IEICE TRANS. INF. & SYST., VOL.E104–D, NO.9 SEPTEMBER 2021

PAPER

Gated Convolutional Neural Networks with Sentence-Related Selection for Distantly Supervised Relation Extraction

Yufeng CHEN[†], ***Member***, **Siqi LI**[†], **Xingya LI**[††], **Jinan XU**[†], ***and*** **Jian LIU**[†a)], ***Nonmembers***

SUMMARY Relation extraction is one of the key basic tasks in natural language processing in which distant supervision is widely used for obtaining large-scale labeled data without expensive labor cost. However, the automatically generated data contains massive noise because of the wrong labeling problem in distant supervision. To address this problem, the existing research work mainly focuses on removing sentence-level noise with various sentence selection strategies, which however could be incompetent for disposing word-level noise. In this paper, we propose a novel neural framework considering both intra-sentence and inter-sentence relevance to deal with word-level and sentence-level noise from distant supervision, which is denoted as Sentence-Related Gated Piecewise Convolutional Neural Networks (SR-GPCNN). Specifically, 1) a gate mechanism with multi-head self-attention is adopted to reduce word-level noise inside sentences; 2) a soft-label strategy is utilized to alleviate wrong-labeling propagation problem; and 3) a sentence-related selection model is designed to filter sentence-level noise further. The extensive experimental results on NYT dataset demonstrate that our approach filters word-level and sentence-level noise effectively, thus significantly outperforms all the baseline models in terms of both AUC and top-n precision metrics.

key words: *relation extraction, distant supervision, gated convolutional neural networks, multi-head self-attention, soft-label, sentence-related selection*

1. Introduction

Relation extraction (RE), aiming to describe the relationship between two specific entities by mining semantic associations in natural language texts, is a vital task to build knowledge bases (KB). Typically, supervised methods for RE require large-scale labeled corpus, which usually need to be manually annotated by experts in related fields. However, the annotating process is too time-consuming and laborious. To address this, Mintz et al. [1] proposed a distant supervision method to obtain a large-scale labeled training set automatically, which assumes that "any sentence containing the two entities in the external corpus reflects their relationship defined in KB". For example, given a triplet in a knowledge base, also known as a relational fact, (*"Jackie Chan"*, **/place_of_birth/**, *"Hong Kong"*), all sentences containing the above named entity pair will be labeled as relation **/place_of_birth/**.

Manuscript received November 26, 2020.
Manuscript revised April 3, 2021.
Manuscript publicized June 1, 2021.

[†]The authors are with Beijing Key Lab of Traffic Data Analysis and Mining, the School of Computer and Information Technology, Beijing Jiaotong University, Beijing, 100044 China.
[††]The author is with China Institute of Marine Technology & Economy, Beijing, 100081 China.
a) E-mail: jianliu@bjtu.edu.cn

DOI: 10.1587/transinf.2020EDP7249

Although distant supervision is a fast and effective means of obtaining training data automatically, it is plagued by the wrong labeling problem, since the same two entities in different sentences with various contexts cannot express a consistent relationship as described in a known KB. For example, in the sentence *"The action star Jackie Chan is in Hong Kong to promote his new film."*, there is no **/place_of_birth/** relation between *"Jackie Chan"* and *"Hong Kong"*, but it would still be regarded as a positive instance. An effective solution to reducing such false positive is multi-instance learning (MIL) methods [2], which relax the strong distant supervision assumption to expressed-at-least-one assumption. This means that any possible relation between two entities holds true in at least one sentence rather than all sentences containing those two entities. In MIL, sentences with the same entity pairs are integrated into one bag, to which only one label is assigned. Thus it is expected that sentences with incorrect labels are assigned less weight and the bag is represented mainly based on sentences with correct labels. Recently, the existing MIL studies on distant supervision are mainly divided into two categories: the first category focuses on weight distribution of sentences in the same bag [2]–[7]. The other category introduces extra knowledge to help models obtain more valid features [8], [9].

The above researches have achieved good results in reducing the impact of noisy sentences with wrong labels. However, they all ignored the negative influence of noisy words in predicting relations, since not all words in a sentence contribute to judging relation labels. For example, in a sentence *"Ms. Bryant was born in New York, and moved to Connecticut when she was a young child"*, where *"Ms. Bryant"* and *"Connecticut"* are two corresponding entities, the sentence describes the **/place_lived/** relation between *"Ms. Bryant"* and *"Connecticut"*, but the sub-sentence *"was born in New York"* has little effect or even negative effect on judging the relation **/place_lived/**, which could be regarded as noisy words or word-level noise. Features based on those noisy words will cut down the precision of the relation extraction model. For a popular distantly supervised relation extraction benchmark, e.g., NYT-10 dataset, there are about 12 noisy words in each sentence on average [10]. To address this problem, we propose a novel neural network to filter both word-level and sentence-level noise, in which a gate mechanism [11] is used to screen out the features extracted by the convolution layer. Moreover, it's well known that multi-head self-attention (MSA) mechanism originated

from Transformer model [12] is widely used in machine translation tasks to assign weights dynamically to different words based on their correlation with each other, filtering out the negative effects of irrelevant words. Thus we are motivated to introduce MSA mechanism after the convolution layer to filter word-level noise. Also, inspired by TransE model [13], we introduce a soft-label strategy in our framework to make full use of prior knowledge from a KB and reduce wrong-labeling propagation. Other than TransE model, we use bilinear transformation instead of linear computation to capture deeper relevance between the head entity and the tail entity.

On the other hand, most of the existed work deals with sentences in one bag independently. However, sentences in the same bag are more or less related, for example, in Table 1, sentence s1 clearly expresses the relation **/contains/**, while sentences s2 and s3 imply the relation **/contains/** between "*Latin America*" and "*Venezuela*" at semantic level, which is not easy to be extracted by models if sentences are treated independently. Therefore, we design a sentence-related selection method which models the relevance among sentences in the same bag to obtain more valid features with less sentence-level noise.

