the parameters of the model. The objective function can be minimized by means of stochastic gradient descent, and use dropout to avoid overfitting. The objective function calculates the loss value on the basis of the vector representation B of the package containing more semantic information, which can effectively alleviate the influence of incorrectly labeled data and improve the anti-noise ability of the model. The entire training procedure is described in Algorithm 1.

3. Experiments

Our experiments are intended to show that our model can capture sentence semantics at the word level and sentence level respectively. In this part, we first introduce the dataset and experimental parameters used. Then, we separately evaluated the effects of sentence-level attention and word-level attention. Finally, we combine the two parts and compare them to some classic methods.

3.1 Datasets

Similar to Lin et al. [14] and Yuan et al. [30], the dataset used in the experiment is the New York Times (NYT) corpus. The corpus was developed by Riedel et al. [20], and it was produced by aligning entities in Freebase and the *New York*

Algorithm 1 A distant supervision noise reduction algorithm based on mutual information.

Step 1: Initialization
1.1 Enter any set of sentences X;
1.2 Initialize the representation vector b of the package to 0;
Step 2: Word-level noise reduction using mutual information
for each $x_i \epsilon X$ **do**
2.1 $c_i^1 = CNN_1(x_i)$ //Use the first layer of convolutional neural network to calculate the encoded information c_i^1;
2.2 $p_i^1 = MaxPooling(c_i^1)$ //Use the pooling layer to generate low-level sentence information to represent p_i^1;
2.3 $Add(p_i^1, P^1)$ //Add p_i^1 to set P^1
2.4 $Add(c_i^1, C^1)$ //Add c_i^1 to set C^1
end for
for each $p_i^1 \epsilon P^1$ **do**
2.5 $v_i^1 = V(p_i^1, P^1)$ //Calculate mutual information v_i^1 by formula 4
2.6 $\gamma_i = \gamma(v_i^1, c_i^1)$ //Calculate the word weight γ_i by formula 9
end for
Step 3: Sentence-level noise reduction using mutual information
for each $c_i^1 \epsilon C^1$ **do**
3.1 $c_i^2 = CNN_2(\gamma[c_i^0, c_i^1])$ //Use the second layer of convolutional neural network to calculate the encoded information c_i^2
3.2 $p_i^2 = PicewiseePooling(c_i^2)$ //Use the segmentation pooling layer to generate low-level sentence information to represent p_i^2
3.3 $Add(p_i^2, P^2)$ //Add p_i^2 to set P^2
end for
for each $p_i^2 \epsilon P^2$ **do**
3.4 $v_i^2 = V(p_i^2, P^2)$ //Calculate mutual information v_i^2 by formula 4
3.5 $g_i = [p_i^2 + v_i^2, p_i^2 \cdot v_i^2]$ //Generate a new sentence vector representation g_i
3.6 $b = b + ATT(g_i)$ //Use the attention mechanism to calculate the vector representation b of the package
end for

Times. NYT corpus is also the standard data for distant supervision relation extraction. Freebase is a knowledge base similar to Wikipedia, both of which store large amounts of structured data in the form of triples. In the NYT corpus, the training dataset contains 522,611 sentences, 281,270 pairs of entities, and 18,252 relation triples and the test dataset includes 172,448 sentences, 96,678 pairs of entities and 1,950 relational triples. The dataset consists of 53 relation types, including NA relation.

3.2 Experimental environment and settings

In this paper, we employ the Skip-gram model(word2vec) [16] to train the word embeddings on the NYT corpus. Word2vec first constructs a vocabulary from the training text data and then learns vector representations of the words. Our experiments utilize 50-dimensional vectors.

The Tab. II is the operating environment of this experiment.

type	versions
OS	Win10
Tensorflow	1.3.6
Python	3.6.6
CPU	i7-8700k
GPU	RTX 2080
RAM	16G

Tab. II *Experimental environment.*

Since this experiment introduces a dense network, the depth of the neural network is increased by one layer compared to Zeng et al. [31]. For the parameters of the convolutional layer, we follow the settings used in Lin et al. [14] and Yuan et al. [30]. For other parameters, we continue to follow the settings of Zeng et al. [31]. The number of iterations in this experiment is 50 rounds. The specific parameter settings are shown in Tab. III.

Setting	Number
Window size	3
CNN-1 convolution kernel size	80
CNN-2 convolution kernel size	110
Word dimension	50
Position dimension	5
Batch size	160
Dropout probability	0.5
Learning rate	0.05

Tab. III *Specific parameters.*

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. The held-out evaluation provides an accurate measurement and does not require expensive manual evaluation. By convention, we report both the precision/recall curves and Precision@N (P@N) in our experiments.

