Applied Intelligence
https://doi.org/10.1007/s10489-021-02632-8

# Piecewise convolutional neural networks with position attention and similar bag attention for distant supervision relation extraction

Weijiang Li[1,2] · Qing Wang[1,2] · Jiekun Wu[1,2] · Zhengtao Yu[1,2]

Accepted: 9 June 2021

## Abstract

In relation to extraction tasks, distant supervision is a very effective method, which can automatically generate training data via aligning KBs and texts, thereby solving the problem of manually labeling data. However, distant supervision inevitably accompanies the wrong labeling problem. The paper presents a neural relation extraction method to deal with the problem of noisy words and poor feature information in the one-sentence bags generated by distant supervision. Previous studies mainly focus on sentence-level denoising and even bag-level denoising by designing neural networks. In the paper, we propose a piecewise convolutional neural network with position attention and similar bag attention for distant supervision relation extraction(PCNN-PATT-SBA). First, we propose a position attention based on Gaussian distribution, by modeling the position relationship between non-entity words and entity words to assign weights for the words of the sentence, which is expected to reduce the influence of those noisy words. In addition, we propose similar bag attention based on the feature similarity between different bags, which merges the features of similar bags to solve the problem of poor feature information in one-sentence bags. Experimental results on the New York Times dataset demonstrate the effectiveness of our proposed position attention and similar bag attention modules. Our method also achieves better relation extraction accuracy than state-of-the-art methods on this dataset. And compared to the bag-of-sentence attention model, the P value is increased by 6.9%, Compared with selective attention over instances (PCNN-ATT), an increase of 25.6%, compared to Instance-Level Adversarial Training (PCNN-HATT), an increase of 12.1%.

**Keywords** Distant supervision · Position attention · Similar bag attention · Gaussian distribution

## 1 Introduction

Relation extraction [1, 2] is one of the most important tasks in NLP, which aims to extract semantic relations between entities. For example, the sentence "[Barack Obama]e1 was born in [Hawaii]e2" expresses the BornIn relation between the entity pair Barack Obama and Hawaii. Relation extraction can apply such as question answering [3, 4] and web search [5].

In recent years, many efforts have been invested in relation extraction, especially in supervised relation extraction, either early works based on handcrafted features [1, 2] or recent works based on neural networks [6–8]. These methods all follow a supervised learning approach which suffers from the lack of large-scale manually labeled data. To address this issue, Mintz [9] proposed the distant supervision method that automatically generates a largescale, labeled training set by aligning entities in knowledge graph [10] to corresponding entity mentions in natural language sentences, solving the problem of tedious manual annotation. the distant supervision method is based on an assumption: if two entities participate in a relation, all sentences that mention these two entities express that relation. But this assumption is not always true, sometimes, different sentences with the same entity pairs have different contexts expressing different relation. It is inevitable that there exists noise sentences in the data labeled by distant supervision.

✉ Weijiang Li
hrbrichard@126.com

[1] School of Information Engineering and Automation, Kunming University of Science and Technology, Kunming, Yunnan, China

[2] Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology, Kunming, 650500, China

Published online: 24 July 2021

To solve the problem of the above method, we propose a piecewise convolutional neural networks with position attention and similar bag attention for distant supervision extraction of relation(PCNN-PATT-SBA). At the word level, in general, the closer the words are to the entity words, the more important they are. In this paper, based on this feature, we use Gaussian distribution to reweight the words in a sentence to increase the weight of important words and decrease the weight of noisy words. At the sentence level, the selective attention [11] is used to reduce the influence of noisy sentences. At the bag level, since the feature information of one-sentence sentence bag is too little, this paper proposes the similar bag attention method, which solves the problem of poor feature information of one-sentence bag by fusing the sentence bags similar to one-sentence bag to improve the sentence bag feature. The main contributions of this paper are summarized as follows:

- We propose a position attention mechanism based on Gaussian distribution, which reduces the influence of noise words on the effect of relation extraction by assigning the weight of each word in the sentence.
- We propose a similar bag attention mechanism, which solves the problem of poor feature information in one-sentence bags by merging the features of similar bags.
- Our methods achieve better relation extraction performance than state-of-the-art models on the widely used New York Times (NYT) dataset.

## 2 Related work

Supervised relation extraction task aims to extract relations from sentences with supervised data, but most of these methods need Large-scale manually labeled data, which is time consuming and labor intensive. To address this issue, Mintz [9] proposed distant supervision by aligning plain text with Freebase to get labeled data. However, distant supervision inevitably accompanies with the wrong labeling problem. To alleviate the wrong labeling problem, Riedel [12] regard distant supervision relation extraction as a multi-instance learning problem, which extracts relation from bag instead of sentence. Later, [13, 14] found that an entity pairs may have more than one relation, so the method of multi-label learning was introduced on the basis of multi-instance learning. In recent years, with the development of deep learning [15], neural network-based methods have been widely used in distant supervision relation extraction. Socher [16] first used recurrent neural networks (RNN) in relation extraction tasks. Afterwards. [6] used Convolutional Neural Networks (CNN) in relation extraction tasks. Zeng [17] proposed a piecewise convolutional neural network (PCNN) to model sentence representation and select the most reliable sentence as the bag representation. Methods based on neural networks are mainly divided into follows: methods based on attention mechanism, methods based on data filtering, and methods based on model structure and loss function.

Lin [11] used PCNN as a sentence encoder, and proposed a selective attention mechanism, which calculates the bags representation by the weighted sum of all sentence representations. Ji [18] adopted a similar strategy of selective attention and combined the entity description to calculate the sentence weight. Similar strategy of selective attention used entity to vector subtraction to generate an attention matrix. The entity description method is to approximate the entity pair vector to the description vector. Liu [19] proposed a soft-label method to reduce the influence of noisy sentences, which combines the relational scores based on the entity-pair representation and the confidence of the hard label to obtain a new label. [20] extends the selective attention to cross-relation cross-bag selective attention and trains the model more noise-robust. [8] used cross lingual attention to consider the information consistency and complementarity among cross-lingual texts. The representation of bags in all the above methods is obtained by the weighted sum of sentence embedding representations without first reduce the influence of noisy words. These methods only consider the sentence-level attention, but do not consider the problem of noisy words. [21] proposed entity position-aware attention to reduce the influence of noisy words, which combined the location embedding and word features to form an attention matrix. Different from the method, our position attention in this paper directly maps this distance relationship into the corresponding attention weights, considering more the importance of the distance relationship rather than the position features.