Combining the above aspects, we propose a novel framework for distantly supervised relation extraction with an MSA based gate mechanism and a sentence-related selection model, which is named as **S**entence-**R**elated **G**ated **P**iecewise **C**onvolutional **N**eural **N**etworks (**SR-GPCNN**). We test on the public data sets and compare the framework with baselines and competitive approaches, and the experimental results show that the performance of our proposed model is significantly better than that of all baseline systems, which verifies the effectiveness of the proposed approach. Our main contributions are summarized as follows:

1) To handle the problem of noisy words, a gate convolution with multi-head self-attention mechanism is proposed to filter word-level noisy features. Moreover, a soft-label strategy is utilized to further weaken the impact of wrong labeling problem caused by distant supervision.

2) Most of the existing methods regard sentences as independent individuals when assigning weights to the sentences. In order to make full use of the relevance of sentences in the same bags, we propose a sentence-related selection method, which utilizes semantic similarity calculation method to model the relevance of the sentences, and assigns weights to sentences according to the relevance, thus obtains better bag-level features to improve the final performance.

3) The proposed framework achieves a new state-of-the-art performance in terms of AUC for distantly supervised relation extraction. Furthermore, the gate mechanism could be adopted by other neural networks and enhance the performance of the corresponding tasks.

Table 1 An example of a bag in NYT-10 dataset.

Entities and Relation	Sentences
"*Latin America*" "*Venezuela*" **/contains/**	s1. ***Venezuela*** *has the fastest-growing economy in* ***Latin America****, with growth rates in the first two quarters of 7.5 percent and 11 percent, respectively.*
	s2. … *Chinese interest in* ***Venezuela****, a senior committee aide said, underlines Washington's lack of attention toward* ***Latin America****.*
	s3. *Mr. Chávez's rise to international prominence in* ***Latin America*** *and beyond has much to do with his supremacy in* ***Venezuela****.*

2. Proposed Approach

As illustrated above, we propose a novel SR-GPCNN framework for distantly supervised relation extraction, which is illustrated in Fig. 1.

Our model mainly consists of four parts:

(1) Vector Representation: each sentence in a given bag is fed into a Vector Representation Layer to generate its corresponding vector representation;

(2) MSA-based GPCNN: it is proposed to generate valid sentence-level features. Based on sentences' vector representations, MSA is adopted to reduce negative effect of noisy words within sentences, and soft-label strategy is used to make full use of prior knowledge from a KB and reduce wrong-labeling propagation;

(3) Sentence-Related Selection: this Layer generates the bag-level vector representation from features after dimension reduction;

(4) Classifier: it is used to predict the final relation labels.

Compared with the traditional PCNNs model, we improve the convolution layer by adding a gate mechanism and multi-head self-attention layer to reduce word-level noise. Furthermore, we replace the relation labels with a soft-label strategy in the training process to weaken the impact of incorrect labels. In addition, we adopt a sentence-related selection method to reduce sentence-level noise. We will

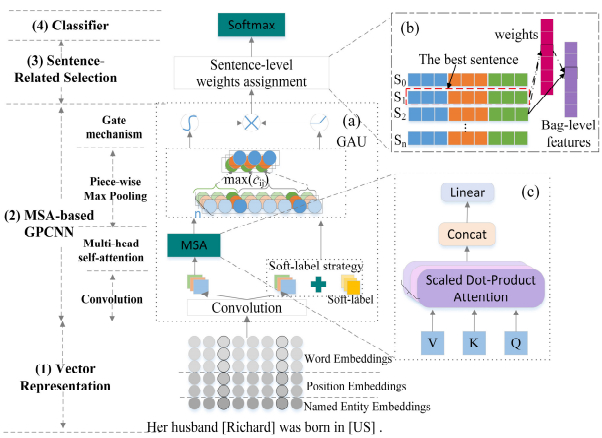


Fig. 1 The architecture of our approach (SR-GPCNN) to distantly supervised relation extraction. The part (a) describes the gated convolution layer, which takes soft labels as supervised information. The part (b) illustrates the modeling of sentence-related selection. And the part (c) describes the structure of multi-head self-attention mechanism.

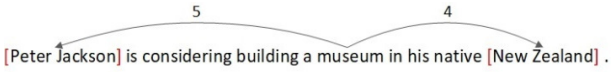


Fig. 2 An example of relative distances

Word	On	July	28	,	Bill	Gates	sold	2	million	Micrsoft	shares
Tag	O	DATE-B	DATE-I	O	PER-B	PRE-I	O	O	O	ORG-B	O

Fig. 3 An example of BIO tag for named entity embeddings

describe the MSA based gate mechanism and the sentence-related selection method in details in Sects. 2.2 and 2.3 respectively.

2.1 Vector Representation

The input of our proposed model is represented by embeddings, which are composed of three parts: word embeddings, position embeddings and Named Entity embeddings.

- **Word Embeddings**

Word embeddings are distributed representation of words, aiming at mapping each word into a low-dimensional vector. The vector is obtained by looking up a pre-trained embedding matrix $\boldsymbol{V} \in \mathbb{R}^{|V| \times d_w}$ (or lookup table), where $|V|$ is the vocabulary size and d_w is the dimension of word embeddings.

- **Position Embeddings**

Following Zeng et al. [2], we employ position features to track the relative distances of the current word to the head entity e_1 and the tail entity e_2. Figure 2 shows an example of the relative distances. The relative distances from word *"museum"* to *"Peter Jackson"* (e_1) and *"New Zealand"* (e_2) are 5 and −4 respectively. We can transfer the two relative distance to real-value vectors $\boldsymbol{PF}_1$ and $\boldsymbol{PF}_2$ by looking up in a randomly initialized position embedding matrix, where $\boldsymbol{PF}_i \in \mathbb{R}^{d_p}$ and d_p is the dimension of position embeddings.

- **Named Entity Embeddings**

To enrich the representation of the input sentence with different types of named entity words, we introduced BIO tag information as named entity embeddings. If the entity type is T, we use the label "T-B" to mark the start of the entity, and the label "T-I" to mark the rest of the entity, the word that is not any part of an entity is marked as "O". Figure 3 shows an example of the BIO tag. Given an input sentence s, we can transfer its named entity label to real-value vectors $\boldsymbol{NE} \in \mathbb{R}^{d_t}$, where d_t is the dimension of named entity embeddings.

Finally, the input representation of a word is a vector concatenated by word embeddings, position embeddings and named entity embeddings. With the vector representation of words, we transfer the sentence s into a matrix $\boldsymbol{S} \in \mathbb{R}^{|s| \times d}$, where $|s|$ is the length of s, and the dimension of each word is $d = d_w + d_p \times 2 + d_t$.