3.3 Experimental results of sentence-level distant supervision relation extraction

To prove the influence about our sentence-level attention method , we compared with different methods by held-out evaluation. We choose the non-IID relevance embedding(PCNNs+N) of Yuan et al. [30] as the baseline. Similar to our method, PCNNs+N uses the connection of sentences to guide the model to reduce noise. Compared with the original attention method of Lin et al.[14], the model effect has been greatly improved. PCNNs+V represents our sentence-level attention method. Fig. 3 shows the comparative experimental results of the two methods on the same dataset.

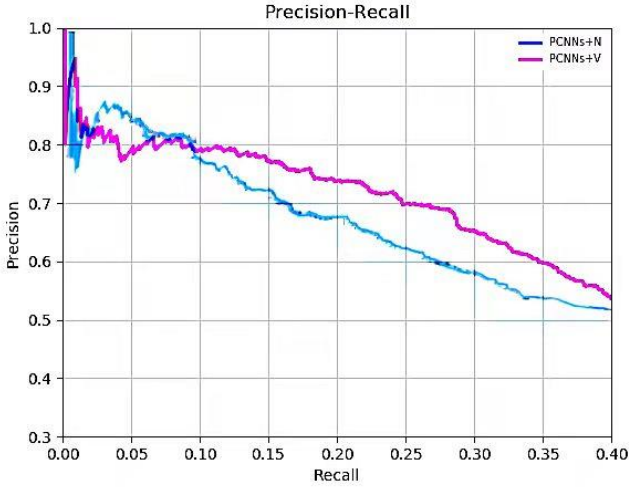

Fig. 3 *PR curves of sentence-level experiment.*

Fig. 3 shows that the experimental effect of PCNNs+V is significantly better than PCNNs+N. These results indicate that PCNNs+V utilizes the commonalities between instances more effectively than PCNNs+N. In our sentence-level attention method(PCNNs+V), mutual information is attached to the vector representation of the sentence, which more effectively expresses the connection of the instances.

Tab. IV reports the Precision@N of the two methods. The effect of our method at P@100 is lower than that of PCNNs+N, but as the recall rate increases, the performance of PCNNs+V improves significantly. This proves that the model has good generalization ability.

P@N(%)	P@100	P@200	P@300	Average
PCNNs+N	81.0	79.5	76.7	79.1
PCNNs+V	78.2	80.5	79.0	79.2

Tab. IV *Results of sentence-level experiment.*

We conduct ablation study to further verify the role of mutual information in the sentence representation g. Here we split g into two parts, namely PCNNs+Vv that only supplements a single mutual information ($g_1 = [p^2 + v^2]$) and PCNNs+Vs that only splice the emphasis vector ($g_2 = [p^2 \cdot v^2]$). We compare the PCNNs+Vv, PCNNS+Vs, PCNNs+V and PCNNs+ATT methods. PCNNs+ATT represents the original attention mechanism of Lin et al. [14].

The results are shown in Fig. 4. The PCNNs+Vv and PCNNS+Vs methods have better performance than the PCNNs+ATT. Because both methods are effective expressions of the connection of instance. But the effect of PCNNs+V is

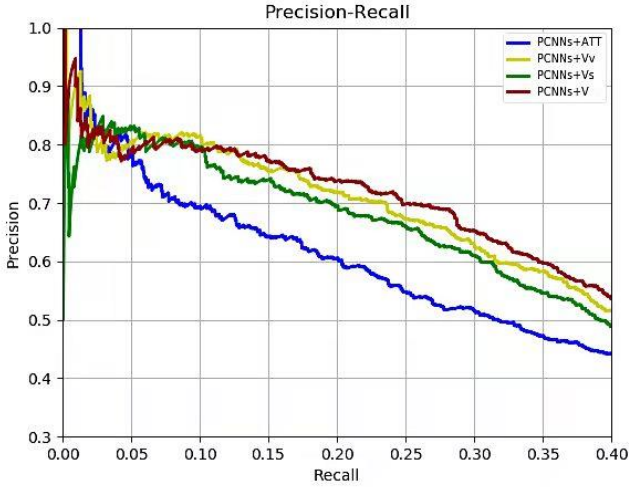

Fig. 4 *PR curves of sentence-level ablation study.*

better than PCNNs+Vv and PCNNS+Vs. It demonstrates the combination of the two methods further improves the expressive ability of the model.