Statistics show that up to 80% of the bags in the NYT dataset contain only one sentence, so each one-sentence bag contains very limited features. Li [22] proposed a self-attention enhanced Selective Gate(SeG) method to overcome the problem occurring in selective attention, which is caused by one-sentence bags. SeG is a pool-based gate that uses rich context representation as an aggregator to generate a bag-level representation for final relationship classification. Although the method overcomes the problem occurring in selective attention, but cannot solve the problem of poor feature information in one-sentence bags. Ye [23] proposed the inter-cross-bag method, which Obtained the attention matrix by calculating the similarity between the target bag and other bags, puts bags with the same label in a group, and then uses the attention mechanism to merge all bags in the group to obtain a super-bag, thereby reducing the influence of noise bag. The method solves the problem of noisy bags, but depends on the features of the same label bags. Above these method can solve the problem of noise sentences or noise bags, but they cannot solve the problem of poor feature information in one-sentence bags. Therefore, we propose a similar bag attention mechanism, different from the inter-cross-bag method [23], which does not depend on whether the bags have the same label, as long as there is sufficient feature similarity between the bags, these similar feature bags are merged into the current bag. The method solves the problem of poor feature information by fusing the characteristics of similar sentence bags.

## 3 PCNN-PATT-SBA model

In this section, we introduce a piecewise convolutional neural network with position attention and similar bag attention model. Let $bag_i = \{s_1, s_2, \cdots, s_m\}$ denote a set of sentences which have the same relation label given by distant supervision, and $m$ is the number of sentences within this set. Let $s_i = \{w_1, \cdots, e_1, \cdots, e_2, \cdots, w_n\}$ indicates the sequence of words in the sentence $s_i$. $e_1$ and $e_2$ are the entity pairs in the sentence $s_i$, and $n$ is the number of words in sentence. First, each word $w_i$ within the sentence $s_i$ use word embeddings mapped into a multi-dimensional feature vector sequence $W_i$, and then these vectors $W_i$ are concatenated to obtain the word vector matrix $W^{d_w} = W_1 \oplus W_2 \oplus \cdots \oplus W_n$, and the dimension is $d_w$.

To describe the position information of two entities, we uses position embeddings [17] to map the relative distances between the i-th word and the first entity $e_1$ and $e_2$ into feature vector $P_1^i$ and $P_2^i$, and concatenate the position vectors to obtain the position vector matrix $P_1^{d_p} = P_1^1 \oplus P_1^2 \oplus, \ldots, \oplus P_1^n$ , $P_2^{d_p} = P_2^1 \oplus P_2^2 \oplus, \ldots, \oplus P_2^n$. $P^{d_p} = P_1^{d_p} \oplus P_2^{d_p}$, and the dimension is $2d_P$.

Finally, these three vector sequences are concatenated to get the sentence representation $S_i = W^{d_w} \oplus P^{d_p}$, the dimension is $d_w + 2d_p$.

Figure 1 is a model diagram of sentence representation. We first re-assign different weights to each word representation through the location attention mechanism, and then use PCNN to extract the three local features of the input vector and connect these three features together. Figure 2 is the framework diagram of the PCNN-PATT-SBA model. The sentences representation obtained in Fig. 1 are spliced together to form a bag. By calculating the degree of matching between the target bag and each bag, the top N most matching bags are obtained, and each bags are assigned appropriate weights, and bags representation is added together to obtain the bag representation G, and finally G is subjected to softmax to obtain the final relationship classification result.

### 3.1 Position attention (PATT)

In order to reduce the influence of noisy words, we propose the position attention ,which is applied to the input layer of the neural network to re-assign the weights to the word representations in the sentence.

Gaussian distribution, also known as normal distribution, is generally used in the field of probability and statistics. Vilis [24] proposed using Gaussian embedding to obtain word vectors, being inspired by it, and we use Gaussian distribution function to calculate the importance weight of words. Let sentence $s_i = \{w_1 \ldots e_1 \ldots e_2 \ldots w_n\}$, $e_1$ and $e_2$ are two entities in the sentence. The task of relation extraction is to extract the relation of the entity pair ,but there are only a few words that express the entity-pair relation in sentence and the other words called noise word may affect the performance of relation extraction. Therefore, it is necessary to reduce the weight of these noise words. Generally speaking, words that are closer to the entity word are more likely to express the relation between the entity pairs, and correspondingly, the higher the importance. Therefore, the words in the sentence can be re-weighted according to the distance relationship with the entity word. The position attention algorithm is as follows.

**Algorithm 1** PATT algorithm.

```
Input: the word vector and position feature vector are
    concatenated $S_i$ as input;
Output: return $S_i'$ , where $S_i'$ is a vector weighted by PATT.

for $iteration\ t$ do
    Using (1) to calculate the importance weights
    of each word relative to $entity_1$ and $entiry_2$ , then
    get $G_1$ and $G_2$;
    Using (2)) softmax normalization $(G_1 + G_2)$ ;
    Using (3)) to re-weight the words of the
    sentence $S_i$ ,getting $S_i'$ .
end for
```

As shown in Table 1, it shows the position distribution of each word in sentences S1 and S2 relative to entity pair $e_1$ and $e_2$, and $d_1$ is distance sequence relative to $entity_1$ and $d_2$ is distance sequence relative to $entity_2$. We use Gaussian function to model the position relationship between words. The definition of Gaussian function is shown in (1). In the paper, we experimentally conclude that $\mu = 0$ and $\sigma = 0.5$, model achieves best performance, G(x)∼(0,0.5$^2$), and specific comparison experimental results are shown in the experimental section. Let sentence vector representation $S_i = \{x_1, x_2, \cdots, x_n\}$, and n is the length of the sentence. We map the distance sequence $d_1$ and $d_2$ into the function