2.2 MSA Based Gate Mechanism

- **Convolutional Neural Network**

Convolutional neural networks (CNNs) can effectively extract all local features of the input and perform global predictions. Specifically, convolution is an operation between a weight matrix (also known as filter) $\boldsymbol{W}_c \in \mathbb{R}^{w \times d}$ and the vector matrix $\boldsymbol{S} = \{\boldsymbol{q}_1, \boldsymbol{q}_2, \ldots, \boldsymbol{q}_{|s|}\}$ of a sentence s, where w is the window size. Let $\boldsymbol{q}_{i:j}$ refer to the concatenation of $\boldsymbol{q}_i$ to $\boldsymbol{q}_j$. The result of convolution operation is $\boldsymbol{c} = \{c_1, c_2, \ldots, c_{|s|-w+1}\}$, and we can obtain c_i by:

$$c_i = \boldsymbol{W}_c \cdot \boldsymbol{q}_{(i-w+1):i} + b_c \tag{1}$$

where $1 \leq i \leq |s| - w + 1$, and b_c is bias.

- **Gate Mechanism**

Considering the negative impact of noisy words, we utilize a gate mechanism to select positive features at word level. The gate mechanism has shown effectiveness in language modeling [14], [15]. We improve the gate mechanism based on GTU (Gated Tanh Units) and name it as GAU (Gated Activation Units, as shown in Fig. 1 (a)), which is represented by:

$$g_i = \text{relu}(\boldsymbol{W}_{GAU} \cdot \boldsymbol{q}_{(i-w+1):i} + b_{GAU}) \tag{2}$$

where $\boldsymbol{W}_{GAU} \in \mathbb{R}^{w \times d}$ is a weight matrix, and b_{GAU} is a bias. The relu gates control features extracted by the tanh units according to its own outputs to achieve the purpose of selecting the important word-level features.

- **Soft-Label Strategy**

Generally, the relation labels of entity pairs are unchangeable during training, leaving out whether they are true or not, which would enlarge the negative impact of the wrong labeling problem on the feature selection process. For this, we introduce a soft-label strategy into GAU to weaken the impact of wrong labels on the model performance, i.e., we replace hard labels with soft labels generated from the entity pairs to guide feature selection and cut down word-level noise during training.

As shown in Fig. 1, GAU is connected to two convolutional networks (one is the original CNN and the other has label features). We use the bilinear transformation:

$$L_{relation} = \boldsymbol{e}_1{}^T \boldsymbol{W}_l \boldsymbol{e}_2 \tag{3}$$

as the soft label between two entities (e_1, e_2) to help select important features. $\boldsymbol{e}_1, \boldsymbol{e}_2 \in \mathbb{R}^{d_w}$ are the word vectors of two entities respectively, $\boldsymbol{W}_l \in \mathbb{R}^{d_w \times d_w}$ is the parameter matrix. We can get the second convolution result with soft label features $\boldsymbol{y} = \{y_1, y_2, \ldots, y_{|s|-w+1}\}$, and we can obtain y_i by:

$$y_i = (\boldsymbol{W}_g \cdot \boldsymbol{q}_{(i-w+1):i} + b_g) + L_{relation} \tag{4}$$

where $\boldsymbol{W}_g \in \mathbb{R}^{w \times d}$ is a parameter matrix, b_g is a bias.

The ability to capture different features typically requires the use of multiple filters in the convolution, so we

adopt n filters $\boldsymbol{W} = \{(\boldsymbol{w}_c^1, \boldsymbol{w}_g^1), (\boldsymbol{w}_c^2, \boldsymbol{w}_g^2), \ldots, (\boldsymbol{w}_c^n, \boldsymbol{w}_g^n)\}$. The final original convolution output is $\boldsymbol{C} = \{\boldsymbol{c}^1, \boldsymbol{c}^2, \ldots, \boldsymbol{c}^n\}$ and the output of soft-label convolution is $\boldsymbol{Y} = \{\boldsymbol{y}^1, \boldsymbol{y}^2, \ldots, \boldsymbol{y}^n\}$.

• **Multi-head Self Attention Mechanism**

Traditional methods often directly use distant supervision labels as the basis of attention calculation, without considering the influence of wrong relation labels during word weight distribution. To deal with this issue, the multi-head self-attention mechanism is introduced here to filter the word-level noise information, which does not need relation labels. The MSA assigns high weights to related words and low weights to irrelevant words by calculating the correlation among words.

The above feature $\boldsymbol{C}$ obtained by the convolutional layer is processed by three different linear transformations to obtain three matrices, namely $\boldsymbol{Q}$, $\boldsymbol{K}$ and $\boldsymbol{V}$. Then the sentence after the single self-attention is obtained through the scaled doc-product attention calculation. The calculation of scaled doc-product attention is:

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{5}$$

where d_k indicates the dimension of $\boldsymbol{K}$.

After many times of attention calculation, the result is concatenated to get the final sentence expression:

$$\mathbf{head}_\text{i} = \text{Attention}(\boldsymbol{C}\boldsymbol{W}_i^q, \boldsymbol{C}\boldsymbol{W}_i^k, \boldsymbol{C}\boldsymbol{W}_i^v) \tag{6}$$

$$\boldsymbol{H} = \text{Concat}(\mathbf{head}_1, \mathbf{head}_2, \ldots, \mathbf{head}_h)\boldsymbol{W}_{sa} \tag{7}$$

Where $1 \le i \le h$, h is the number of the heads, $\boldsymbol{W}_i^q$, $\boldsymbol{W}_i^k$, $\boldsymbol{W}_i^v \in \mathbb{R}^{n\times n}$ and $\boldsymbol{W}_{sa} \in \mathbb{R}^{n\times n}$ are parameter matrices.