3.4 Experimental results of word-level distant supervision relation extraction

This part mainly verifies the effectiveness of our word-level attention method (PCNNs+D) based on dense network and mutual information. We select the linear attenuation simulation (PCNNs+W) of Yuan et al. [30] as the baseline. In order to further prove the role of mutual information at the word level, we add a new algorithm for comparison. Compared with PCNNs+D, this method removes the mutual information part and only contains the dense network part. We use DensePCNNs+ATT to represent this method. The results are shown in Fig. 5.

As shown in Fig. 5, the results of DensePCNNs+ATT using only dense networks are also better than PCNNs+ATT. This indicate that dense networks can better capture the key information in sentences compared to single-layer convolutional neural networks. However, the results of PCNNs+D that uses mutual information at the word level are much better than DensePCNNs+ATT. This proves the effectiveness of mutual information in word-level noise reduction. The method combining dense network and mutual information will have stronger performance in word-level.

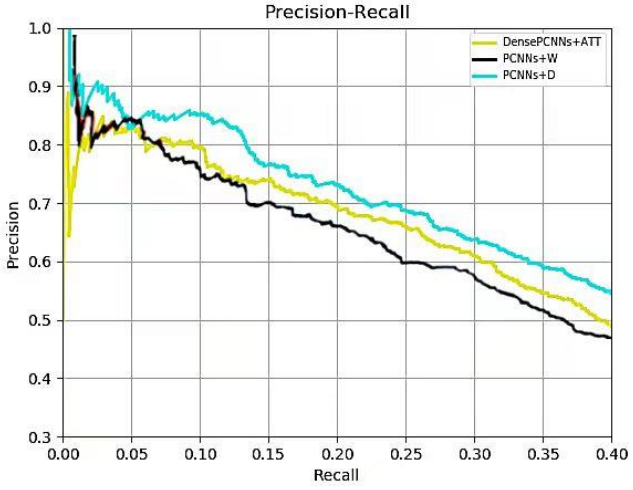

Fig. 5 *PR curves of word-level experiment.*

3.5 Comprehensive comparison and evaluation

To evaluate the proposed method, we select the following five traditional methods for comparison.

- Mintz Mintz et al.[17] proposed a original distant supervision model.
- MIMLRE Surdeanu et al.[23] proposed a multi-instance and multi-label model.
- PCNNs+MIL Zeng et al.[31] proposed a model of piecewise convolutional neural networks (PCNNs) with multi-instance learning.
- PCNNs+ATT Lin et al.[14] proposed a selective attention model combined with PCNNs and CNNs.
- PCNNs+WN Yuan et al.[30] proposed a linear attenuation simulation and Non-IID relevance embedding model.
- PCNNs+DV is our method.

PCNNs+W, PCNNs+N and DensePCNNs+ATT have already appeared in previous experiments. We also add them to the comparison.

Fig. 6 shows that the precision-recall curves for each method. In order to ensure the accuracy of the results, the comparison experiments all use the same dataset. PCNNs+DV is a combination of PCNNs+V and PCNNs+D. It uses mutual information to build attention at the word level and sentence level. The results show that our PCNNs+DV is generally better than all baseline. This proves that mutual information is an effective representation of the commonality of instance, and the

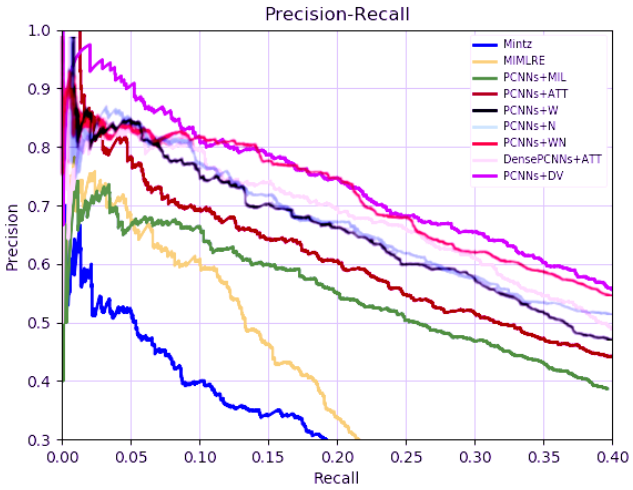

Fig. 6 *PR curves of comprehensive experiment.*

multi-level attention mechanism is a reasonable method to extract the semantics of the package. The results demonstrate that our method is an effective way to distant supervised relation extraction.