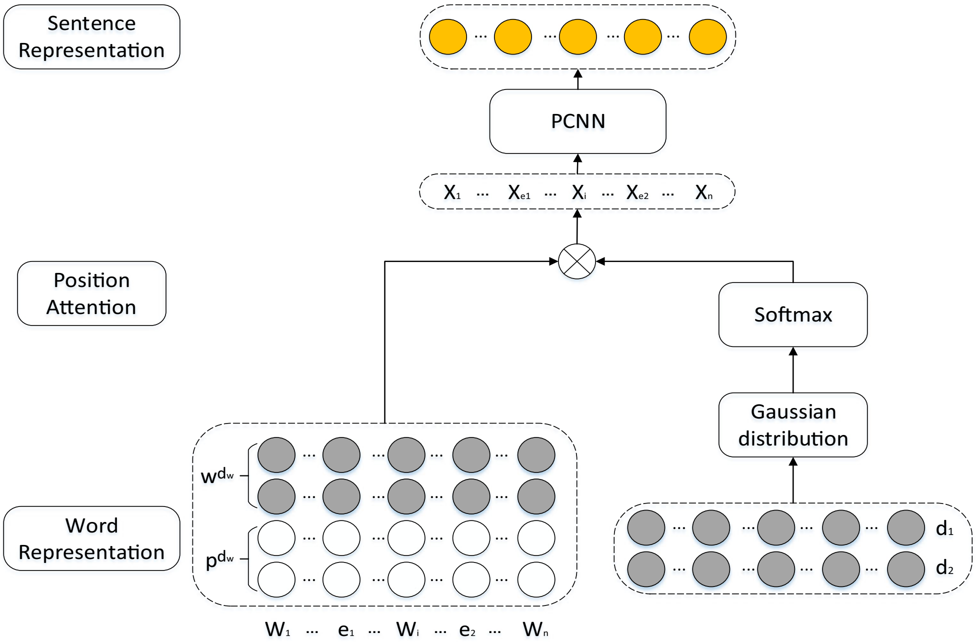


**Fig. 1** Sentences representation model structure

$G(x)$, and x is distance value between words and entities, getting the corresponding Gaussian sequence value $G_1$ and $G_2$. First we add the values of $G_1$ and $G_2$ at the corresponding positions, then using the softmax function to normalize, as shown in (2), finally, re-weighting the vector representation of words in the sentence, as shown in (3).

$$G(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1}$$

$$a = softmax(G_1 + G_2) \tag{2}$$

$$S_i' = S_i \cdot a \tag{3}$$

### 3.2 Piecewise convolutional neural network (PCNN)

In the task of relation extraction, the main challenge is that the length of the sentence is variable and important information may appear anywhere in the sentence, so it is necessary to use all local information to predict global relations. We use convolutional neural networks to extract the features of sentences. The convolutional layer uses a sliding window to extract the local features of the sentence,

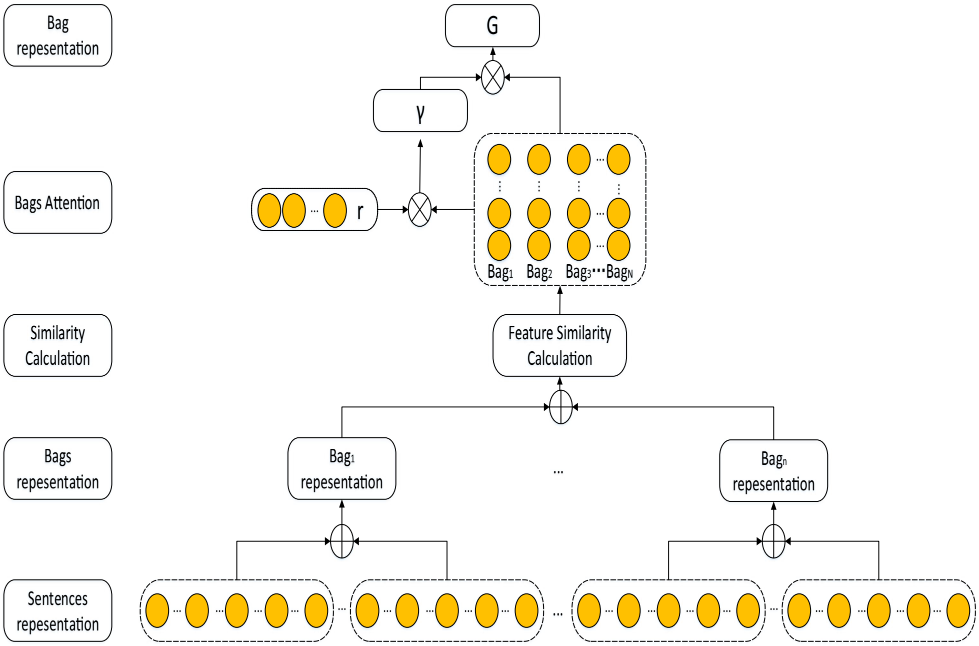


**Fig. 2** PCNN-PATT-SBA model structure

**Table 1** The position distribution of words relative to entities in sentences

| Sentence | The distance $d_1$ of the word relative to the entity $e_1$ | The distance $d_2$ of the word relative to the entity $e_2$ |
|---|---|---|
| S1:**Steve Jobs** was the co-founder and CEO of **Apple** and formerly Pixar | 0,1,2,3,4,5,6,7,8,9,10 | −7,-6,-5,-4,-3,-2,-1,0,1,2,3 |
| S2:**Barack Obamna** was the 44th president of **the United States** | 0,1,2,3,4,5,6 | −6,-5,-4,-3,-2,-1,0 |

and then combines all the local features through a max pooling operation to obtain a fixed-size vector for the input sentence.