• **Piecewise Max Pooling**

Max pooling operation is usually used to extract the most dominant features in feature maps, but it ignores the structure information and fine-grained information. Thus, PCNNs divide an instance into three segments according to the given entity pair and conduct max pooling operation on each segment. For each multi-head self-attention result $\boldsymbol{H}$, it can be divided into three parts:

$$\boldsymbol{H} = \{\boldsymbol{H}_{i,1}, \boldsymbol{H}_{i,2}, \boldsymbol{H}_{i,3}\} \tag{8}$$

then the piecewise max pooling process is defined as:

$$\boldsymbol{p}_i = \{p_{i,1}, p_{i,2}, p_{i,3}\} \tag{9}$$

$$p_{i,j} = \max(\boldsymbol{H}_{i,j}) \tag{10}$$

where $1 \le i \le n$, where $j \in \{1, 2, 3\}$. We use the same piecewise pooling operation on the soft-label convolution result $\boldsymbol{Y}$ to get $\boldsymbol{G} \in \mathbb{R}^{3n}$. Then, we apply a non-linear function tanh at the output $\boldsymbol{p}$, and get $\boldsymbol{P} \in \mathbb{R}^{3n}$.

According to Eq. (2) and Eq. (4), we can get the gated value $\boldsymbol{g}$ by:

$$\boldsymbol{g} = \text{relu}(\boldsymbol{G}) \tag{11}$$

Then we can get the final sentence representation, which is calculated by the gated convolution:

$$\boldsymbol{s} = \boldsymbol{g} * \boldsymbol{P} \tag{12}$$

2.3 Sentence-Related Selection and Output

In previous studies, bilinear and nonlinear attention mechanisms were often used to assign weights to different sentences, thus reducing the sentence-level noise. However, they all neglected the relevance among sentences. Considering preserving the characteristics of sentences and computational efficiency, we adopt cosine similarity to calculate the relevance among sentences. Specifically, given a bag $B_i = \{s_1, s_2, \ldots, s_m\}$, we recognize the instance with the highest probability as $\boldsymbol{s}_{max}$, and assign high weights to sentences with high similarity to $\boldsymbol{s}_{max}$, otherwise, assign low weights, the weight β_j^i is calculated as follows:

$$\beta_j^i = \frac{\exp(\varphi^j)}{\sum_m \exp(\varphi^m)} \tag{13}$$

$$\varphi^j = \text{Cos}_\text{Similarly}(\boldsymbol{s}_{max}, \boldsymbol{s}_j) = \frac{\boldsymbol{s}_{max} \cdot \boldsymbol{s}_j}{\|\boldsymbol{s}_{max}\| \, \|\boldsymbol{s}_j\|} \tag{14}$$

where $1 \le j \le m$, $\boldsymbol{s}_{max}$ and $\boldsymbol{s}_j$ are the corresponding sentence-level features obtained by neural network. Then the bag-level features $\boldsymbol{b_i} \in \mathbb{R}^{3n}$ are:

$$\boldsymbol{b_i} = \sum\nolimits_j \beta_j^i \boldsymbol{s}_j \tag{15}$$

Although the above method can effectively model sentences' relevance, the features obtained from other sentences (according to the similarity with $\boldsymbol{s}_{max}$) inevitably introduce noise features because of irrelevant features in $\boldsymbol{s}_{max}$. We assume that the effective features have been highlighted during training. Therefore, instead of getting the bag representation with weighted sum as Eq. (15), we select the maximum features of each dimension from all the sentences, and then integrate them as bag-level features, so as to ensure that effective features can be obtained as much as possible. The specific definition of the bag-level features $\boldsymbol{b}_i$ is as follows:

$$\boldsymbol{b}_i = \left[b_i^1, b_i^2, \ldots, b_i^{3n}\right] \tag{16}$$

$$b_i^j = \max\left(\left[\beta_1^i s_1^j, \beta_2^i s_2^j, \ldots, \beta_m^i s_m^j\right]\right) \tag{17}$$

where the s_k^j denotes the j-th dimension of the k-th sentence. The vector representation $\boldsymbol{b}_i$ of bag B_i is then fed to the softmax classifier to predict the final relation labels as follows:

$$p(r \mid \boldsymbol{b}_i, \theta) = \frac{\exp(o_r)}{\sum_{k=1}^{n_r} \exp(o_k)} \tag{18}$$

where n_r is the number of relations and $\boldsymbol{o}$ is the final output which corresponds to all relation types, which is defined as:

$$\boldsymbol{o} = \boldsymbol{W}_o \boldsymbol{b}_i + \boldsymbol{b}_o \tag{19}$$

where $\boldsymbol{W}_o \in \mathbb{R}^{n_r \times 3n}$ is the relation matrix and $\boldsymbol{b}_o \in \mathbb{R}^{n_r}$ is a bias, and the cross-entropy objective function on all training bags (T) is calculated as:

$$J(\theta) = \sum_{i=1}^{T} \log p(r_i \mid \boldsymbol{b}_i; \theta) \tag{20}$$

3. Experiments and Discussions

3.1 Datasets and Evaluation Metrics

To evaluate our proposed framework, we conduct experiments on a widely used dataset which is developed by Riedel et al. [16]. This dataset is generated by aligning the relations in Freebase with the New York Times corpus (NYT). We use aligned sentences from 2005 to 2006 as training data and sentences from 2007 as testing data. The dataset has 53 kinds of relation labels including label NA which means that there is no relation between entity pairs. The training data includes 570,088 sentences, 281,270 entity pairs and 18,252 relation facts. The testing data includes 172,448 sentences, 96,678 entity pairs and 1,950 relational facts.

We use the held-out evaluation to evaluate our model. It provides an approximate precision measurement method without time-consuming manual evaluation by comparing the relation instances extracted from bags against Freebase relations data automatically. We will use the aggregated precision-recall (P-R) curves, area under curve (AUC) and top-N precision (Precision@N) as metrics in our experiments.

3.2 Parameter Settings

In our experiments, we use the word2vec tool [17] to pretrain the word embeddings on NYT corpus. We tune all of the models using three-fold validation on the training set. We select the dimension of word embeddings d_w among {50, 100, 200, 300}, the dimension of position embeddings d_p among {5, 10, 20}, the window size w among {3, 5, 7}, the number of filters n among {50, 100, 230, 300}, the batch size among {50, 100, 160}, the learning rate λ among {0.001, 0.01, 0.1, 0.2, 0.5}. The best configurations are: $d_w = 50$, $d_p = 5$, $w = 3$, $n = 230$, $\lambda = 0.2$, the batch size is 160 and the number of heads h is 5. We use dropout strategy and Adadelta to train our models. According to experience, the dropout rate is fixed to 0.5.