P@N(%)	P@100	P@200	P@300	Average
Mintz	51.8	50.0	44.8	48.9
MIMLRE	70.9	62.9	60.9	64.9
PCNNs+MIL	72.3	69.7	64.1	68.7
PCNNs+ATT	81.1	71.1	69.4	73.9
PCNNs+W	83.0	77.0	72.0	77.0
PCNNs+N	81.0	79.5	76.7	79.1
PCNNs+WN	83.0	82.0	80.3	81.8
DensePCNNs+ATT	82.1	80.5	74.4	79.0
PCNNs+DV	90.0	83.0	79.7	84.3

Tab. V *Results of comprehensive experiment.*

In Tab. V, we report the P@100,P@200, P@300 and the average of them for Mintz, MIMLRE, PCNNs+MIL,PCNNs+ATT, PCNNs+W, PCNNs+N, PCNNs+WN, DensePCNNs+ATT, and PCNNs+DV. As shown in Tab. V, our model PCNNs+DV achieves 90.0 on p@100 and 82.0 on p@200, which is the best among all methods. Although the effect of PCNNs+DV on p@300 is 79.4 compared to 80.3 of PCNNs+WN, it is not the best. However, the overall effect is still higher than other methods. In terms of mean, our method reached the highest value of 84.3 among all methods, which is about 2.5 higher than PCNNs+WN, about 5.3 higher

than DensePCNNs+ATT, and about 11 higher than PCNNs+ATT. These results demonstrate that our method possesses important effects for distant supervision.

4. Conclusion

In this paper, we propose a novel model of distant supervision relation extraction based on mutual information and multi-level attention. The model uses mutual information to measure the relevance of instances, and builds multi-level attention based on this relevance to alleviate the problem of incorrect labeling in distant supervised relation extraction. Firstly, we use dense network combined with PCNN as the coding layer to ensure the accuracy of mutual information. Then, mutual information is constructed through the attention mechanism to reflect the semantic connection between instances. Finally, mutual information is used to establish multi-level attention at the word level and sentence level. It assigns higher weights to correct instances and suppresses the influence of incorrect labels on the model. The results demonstrated that multi-level attention method based on mutual information is effective, and our method can reduce the side effect of noisy information. In the future, we will explore how to extend our method to more challenging document-level relation extraction.

Acknowledgement

This work was supported by the National Natural Science Foundation of China (NSFC) (grant number 42050103, 62076108, U19A2061).

References

[1] AUER S., BIZER C., KOBILAROV G., LEHMANN J., CYGANIAK R., IVES Z.G. DBpedia: A Nucleus for a Web of Open Data. In: *ISWC/ASWC*, 2007, pp. 722–735.

[2] BAI L., JIN X., ZHUANG C., CHENG X. Entity Type Enhanced Neural Model for Distantly Supervised Relation Extraction (Student Abstract). In: *AAAI*, 2020, pp. 13751–13752.

[3] BOLLACKER K.D., EVANS C., PARITOSH P., STURGE T., TAYLOR J. Freebase: a collaboratively created graph database for structuring human knowledge. In: *SIGMOD*, 2008, pp. 1247–1250.

[4] CHEN J., JIMÉNEZ-RUIZ E., HORROCKS I., ANTONYRAJAH D., HADIAN A., LEE J. Augmenting Ontology Alignment by Semantic Embedding and Distant Supervision. In: *ESWC*, 2021, pp. 392–408.

[5] CHEN L., FENG Y., HUANG S., LUO B., ZHAO D. Encoding implicit relation requirements for relation extraction: A joint inference approach. *Artificial Intelligence*. 2018, 265, pp. 45–66.

[6] FU Z., SHI B., LAM W., BING L., LIU Z. Partially-Aligned Data-to-Text Generation with Distant Supervision. In: B. WEBBER, T. COHN, Y. HE, Y. LIU, eds. *EMNLP*, 2020, pp. 9183–9193.

[7] HAN X., LIU Z., SUN M. Neural Knowledge Acquisition via Mutual Attention Between Knowledge Graph and Text. In: *AAAI*, 2018, pp. 4832–4839.

[8] HE J., ZHAO Y., LUO G. DSREFC: Improving Distantly-supervised Neural Relation Extraction Using Feature Combination. In: *ICMLC*, 2020, pp. 524–529.

[9] HETTINGER L., ZEHE A., DALLMANN A., HOTHO A. EClaiRE: Context Matters! - Comparing Word Embeddings for Relation Classification. In: *GI-Jahrestagung*, 2019, pp. 191–204.

[10] HUANG G., LIU Z., van der MAATEN L., WEINBERGER K.Q. Densely Connected Convolutional Networks. In: *CVPR*, 2017, pp. 2261–2269.