Max pooling can only extract a local max feature of the sentence, while ignoring the context information of the sentence. Therefore, [17] proposed the method of PCNN, which is a variant of CNN. In the relation extraction task, PCNN uses piecewise max pooling, and divides each convolution filter $c_i$ into three parts $(c_{i1}, c_{i2}, c_{i3})$ by entity $e_1$ and $e_2$, $i \in [1, d_c]$, and the max pooling process is executed in three parts, defined as follows:

$$[x]_i = \{max(c_{i1}) : max(c_{i2}) : max(c_{i3})\} \tag{4}$$

$[x]_i$ is the concatenated vector of the three max pooling, $[x]_i \in R^{3d_c}$. The structure of the PCNN model based on position attention is shown in Fig. 3.

### 3.3 Similar bag attention (SBA)

Note that almost 80% of bags from popular relation extraction dataset NYT consist of only one sentence [20]. Due to the one-sentence bag contains very limited feature information and the performance of only using sentence-level attention for one-sentence bags is not significant. Therefore, we propose similar bag attention mechanism to to solve the problem of poor feature information in one-sentence bags.

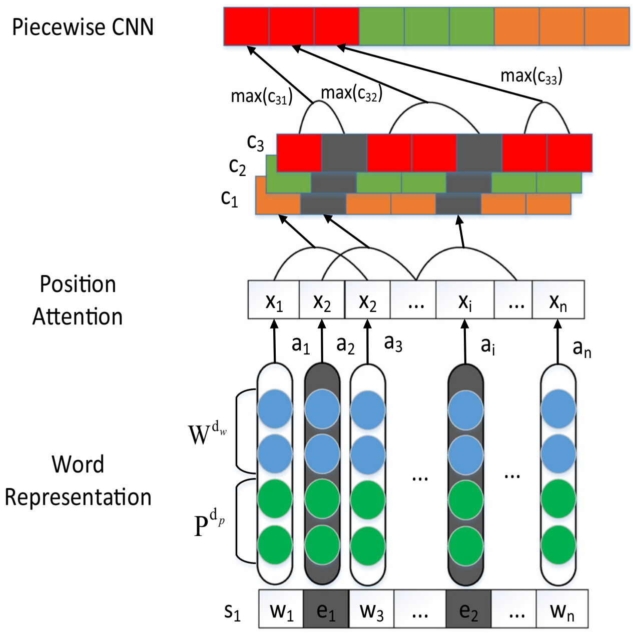


**Fig. 3** PCNN model structure based on position attention

One-sentence bags have poor features, which will inevitably affect the performance of the model during feature training. However, there may be similar features between different bags, so similar features of other bags can be merged in the current bag thereby increasing the features. We propose similar bag attention mechanism based on feature similarity between bags, which is different from the Inter-Bag Attention [23] using bag-level attention to merge same label bags. Similar bag attention mechanism is that regardless of whether the bags have the same label, as long as they have enough the feature similarity put into the bag set $Group^j$, and then these bags features of the $Group^j$ are combined through the attention mechanism. The similar bag attention algorithm as follows.

**Algorithm 2** SBA algorithm.

```
Input: The feature representation of Bag as input;
Output: return $o_r^{n_r}$, where $n_r$ is the number of relation categories.

for $iteration\ t$ do
    Using (5)) to calculate the feature similarity between the current bag and all other bags, and sort the values of similarity from high to low;
    Selecting the bag of similarity TopN and put them in the $Group^j$;
    Using (7)) and (8)) to calculate the attention weight matrix $\gamma_i$ of each bag in $Group^j$;
    Using (6)) to weight each sentence bag of $Group^j$, and get the weighted attention feature vector as $G^j$;
    Using (9)), (10)) and (12)) to predict the relation of $G^j$;
    Finally, use the loss function (11)), Adadelta optimizer to update the model parameters.
end for
```

The similar bag attention mechanism uses $Bag_i$ as the basic unit. The $Bag_i$ merged the features of all sentences in the sentence bag through sentence-level attention [11].

In the realization of the method, similar bag attention first calculates the feature similarity between current $Bag_i$ and all bags in the current $batch_size$. The calculation process is shown in (5). These feature similarity values are sorted in descending order, and the bags corresponding to the topN feature similarity values are selected to combine to $Group^j = \{Bag_1, \cdots, Bag_N\}$. $Bag_1$ is the target bag that we need increse feature, where the label of $Group^j$ and $Bag_1$ are the same.

$$similarity(Bag_1, Bag_i) = Bag_1 Bag_i^T \tag{5}$$

$i = \{1, 2, \ldots, n\}, n = batch\_size$, we first calculate the attention weights of all similar bags in $Group^j$,and then merging the features of these bags according to different attention weights, the merged of all bags in $Group^j$ is expressed as $G^j$.The calculation process is shown in (6).

$$G_j = \sum_i^N \gamma_i Group_i^j \tag{6}$$

Where N is the number of bags in $Group_j$, and $\gamma_i$ is the weight of each bag.

$$\gamma_i = \frac{exp(e_i)}{\sum_k^N exp(e_k)} \tag{7}$$

$e_i$ is a query function. It calculates the matching size between $Group_j^i$ and the label of $Group_j$ to obtain the weight matrix of different bags. $e_i$ is defined as follows:

$$e_i = Group_i^j Br \tag{8}$$

Where $B$ is the weight diagonal matrix, $r$ is the vector representation of the relation label, and $j = \{0, 1, \cdots, N\}$.

Finally, we define a conditional possibility function, and classify relations through the softmax layer:

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

$n_r$ is the number of class of the relation classification, $o_r$ is the final output of the neural network, which is defined as follows:

$$o_r = MG^j + d \tag{10}$$

$d$ is the bias, $d \in R^{n_r}$, and $M$ is relation representation matrix.

### 3.4 Optimization and implementation details

In this part, we will introduce the details of model learning and optimization. The loss function used in our model is cross-entropy loss, which is used to calculate the current overall loss of the training sample to promote further parameter updates of the model.The (11) is defined as:

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

Where $s$ is the set of training samples, $\theta$ is all the parameters in this model, including word-embedding matrix, position-embedding matrix, CNN weight matrix and relation-embedding matrix. To solve the optimization problem of the model, we use the Adadelta optimizer. Since Adadelta can adjust the learning rate adaptively, it helps to overcome this sensitivity to the choice of hyperparameters, and at the same time it also can avoid the continuous decline of the learning rate, which is very suitable for training deep networks.