3.3 Relation Extraction Performance

We compare our approach with extensive previous work, including feature-engineering, competitive and state-of-the-art models, which are listed in the following.

- **PCNN+ATT** [5] employed a selective attention mechanism to combine sentence features for each bag, and is trained under the PCNN model.
- **APCNN+D** [3] adopted sentence-level attention to select multiple valid instances in a bag, and introduced entity descriptions to improve entity representations.
- **RESIDE** [9] employed a graph convolution neural network to encode sentence dependency information and utilized additional side information from KBs to improve relation extraction.
- **Non-IID** [6] utilized a linear attenuation simulation and non-IID (non-independent and non-identically distributed) embeddings to increase valid instances.
- **SeG** [4] introduced an entity-aware embedding module and a self-attention enhanced selective gate mechanism to deal with mislabeled data.

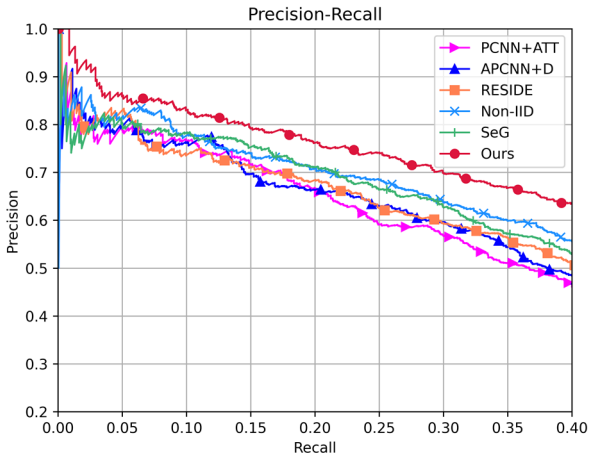


Fig. 4 Aggregate precision/recall curves for our approach and all the baseline models.

Table 2 AUC scores of our approach and all the baseline models.

Model	AUC
PCNN+ATT [5]	0.38
APCNN+D [3]	0.40
RESIDE [9]	0.42
Non-IID [6]	0.44
SeG [4]	0.42
Ours	**0.51**

Table 3 Precision@N of our approach and all the baseline models.

Precision@N(%)	Top 100	Top 200	Top 300	Average
PCNN+ATT [5]	78.21	77.61	74.09	76.64
APCNN+D [3]	78.22	75.62	77.41	77.08
RESIDE [9]	84.00	78.52	75.63	79.41
Non-IID [6]	81.19	81.09	76.41	79.56
SeG [4]	81.18	78.60	77.06	78.95
Ours	**86.13**	**84.57**	**81.39**	**84.03**

Figure 4 shows the aggregated P-R curves, Table 2 shows the AUC scores and Table 3 shows the Precision@N with N = {100, 200, 300} of our approach and all the baselines.

For the sake of clarity, all the curves are showed with different colors and bold lines in Fig. 4. From Fig. 4, Table 2 and Table 3 we can observe that our proposed approach consistently and significantly outperforms all the baselines. Specifically, the precision of our model decreases much slower than that of other baseline models along with higher recall rate, where the whole curve of our proposed model is much smoother. And its improvement on Precision@N is very significant, which is about 6% higher than baselines on average. Also, the empirical results of AUC are coherent with those of Precision@N, which shows that our

proposed approach can significantly improve previous ones and reach a new state-of-the-art performance by filtering both of the word-level and the sentence-level noise. This indicates that our proposed SR-GPCNN is effective because the MSA-based gate mechanism and the sentence-related selection method can select more important word-level features at fine-grained level and sentence relevance can be very useful for sentence weights assignment.

3.4 Ablation Study

In order to verify the effectiveness of each module in the proposed framework, we conduct an extensive ablation study in this section.

SR-GPCNN w/o Gate: denotes removing the gate mechanism introduced in Eq. (2), to verify the effectiveness of the proposed gate mechanism in word-level feature selection;

SR-GPCNN w/o SL: denotes removing the soft-label information in our proposed GAU unit, and adopting bag labels as the supervision signal of the gate mechanism.

SR-GPCNN+CNN-SL w/o Gate: denotes removing the gate mechanism introduced in Eq. (2), but adding extra soft-label information in the convolution part (Eq. (1)).

SR-GPCNN w/o MSA: denotes removing the multi-head self-attention mechanism before the pool layer of GPCNN;

SR-GPCNN-Att: denotes replacing the multi-head self-attention mechanism with the general attention mechanism to explore the influence of attention mechanism further.

As for the sentence selection module, we also set up the following comparative experiments to verify the effectiveness of the proposed sentence-related selection method. First we adopt GPCNN as the baseline to extract sentence-level features, and then use different sentence selection methods to obtain bag-level features, including: 1) **MIL-GPCNN** model selects the sentence with the highest probability as the bag-level feature; 2) **ATT-GPCNN** model uses the attention mechanism at the sentence level, and 3) **SR-GPCNN** represents our proposed sentence-related modeling method. By setting up all the above comparative experiments, the corresponding AUC scores and PR curves are shown in Table 4 and Fig. 5.

3.4.1 Effectiveness of Gate Mechanism

From Table 4 and Fig. 5, we can see:

1) The performance of SR-GPCNN w/o Gate declines significantly after removing the gate mechanism, indicating that GAU module can effectively improve the performance of the model, verifying the effectiveness of the gate mechanism and reflecting the robustness of the GAU module.
2) The performance of SR-GPCNN is significantly improved compared with SR-GPCNN w/o Gate, which demonstrates that the word-level noise has a great impact on the performance of the relation extraction models, and also indicates the GAU module can effectively filter word-level noise by obtaining more important features.