[11] HUANG Y.Y., WANG W.Y. Deep Residual Learning for Weakly-Supervised Relation Extraction. In: *EMNLP*, 2017, pp. 1803–1807.

[12] JAT S., KHANDELWAL S., TALUKDAR P.P. Improving Distantly Supervised Relation Extraction using Word and Entity Based Attention. In: *AKBC@NIPS*, 2017.

[13] JI G., LIU K., HE S., ZHAO J. Distant Supervision for Relation Extraction with Sentence-Level Attention and Entity Descriptions. In: *AAAI*, 2017, pp. 3060–3066.

[14] LIN Y., SHEN S., LIU Z., LUAN H., SUN M. Neural Relation Extraction with Selective Attention over Instances. In: *ACL*, 2016.

[15] LUO B., FENG Y., WANG Z., ZHU Z., HUANG S., YAN R., ZHAO D. Learning with Noise: Enhance Distantly Supervised Relation Extraction with Dynamic Transition Matrix. In: R. BARZILAY, M. KAN, eds. *ACL*, 2017, pp. 430–439.

[16] MIKOLOV T., CHEN K., CORRADO G., DEAN J. Efficient Estimation of Word Representations in Vector Space. *Computer Science*. 2013.

[17] MINTZ M., BILLS S., SNOW R., JURAFSKY D. Distant supervision for relation extraction without labeled data. In: *ACL/IJCNLP*, 2009, pp. 1003–1011.

[18] MOREIRA J., OLIVEIRA C., MACÊDO D., ZANCHETTIN C., BARBOSA L. Distantly-Supervised Neural Relation Extraction with Side Information using BERT. In: *IJCNN*, 2020, pp. 1–7.

[19] OUYANG X., CHEN S., WANG R. Semantic Enhanced Distantly Supervised Relation Extraction via Graph Attention Network. *Information (Switzerland)*. 2020, 11(11), pp. 528.

[20] RIEDEL S., YAO L., MCCALLUM A. Modeling Relations and Their Mentions without Labeled Text. In: *ECML/PKDD*, 2010, pp. 148–163.

[21] SHEN S., DUAN S., GAO H., QI G. Improved distant supervision relation extraction based on edge-reasoning hybrid graph model. *J. Web Semant.* 2021, 70, pp. 100656.

[22] SUCHANEK F.M., KASNECI G., WEIKUM G. Yago: a core of semantic knowledge. In: *WWW*, 2007, pp. 697–706.

[23] SURDEANU M., TIBSHIRANI J., NALLAPATI R., MANNING C.D. Multi-instance Multi-label Learning for Relation Extraction. In: *EMNLP-CoNLL*, 2012, pp. 455–465.

[24] TAKAMATSU S., SATO I., NAKAGAWA H. Reducing Wrong Labels in Distant Supervision for Relation Extraction. In: *ACL*, 2012, pp. 721–729.

[25] WANG Y., LIU K., LIU J., HE W., LYU Y., WU H., LI S., WANG H. Multi-Passage Machine Reading Comprehension with Cross-Passage Answer Verification. In: *ACL*, 2018, pp. 1918–1927.

[26] XIAO C., YAO Y., XIE R., HAN X., LIU Z., SUN M., LIN F., LIN L. Denoising Relation Extraction from Document-level Distant Supervision. In: B. WEBBER, T. COHN, Y. HE, Y. LIU, eds. *EMNLP*, 2020, pp. 3683–3688.

[27] XU K., REDDY S., FENG Y., HUANG S., ZHAO D. Question Answering on Freebase via Relation Extraction and Textual Evidence. In: *ACL*, 2016.

[28] XU W., HOFFMANN R., ZHAO L., GRISHMAN R. Filling Knowledge Base Gaps for Distant Supervision of Relation Extraction. In: *ACL*, 2013, pp. 665–670.

[29] YANG R., ZHANG J., GAO X., JI F., CHEN H. Simple and Effective Text Matching with Richer Alignment Features. In: *ACL*, 2019, pp. 4699–4709.

[30] YUAN C., HUANG H., FENG C., LIU X., WEI X. Distant Supervision for Relation Extraction with Linear Attenuation Simulation and Non-IID Relevance Embedding. In: *AAAI*, 2019, pp. 7418–7425.

[31] ZENG D., LIU K., CHEN Y., ZHAO J. Distant Supervision for Relation Extraction via Piecewise Convolutional Neural Networks. In: *EMNLP*, 2015, pp. 1753–1762.