In the implementation, we employ dropout on the output layer to prevent overfitting. The dropout layer is defined as an element-wise multiplication with a vector $h$ of Bernoulli random variables with probability $p$ [11]. Then (10) is rewritten as (12):

$$o = M(G^j \circ h) + d \tag{12}$$

In the test phase, the learnt set representations are scaled by $p$, $\hat{G^j} = pG^j$. And the scaled set vector $\hat{r_i}$ is finally used to predict relations.

## 4 Experiment

### 4.1 Dataset

Our experiment uses the New York Times (NYT) dataset, which was first released by [12], and then used as a standard dataset for distant supervision relation extraction. The dataset contains a total of 53 relations, including a special relation NA which indicated there was no relation between two entities, and the other 52 relations have specific meanings. The training set contains 522,611 sentences, 281,270 entity pairs, and 18,252 relational facts. The testing set contains 172,448 sentences, 96678 entity pairs, and 1950 relational facts, as shown in Table 2.

### 4.2 Evaluation metrics

Similar to previous work [11, 17, 22, 25, 26], we evaluate our model using precision/recall curves and Precision@N (P@N) in our experiments.

**Table 2** Distant supervision dataset NYT

| Dataset | Sentences | Entity pair | Relational facts |
|---|---|---|---|
| NYT-Train | 522611 | 281270 | 18252 |
| NYT-Test | 172448 | 96678 | 1950 |

### 4.3 Parameter settings

We use Word2vec tool [6] to embed the words as initial word embeddings. In the experiment, the word vector dimension is 50 dimensions, and the position feature vector is 5 dimensions. The detailed parameter settings are shown in Table 3. Since other parameters except Table 3 have relatively little influence on the model performance, we use [11] parameters in our paper. Group_size represents the TopN value selected by the similar bag attention mechanism, and Window size $l$ is the length of the sliding window of the convolutional layer. Sentence embedding size $d^c$ is the number of filters.

### 4.4 Comparative model analysis

#### 4.4.1 The influence of the number of sentences

In the original test dataset, there are 74,857 bags with only one sentence, accounting for almost 3/4 of all bags. Obviously, the more sentences in a bag, the more features it contains, at the same time, the more noisy sentences, so the number of sentences in the bag will have a certain impact on the performance of relation extraction. The name and method of the baseline model are briefly summarized in the following.

1. Related introduction of the baseline model: **Mintz** (Mintz et al. [9]): The method is the original distantly supervised approach to solve relation extraction problems with distantly supervised data.. **MultiR** (Hoffmann et al. [13]): The method is a graphical model within a multi-instance learning framework that is able to handle problems with overlapping relations. **MIML** (Surdeanu et al. [14]): The method is a multi-instance, multilabel learning framework that jointly models both multiple instances and multiple relations. **PCNN+ATT**The method employs a selective attention over multiple instances to alleviate the wrongly labeled problem, which is the principal baseline of our work. **PCNN+ATT+SL** (Liu et al. [19]): The method introduces an entitypair level denoising method, namely employing a soft label to alleviate the impact of wrongly labeled problem. **PCNN-HATT** (Han et al. [25]): The method employs hierarchical attention to exploit correlations among relations. **PCNN-BAG-ATT** (Ye et al. [23]): The method uses an intra-bag to deal with the noise at sentence-level and an inter-bag attention to deal with noise at the bag-level. **SeG** (Li et al. [11]): The method uses an Self-Attention Enhanced Selective Gate to overcome problem occurring in selective attention, which is caused by one-sentence bags.

Table 3 Model parameter settings on NYT

| Parameter | NYT |
|---|---|
| Word dimension $d_w$ | 50 |
| Position dimension $d_p$ | 5 |
| Batch size $B$ | 128 |
| Learning rate λ | 0.0001 |
| Dropout probability $p$ | 0.5 |
| Group_size | 5 |
| Window size $l$ | 3 |
| Sentence embedding size $d_c$ | 230 |

We will compare our experimental models with the baseline model and evaluate our experimental model in three ways: One, Two, and All.

- **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.

In the training process, we uses all sentences to train. In the test process, we will use the P@100, P@200, P@300 and the mean of them for each model as the evaluation criteria. In Table 4, the best performance of our model is compared with the best performance of other baseline models. P@N means the precision of the relation classification results with the topN highest probabilities in the test set.

Table 4 compares the results of P@N on the same dataset NYT of different models. The data of the first four methods are from the paper by [23].

It can be seen from Table 4 that the P@N value of the PCNN-PATT model on One, Two and All exceeds the baseline model PCNN-ATT, and under the condition of ALL, the mean value of P@N increases by 11.5% comparing PCNN-ATT model. It shows that the position attention mechanism can effectively reduce the influence of noisy words.

In the PCNN-PATT-SBA model, under the condition of One, the values of three items have achieved better performance than other models. under the condition of Two, the values of two items have achieved better performance than other models. under the conditions of ALL, the values of the four items far exceed the values of other methods. The Mean value of P@N is increased by 25.6% compared to the PCNN-ATT model, and 11.1% compared with the PCNN-PATT model. The result shows the effectiveness of the similar bag attention mechanism.