Table 4 Ablation study regarding the AUC value

Model	AUC
SR-GPCNN w/o Gate	0.41
SR-GPCNN w/o SL	0.42
SR-GPCNN+CNN-SL w/o Gate	0.47
SR-GPCNN w/o MSA	0.49
SR-GPCNN-Att	0.40
MIL-GPCNN	0.46
ATT-GPCNN	0.47
SR-GPCNN	**0.51**

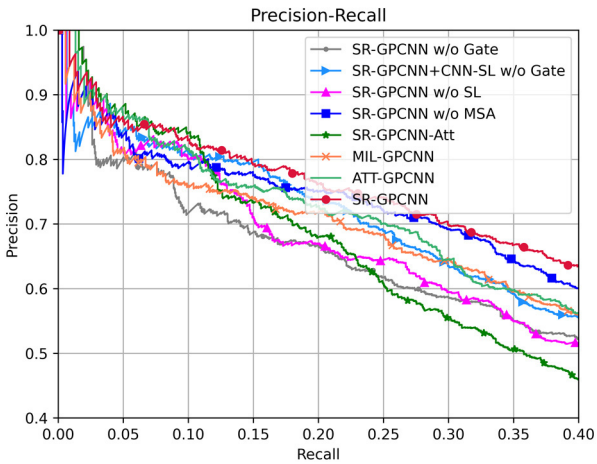


Fig. 5 Performance comparison for ablation study regarding precision/recall curves.

3.4.2 Effectiveness of Soft-Label Strategy

From Fig. 5, we observe a notable performance drop after removing the soft label mechanism, which proves the effectiveness of the soft-label strategy. Moreover, the curve indicates that the soft-label strategy not only improves the precision rate under high recall rate, but also greatly improves the accuracy rate of the model when the recall rate is low, which indicates that the soft label mechanism plays a positive role in the gate mechanism. In addition, in the case of removing the gate mechanism, the performance can also be greatly improved after adding additional soft label information at the CNN layer.

3.4.3 Effectiveness of Multi-Head Self-Attention Mechanism

It can be seen from Table 4 and Fig. 5 that the performance of SR-GPCNN w/o MSA declines after removing the multi-head self-attention mechanism, and there is a notable performance drop after replacing the MSA mechanism with normal attention mechanism, which shows that the multi-head self-attention mechanism is superior in filtering out the negative effects of irrelevant words and can further filter noisy features in sentences.

Table 5 An example of attention weights distribution in different models.

Triplet	Instances	APCNN+D	Ours
"*Richard Epstein*", /**company**/ "*University of Chicago*"	1. *"You got a very strong, visceral response," said **Richard Epstein**, a law professor at the **University of Chicago**, "because this is something on which everyone is a constitutional expert."*	0.094	0.341
	2. *"The Kelo decision wasn't compelled by legal precedents," says **Richard Epstein**, a law professor at the **University of Chicago**."*	0.814	0.328
	3. *"We changed from a court split 4 to 3, with two in the middle," said **Richard Epstein**, a law professor at the **University of Chicago**, referring to the dual swing votes of justices O'Connor and Kennedy."*	0.092	0.331

3.4.4 Effectiveness of Sentence-Related Selection

From Table 4 and Fig. 5, we can see that our sentence-related selection method (SR-GPCNN) has achieved the best performance on the premise of the same feature extractor, which verifies its effectiveness of modeling the relevance of sentences. In previous MIL and ATT methods, sentences are regarded as independent individuals, ignoring the relevance among sentences. Compared with them, the sentence-related selection captures the relevance of sentences at the semantic level, and thus assigns more reasonable weights to each sentence in a bag.

3.5 Case Study

To explicitly explain the negative influence of noisy words on feature selection and the effectiveness of our proposed approach, we illustrate an example of attention weights distribution in a bag during testing. As shown in Table 5, the bold strings are head/tail entities and the underlined strings are keywords to predict the relation. The relation **/company/** corresponds to **/business/person/company/** in Freebase. all sentences contain the phrase *"a law professor at"*, which clearly indicates the **/company/** relation between the entities *"Richard Epstein"* and *"University of Chicago"*, so they are all positive instances.

It can be find out that, the phrase *"a law professor at"* in sentences 1∼3 has direct connection with the entity pair relation **/company/**, while other words rarely imply this relation, which can be denoted as noisy words. According to the length of those sentences, sentence 2 might contain the least amount of noisy information, sentence 1 contains the second, and sentence 3 contains the most. We adopt **APCNN+D** [8] model for comparison, which utilized sentence-level attention. Due to the lack of effective intra-sentence (word-level) noise filter, APCNN+D is inevitably interfered by noisy words and thus assigns higher weight to sentence 2 (0.814) and lesser weights to sentence 1 (0.094) and sentence 3 (0.092). In contrast, our SR-GPCNN framework is barely affected by noisy information, which is proposed to filter word-level noisy features, and thus assigns similar weights to the three sentences (0.341, 0.328, 0.331). The attention weights comparisons verify that the proposed gate mechanism with MSA and soft-label strategy can effectively filter word-level noise and select more important intra-sentence features to make full use of the supervision information in a bag.

Table 6 An example of attention weights distribution of our approach with sentence-related selection.

Triplet	Instances	Relevance	Weight
"*Boston*", /**contains**/ "*Fenway_Park*",	1. *yet two of the most popular parks remain two of the most historic : Yankee Stadium in the Bronx and **Fenway_Park** in **Boston** .*	1	0.26
	2. *Yankees manager Joe Torre has said that it is unlikely clemens will start in the three-game weekend series against **Boston** at **Fenway_park** .*	0.49	0.15
	3. *some time after midnight near **Fenway_Park** in **Boston** , where myers was scheduled to start the next game for the phillies ...*	0.85	0.22
	4. *of retaliation protocol , pettitte recalled hitting **Boston** 's kevin youkilis on friday night at **Fenway_Park** in his last start .*	0.52	0.16
	5. *... at **Boston**'s **Fenway_Park**, the green monster is a tall, hard wall in left field .*	0.69	0.19

On the other hand, we show another example of attention weights distribution in a bag during training, so as to better explain the weights distribution process of the sentence-related selection.

As shown in Table 6, the bold strings are head/tail entities. The relation **/contains/** corresponds to **/location/location/contains/** in Freebase. All sentences in Table 6 are positive instances, in which sentences 1, 3, and 5, directly indicate the relationship **/contains/** between *"Boston"* and *"Fenway_Park"* according to the word *"in"* or the collocation *"'s"*, while the other two sentences imply the relationship between entity pairs at the semantic level. Obviously, sentence 1 reflects the relationship between entities concisely and intuitively, so the sentence-related selection model takes it as the best instance and sets its relevance as 1. Then, according to the feature vector, the relevance of other sentences with sentence 1 can be calculated based on the cosine scores. Finally, the scores are normalized to get the final weights. It can be seen that with the sentence-related selection, our model can effectively capture the potential valid inter-sentence features, and thus improve the final performance of relation extraction.