**Table 4** P@N values of entity pairs with different number of sentences

| Test Settings | One | | | | Two | | | | All | | | | Increase ratio |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| P@N(%) | 100 | 200 | 300 | mean | 100 | 200 | 300 | mean | 100 | 200 | 300 | mean | |
| PCNN-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 | 25.6 |
| PCNN-ATT-SL | 84.0 | 75.5 | 68.3 | 75.9 | 86.0 | 77.0 | 73.3 | 78.8 | 87.0 | 84.5 | 77.0 | 82.8 | 9.5 |
| PCNN-HATT | 84.0 | 76.0 | 69.7 | 76.6 | 85.0 | 76.0 | 72.7 | 77.9 | 88.0 | 79.5 | 75.3 | 80.9 | 12.1 |
| PCNN-BAG-ATT | 86.8 | 77.6 | 73.9 | 79.4 | **91.2** | 79.2 | 75.4 | 81.9 | 91.8 | 84.0 | 78.7 | 84.8 | 6.9 |
| SeG | **94.0** | **89.0** | **85.0** | **89.3** | 91.0 | **89.0** | **87.0** | **89.0** | **93.0** | 90.0 | 86.0 | 89.3 | 1.6 |
| PCNN-PATT | 84.0 | 79.5 | 74.6 | 79.4 | 86.0 | 78.5 | 75.0 | 79.8 | 88.0 | 82.5 | 74.0 | 81.6 | 11.1 |
| PCNN-PATT-SBA | 86.0 | 81.0 | 76.7 | 81.2 | 85.0 | 82.0 | 76.3 | 81.1 | 92.0 | **91.0** | **89.3** | **90.7** | |

Compared with PCNN-BAG-ATT, PCNN-PATT-SBA achieves better performance with three values under the condition of One. Under the conditions of Two, two values achieve better performance. under the condition of All, there are four values to achieve better performance, and the Mean value is 6.9% higher. The PCNN-BAG-ATT method uses the inter-bag attention between the same label bag to solve the problem of noise bags, but this method relies on the features of the same label bag.

SeG uses Self-Attention Enhanced Selective Gate to overcome problem occurring in selective attention, which is caused by one-sentence bags. However, this method uses the features of the sentence itself, but the features of the one-sentence bags are limited after all, so this method has certain limitations in some cases. The SBA method adopted in this paper is to merge the features of external bags to increase the features of the current bags. As long as the corresponding external bags are selected reasonably, the problem of too little information in a one-sentence bag can be solved. The model in this paper is lower than the SeG model on One and Two, because One and Two randomly select one and two sentences in a bag, which will lead to a significant reduction in the information of the current bag, and calculating the similarity between current bag and other bags, it will also affect the choice of similar sentence bags, and ultimately have a great impact on the performance of the model, so this model does not perform well on One and Two. On All, PCNN-PATT-SBA has a better effect than SeG with three values. The model proposed in this paper is 1.6 higher than SeG.

However, our model is based on the feature similarity between the bags to solve the problem of poor feature information in one-sentence bags, then merging similar feature bags. Our method does not rely on bags with the same label, so PCNN-PATT-SBA achieves better performance.

PCNN-ATT-SL uses soft label instead of hard label to solve the problem of incorrect label, but it fails to solve the problem of noisy bags and poor feature information in one-sentence bags. Compared with PCNN-ATT-SL, our model is on the Mean value of All increased by 9.5%.

PCNN-HATT uses hierarchical attention to obtain connections between relations, but it fails to solve the problem of noisy bags and poor feature information in one-sentence bags. Compared with PCNN-HATT, our model is on the Mean value of All increased by 12.1%.

In the test set setting, One and Two randomly select one and two sentences from each bag. This setting may avoid introducing noisy sentences, but the information contained in each bag is greatly reduced. Therefore, the performance of All selecting all sentences for testing is significantly better than that of One and Two.

The results of the comparison show that PCNN-PATT-SBA uses position attention combined with similar bag attention to perform more better.

#### 4.4.2 PR curve

In this part, we uses the PR curve to compare the performance of our model and the baseline model. In order to make a fair comparison with the baseline model, we also draws the top 2000 points of the prediction results of the test set to draw the PR curve. Figure 4 shows the PR curve comparison between the baseline model and our model PCNN-PATT-SBA.

It can be seen from Fig. 4: (1) When PCNN is used to encode sentences, the performance of PCNN-PATT exceeds that of PCNN-ATT. This is because the ATT method only considers to improve the performance of relation extraction by reducing the influence of noisy sentences, and does not consider the impact of noisy words. However, PATT first re-assigns weight to each word in the sentence through the position attention mechanism before ATT, thereby reducing the influence of noise words. (2) On the basis of PATT, we add SBA to the model to achieve a better effect. This result proves the effectiveness of the similar

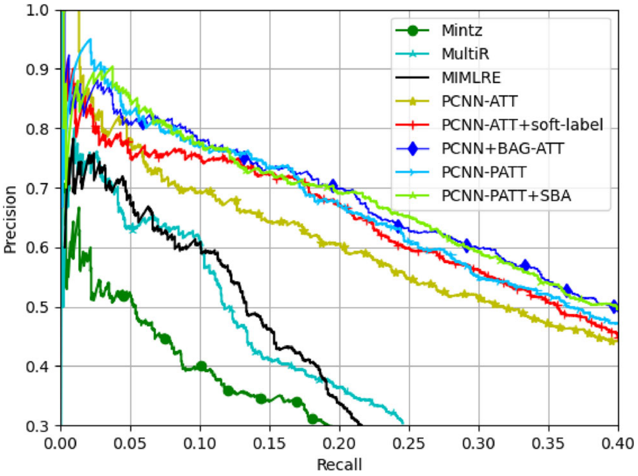


**Fig. 4** Performance comparison between our model and the baseline model

bag attention mechanism in distant supervision relation extraction. (3) The model performance of PCNN-PATT-SBA and PCNN+BAG-ATT did not differ significantly, but better results were achieved than other baseline models, indicating the advanced nature of the present model approach.

### 4.5 The influence of position attention

In this section, we will conduct a set of experiments to evaluate the influence of position attention on the performance of PCNN-PATT.

The position attention is determined by the position relationship sequence D between words and the Gaussian function G. D is determined by the sentence itself, and G obeys the $(\mu, \sigma^2)$ distribution. The weight of the position attention will change with the changes of $\mu$ and $\sigma$. However, the paper focuses on entity words and calculates the importance weights of non-entity words separately, so the paper sets $\mu = 0$. The influence of different $\sigma$ on the model PCNN-PATT performance is mainly compared, and the comparison results are shown in Table 5.

Table 5 compares the model performance under different variances. The variance determines the difference in weight between different words. The smaller the variance, the closer the word to the entity word has greater weight. Relatively speaking, the words that are farther from the entity word, the smaller the weight. We use $\mu = 0$ and $\sigma = \{0.5, 1, 1.5, 2\}$ as a comparative experiment. It can be seen from Table 5 that when $\sigma = 0.5$, the model performance achieves the best performance. Figure 5 compares the position attention PR curves of different $\sigma$. Adding position attention can significantly improve the performance of relation extraction. It can be seen from the Fig. 5 that the results of the four groups of experiments have a large difference. This is because different $\sigma$ directly affects the weight of the word. The larger the $\sigma$, the smaller the difference in the weight between words, the worse the noise reduction effect. The smaller the $\sigma$, the greater the difference between the weights of words, the better the noise reduction effect, but at the same time it will also increase the weight of the noise word, which affects the noise reduction effect. Figure 5 shows that the performance of $\sigma = 0.5$ and $\sigma = 1.5$ obviously exceed the performance of $\sigma = 1$ and $\sigma = 2$. When the recall is less than 0.25, the model with $\sigma = 0.5$ has the highest precision. When the recall is greater than 0.25, the model with $\sigma = 1.5$ has the highest precision. It can be seen from the comprehensive model precision and PR curve that the overall performance of $\sigma = 0.5$ is the best. Consequently, we use the $\mu = 0$ , $\sigma = 0.5$ parameter as the final parameter of position attention.

**Table 5** Model performance corresponding to different $\sigma$

| $\mu = 0$ | $\sigma$ | One | Two | ALL | Mean |
|---|---|---|---|---|---|
| | 0.5 | **88.0** | **82.5** | **74.0** | **81.6** |
| | 1 | 85.0 | 72.0 | 68.0 | 75.0 |
| | 1.5 | 80.0 | 76.0 | 72.6 | 76.2 |
| | 2 | 73.0 | 72.0 | 71.3 | 72.1 |

### 4.6 The influence of similar bag attention mechanism

Since there are a large number of one-sentence bag in the NYT dataset, the performance of relation extraction can be

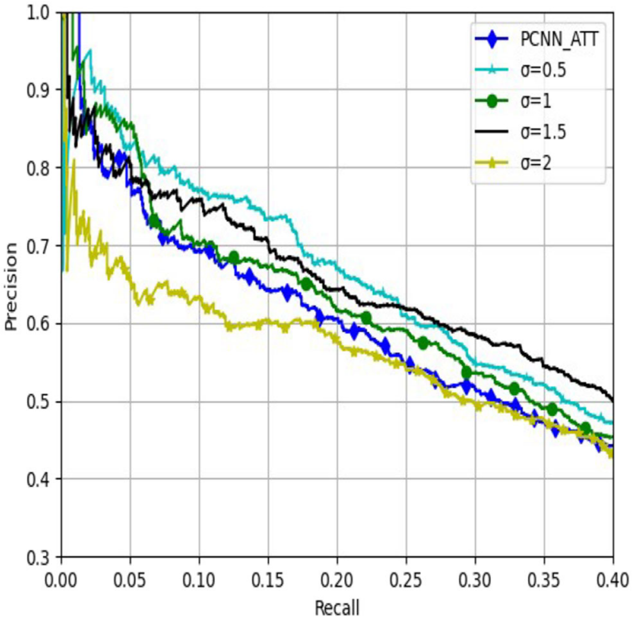


**Fig. 5** Model PR curves corresponding to different $\sigma$

improved by adding additional bag features, but it is not sure how many other bag features are needed to make the model perform best, so we aim to further experimental analysis the number of bag.

As shown in Fig. 6, we show the effect of different Group_size values in the PCNN-PATT-SBA model on the experimental performance. We use Group_size = {3, 4, 5, 6, 7, 8}, Batch_size = 128. It can be seen from Fig. 6 that when the value of Group_size increases, the value of the model on All increases accordingly. When Group_size=5, the model performance reaches the best, and then as Group_size increases, the model performance decreases. According to the results of Fig. 6, we can found that it is not necessarily merge more bag information, the performance of the model is more better, because merging other bag features will also increase the current bag more noise. Therefore, we need an suitable Group_size, which can not only increase the features of sentence bags but also reduce the introduction of noise.

### 4.7 The influence of different Batch_size values

Generally speaking, the size of Batch_size affects the optimization and learning speed of the model, but for our model, the value of Batch_size will also affect the performance of the similar bag attention mechanism. When calculating the similarity between bags, we need to calculated sequ entially all bags of Batch_size. The larger the value of Batch_size, the more bags with similar features may be included, and the performance of the similar bag attention mechanism is better, but at the same time it will also affect the overall performance of the model. Therefore, we further experimental analysis for different Batch_size values.

As shown in Table 6, we show the effect of different Batch_size values in the PCNN-PATT-SBA model on the experimental performance. We use Batch_size = {32, 64, 128, 256}, Group_size = 5. It can be seen from Table 6 that when the value of Batch_size increases, the value of the model on All increases accordingly, and then as Batch_size increases, the model performance decreases. When Batch_size = 128, the model performance reaches the best. According to the results of Table 6, we can found that it is not necessarily that the larger the value of Batch_size, the better the performance of the model. Although the larger the value of Batch_size, the better the performance of the similar bag attention mechanism, but the overall performance of the model is not necessarily the best.

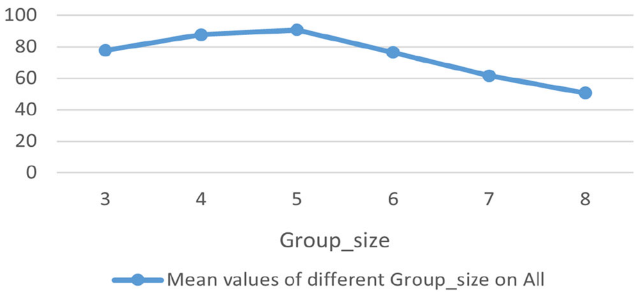


**Fig. 6** Mean values of different Group_size on All

**Table 6** The P@N of different Batch_size

| Test settings | | ALL | | |
|---|---|---|---|---|
| P@N(%) | 100 | 200 | 300 | Mean |
| Batch_size = 32 | 76.0 | 68.0 | 63.0 | 69.0 |
| Batch_size = 64 | 91.0 | 74.5 | 69.7 | 78.4 |
| Batch_size = 128 | **92.0** | **91.0** | **89.3** | **90.7** |
| Batch_size = 256 | 91.0 | 88.5 | 82.0 | 87.2 |