3.6 Error Analysis

In order to analyze the possible reasons of classification errors, we randomly selected 100 misclassified samples from the test set and performed manual analysis on them. The errors can be roughly categorized into following two types:

1) Confusion of Similar Relations (90/100). We observe that it is difficult for our model to distinguish some similar relations, such as **/place_of_birth/** and **/place_lived/**.

The place of birth and the place of residence of a person are likely to be the same, so the model is likely to mistakenly classify such relations. In subsequent studies, we try to solve this problem by incorporating some word/phrase collocation patterns.

2) Lack of Background Knowledge (10/100). For example, for a triple (*All Blacks*, /**location**/, *New Zealand*) and a sentence "*...in New Zealand, the All Blacks is so dominant that...*", Due to the absence of the background knowledge that the *All Blacks* is the name of a basketball team, our model predicts the relationship as /**place_lived**/. We may improve our model by introducing external knowledge such as entity description in the future.

4. Related Work

Distant supervision plays an increasingly important role in RE to solve the lack of annotated training data. However, this method suffers from wrong labeling problem due to the strong assumption of distant supervision. For this, Riedel et al. [16] modeled distantly supervised RE as a single labeling problem by using multi-instance learning. Surdeanu et al. [18] and Hoffmann et al. [19] adopted the multi-instance multi-label learning and used a probabilistic graphical model to select sentences. However, all of the above methods rely heavily on the quality of features generated by NLP tools and are deeply affected by error propagation problem.

With the rapid development of neural networks, Zeng et al. [2] proposed PCNNs with MIL to select the most likely positive sentences; Lin et al. [5] used selective attention over instance with PCNNs to select valid sentences; Ji et al. [3] assigned more precise attention weights by making use of entity descriptions; and Yuan et al. [6] used a non-independent and non-identically distributed (non-IID) relevance embedding to capture the relevance of sentences in bags to extract the relevance between sentences in bags and get better bag vector representation. However, Li et al. [4] found that there are many bags containing only one sentence in the training data, and the selective attention mechanism will not achieve the expected effect at this situation, so they designed a pooling mechanism based on rich context representation as an aggregator to address this problem. To obtain richer information from sentences and make full use of the information of entity pairs, Yuan et al. [7] proposed combined networks containing PCNNs and bidirectional gated recurrent units (BiGRU) with multi-level attention.

On the other hand, reinforcement learning (RL) has been used to select the valid instances before training for relation extraction [20], [21]. RL mainly consists of two parts: case selector and relationship classifier, in which the former is used to select high-quality instances, while the latter is used for relationship prediction and reward feedback. Experimental results show that the case selector can effectively eliminate noisy data, thus improve the performance of relation extraction. Furthermore, Wu et al. [22] introduced an adversarial training into the task of distant supervision relationship extraction for the first time. While Qin et al. [23] adopted a generic adversarial network (GAN) to filter the training data and improved the performance of relationship extraction.

To alleviate the error propagation problem, the soft-label strategy is a feasible method during the feature selection process. Liu et al. [10] proposed a joint scoring function, which dynamically selected the original labels and the relation labels generated based on the entity pair to obtain the new labels of the entity pair. Focused on the imbalance of datasets, a label-free method has been proposed by Wang et al. [24]. Based on a context-dependent rectification strategy, Huang et al. [25] adjusted the labels that might otherwise be wrong in the right direction.

However, all the above approaches filtered noise at the sentence level, ignoring the word-level (intra-sentence) noise, which widely exists inside sentences. Moreover, MIL only selects the sentence with the highest probability to be a valid candidate, so that a large amount of rich information is lost [5], which results in insufficient use of the supervision information in a bag. And the relevance of sentences is ignored consequently in most of the previous studies. Actually, sentences with the same entity pairs are more or less related in the same bag, which can be used to improve performance.

Motivated by aforementioned observations, a novel SR-GPCNN framework is proposed in this paper, which further improve the gate convolution layer [26] with multi-head self-attention mechanism to reduce word-level noise by selecting more important intra-sentence features. Furthermore, we also introduce a soft-label strategy in our model by adopting bilinear transformation results of entity pairs as relation labels, to reduce error propagation. Besides, a sentence-related selection method is adopted to filter the sentence-level noise. Experimental results verify the effectiveness of the SR-GPCNN model. Note that, Liu et al. [8] proposed a word-level noise filtering method. The differences between ours and theirs lie in: Liu et al. used NLP tools to build dependency subtrees, and introduced external knowledge to filter word-level noise through transfer learning; whereas our model only uses two mechanisms (gate mechanism and MSA mechanism), which not only ensures the completeness of sentences, but also avoids using external tools and knowledge.

5. Conclusions and Future Work

Aiming at tackling the low-quality corpus problem, we propose a brand-new distantly supervised approach for relation extraction, named SR-GPCNN, which adopts a gate mechanism with multi-head self-attention and a soft-label strategy to cut down word-level noise by valid inner-sentence feature selection, and also designs a sentence-related selection method to model sentence relevance by cosine similarity. The gate mechanism can effectively select word-level features extracted by convolution layer. The multi-head self-attention mechanism is adopted after convolution layer to

reduce the influence of word-level noise. The soft-label strategy is introduced to improve the accuracy of feature selection and reduce the error propagation problem by using bilinear transformation results between entity pairs. Furthermore, the sentence-related selection method can be used to assign more reasonable weights to sentences in a bag. The experimental results conduct on popular NYT datasets show that our approach is significantly superior to all baseline systems and achieves the best results.

In the future, we will incorporate reinforcement learning to filter noise from different aspects. Meanwhile, we will introduce the external prior knowledge to explore ways to improve the performance of relation extraction further.

Acknowledgments

The research work described in this paper has been supported by the Natural Science Foundation of China under Grant No. 61976016, 61976015, and 61876198, and the Beijing Municipal Natural Science Foundation under Grant No. 4172047.

References

[1] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, "Distant supervision for relation extraction without labeled data," Proc. Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, pp.1003–1011, 2009.

[2] D. Zeng, K. Liu, Y. Chen, and J. Zhao, "Distant supervision for relation extraction via piece-wise convolutional neural networks," Proc. Conference on Empirical Methods in Natural Language Processing, pp.1753–1762, 2015.