## 5 Summary and future work

In the paper, we propose a piecewise convolutional neural networks with position attention and similar bag attention for distant supervision relation extraction(PCNN-PATT-SBA). This model is used to deal with the problems of word-level noise and poor feature information in one-sentence bag in distant supervision relation extraction. First of all, according to the importance of each word in the sentence, we propose a position attention based on Gaussian distribution to model the importance of each word, and then assign corresponding weights to the words, improving the performance of relation extraction by reducing the weight of noise words. After that, we propose that similar bag attention is used to merge the features of similar bag, thereby increasing the features of the current bag. On the NYT dataset, our experimental results show that in most cases, the PCNN-PATT-SBA model is better than some of the most advanced baseline methods in distant supervision relation extraction.

In the future, our work will focus on the research of processing multi-label distant supervision relation extraction and the research of different attention mechanisms. We will mainly focus on the following aspects: (1) further improve the attention mechanism proposed in this paper. (2) design a new neural network model and different attention mechanisms for the task of multi-label distant supervision relations extraction.

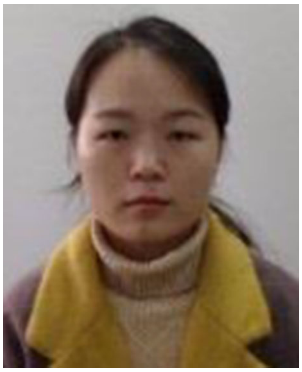
**Acknowledgements** This work is supported by the National Natural Science Foundation of China (62066022).

## References

1. Zelenko D, Aone C, Richardella A (2003) Kernel methods for relation extraction. J Mach Learn Res 3(6):1083–1106
2. Mooney RJ, Bunescu R. (2005) Subsequence kernels for relation extraction, pp 171–178
3. Sadeghi F, Divvala SK, Farhadi A (2015) Viske: Visual knowledge extraction and question answering by visual verification of relation phrases, pp 1456–1464
4. Ravichandran D, Hovy E (2002) Learning surface text patterns for a question answering system, pp 41–47
5. Yan Y, Okazaki N, Matsuo Y, Yang Z, Ishizuka M (2009) Unsupervised relation extraction by mining wikipedia texts using information from the web, pp 1021–1029
6. Zeng D, Liu K, Lai S, Zhou G, Zhao J (2014) Relation classification via convolutional deep neural network, pp 2335–2344
7. Santos CND, Xiang B, Zhou B (2015) Classifying relations by ranking with convolutional neural networks 1:626–634
8. Miwa M, Bansal M (2016) End-to-end relation extraction using lstms on sequences and tree structures 1:1105–1116
9. Mintz M, Bills S, Snow R, Jurafsky D (2009) Distant supervision for relation extraction without labeled data, pp 1003–1011
10. Bollacker K, Evans C, Paritosh P, Sturge T, Taylor J (2008) Freebase: a collaboratively created graph database for structuring human knowledge, pp 1247–1250
11. Lin Y, Shen S, Liu Z, Luan H, Sun M (2016) Neural relation extraction with selective attention over instances 1:2124–2133
12. Riedel S, Yao L, Mccallum A (2010) Modeling relations and their mentions without labeled text, pp 148–163
13. Hoffmann R, Zhang C, Ling X, Zettlemoyer L, Weld DS (2011) Knowledge-based weak supervision for information extraction of overlapping relations, pp 541–550
14. Surdeanu M, Tibshirani J, Nallapati R, Manning CD (2012) Multi-instance multi-label learning for relation extraction, pp 455–465
15. Lecun Y, Bengio Y, Hinton GE (2015) Deep learning. Nature 521(7553):436–444
16. Socher R, Huval B, Manning CD, Ng AY (2012) Semantic compositionality through recursive matrix-vector spaces, pp 1201–1211
17. Zeng D, Liu K, Chen Y, Zhao J (2015) Distant supervision for relation extraction via piecewise convolutional neural networks, pp 1753–1762
18. Ji G, Liu K, He S, Zhao J (2017) Distant supervision for relation extraction with sentence-level attention and entity descriptions, pp 3060–3066
19. Liu T, Wang K, Chang B, Sui Z (2017) A soft-label method for noise-tolerant distantly supervised relation extraction, pp 1790–1795
20. Yuan Y, Liu L, Tang S, Zhang Z, Zhuang Y, Pu S, Wu F, Ren X (2019) Cross-relation cross-bag attention for distantly-supervised relation extraction 33(01):419–426
21. Zhang Y, Zhong V, Chen D, Angeli G, Manning CD (2017) Position-aware attention and supervised data improve slot filling, pp 35–45
22. Li Y, Long G, Shen T, Zhou T, Yao L, Huo H, Jiang J Self-attention enhanced selective gate with entity-aware embedding for distantly supervised relation extraction, arXiv:Computation and Language
23. Ye Z, Ling Z (2019) Distant supervision relation extraction with intra-bag and inter-bag attentions, pp 2810–2819
24. Vilnis L, Mccallum A Word representations via gaussian embedding, arXiv: Computation and Language
25. Du J, Han J, Way A, Wan D (2018) Multi-level structured self-attentions for distantly supervised relation extraction, pp 2216–2225
26. Zhou P, Xu J, Qi Z, Bao H, Chen Z, Xu B (2018) Distant supervision for relation extraction with hierarchical selective attention, vol 108

**Publisher's note** Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

**Weijiang Li** was born in 1969. He is a professor. Supervisor at Kunming University of Science And Technology. His research interests include Natural language processing, etc. Yunnan Key Laboratory of Artificial Intelligence, Kunming University of Science and Technology.



**Qing Wang** was born in 1994. She is an M.S. candidate at Kunming University of Science And Technology. Her research interests include Natural language processing and relation extraction, etc. School of Information Engineering and Automation, Kunming University of Science and Technology.