[3] G. Ji, K. Liu, S. He, L. Xu, and J. Zhao, "Distant supervision for relation extraction with sentence-level attention and entity descriptions," Proc. AAAI-2017, pp.3060–3066, 2017.

[4] Y. Li, G. Long, T. Shen, T. Zhou, L. Yao, H. Huo, and J. Jiang, "Self-Attention Enhanced Selective Gate with Entity-Aware Embedding for Distantly Supervised Relation Extraction," Proc. AAAI-2020, pp.8269–8276, 2020.

[5] Y. Lin, S. Shen, Z. Liu, H. Luan, and M. Sun, "Neural relation extraction with selective attention over instances," Proc. 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp.2124–2133, 2016.

[6] C. Yuan, H. Huang, C. Feng, X. Liu, and X. Wei, "Distant Supervision for Relation Extraction with Linear Attenuation Simulation and Non-IID Relevance Embedding," Proc. AAAI Conference on Artificial Intelligence, pp.7418–7425, 2019.

[7] H. Yuan, "Combined Networks with Multi-level Attention for Distantly-Supervised Relation Extraction," Journal of Physics Conference Series, vol.1550, pp.1–4, 2020.

[8] T. Liu, X. Zhang, W. Zhou, and W. Jia, "Neural relation extraction via inner-sentence noise reduction and transfer learning," Proc. Conference on Empirical Methods in Natural Language Processing, pp.2195–2204, 2018.

[9] S. Vashishth, R. Joshi, and S. Prayaga, "Reside: Improving distantly-supervised neural relation extraction using side information," Proc. Conference on Empirical Methods in Natural Language Processing, pp.1257–1266, 2018.

[10] T. Liu, K. Wang, B. Chang, and Z. Sui, "A soft-label method for noise-tolerant distantly supervised relation extraction," Proc. Conference on Empirical Methods in Natural Language Processing, pp.1790–1795, 2017.

[11] S. Hochreiter and J. Schmidhuber, "Long short-term memory," Neural computation, vol.9, no.8, pp.1735–1780, 1997.

[12] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. Gomez, and L. Kaiser, "Attention is all you need," Advances in Neural Information Processing Systems, pp.5998–6008, 2017.

[13] Bordes, N. Usunier, A. Garcia-Duran, J. Weston, and O. Yakhnenko, "Translating embeddings for modeling multi relational data," Proc. International Conference on Neural Information Processing Systems, pp.2787–2795, 2013.

[14] N. Kalchbrenner, L. Espeholt, K. Simonyan, A. Oord, A. Graves, and K. Kavukcuoglu, "Neural machine translation in linear time," arXiv preprint arXiv, 1610.10099, 2016.

[15] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. Dauphin, "Convolutional sequence to sequence learning," Proc. 34th International Conference on Machine Learning, pp.1243–1252, 2017.

[16] S. Riedel, L. Yao, and A. McCallum, "Modeling relations and their mentions without labeled text," Proc. Machine Learning and Knowledge Discovery in Databases, pp.148–163, 2010.

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," arXiv preprint arXiv, 1301.3781, 2013.

[18] M. Surdeanu, J. Tibshirani, R. Nallapati, and C.D. Manning, "Multi-instance multi-label learning for relation extraction," Proc. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, pp.455–465, 2012.

[19] R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D.S. Weld, "Knowledge-based weak supervision for information extraction of overlapping relations," Proc. ACL-2011, pp.541–550, 2011.

[20] J. Feng, M. Huang, L. Zhao, Y. Yang, and X. Zhu, "Reinforcement learning for relation classification from noisy data," Proc. AAAI-2018, pp.5779–5786, 2018.

[21] P. Qin, W. Xu, and W.Y. Wang, "Robust distant supervision relation extraction via deep reinforcement learning," Proc. ACL, pp.2137–2147, 2018.

[22] Y. Wu, D. Bamman, and S. Russell, "Adversarial training for relation extraction," Proc. Conference on Empirical Methods in Natural Language Processing, pp.1778–1783, 2017.

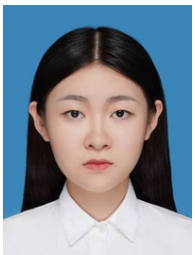
[23] P. Qin, W. Xu, and W.Y. Wang, "DSGAN: Generative Adversarial Training for Distant Supervision Relation Extraction," Proc. 56th Annual Meeting of the Association for Computational Linguistics, pp.496–505, 2018.

[24] G. Wang, W. Zhang, R. Wang, Y. Zhou, X. Chen, W. Zhang, H. Zhu, and H. Chen, "Label-free distant supervision for relation extraction via knowledge graph embedding," Proc. Conference on Empirical Methods in Natural Language Processing, pp.2246–2255, 2018.

[25] X. Huang, B. Zhang, Y. Ye, X. Chen, and X. Li, "A Noise Adaptive Model for Distantly Supervised Relation Extraction," Proc. CCF International Conference on Natural Language Processing and Chinese Computing, pp.519–530, 2020.

[26] X. Li, Y. Chen, J. Xu, and Y. Zhang, "Attention-Based Gated Convolutional Neural Networks for Distant Supervised Relation Extraction," Proc. Chinese Computational Linguistics, pp.246–257, 2019.

Yufeng Chen received the B.S. degree in Mechanical Electrical Engineering from Beijing Jiaotong University in 2003, and the Ph.D degree in Pattern Recognition and Intelligent Systems from National Lab of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, in June, 2008. From July 2008 to September 2014, she worked in NLPR. From September 2014, she joined School of Computer and Information Technology, Beijing Jiaotong University and now works as an associate professor. Her research interests include natural language processing, machine translation, information extraction and so on.



Siqi Li is currently pursuing the master's degree with the School of Computer and Information Technology, Beijing Jiaotong University, Beijing, China. Her current research interests include natural language processing and information extraction.

Xingya Li received the M.S. degree in computer science from Beijing Jiaotong University in 2020. During the master's degree, he mainly focused on distantly supervised relation extraction in NLP, and devoted himself to reducing the word-level noise so as to obtain better semantic representation of the sentences.

Jinan Xu is currently a Professor in School of Computer Science and information technology, Beijing Jiaotong University, Beijing, China. He received the B.S. degree from Beijing Jiaotong University in 1992, the MS degree and Ph.D. degree in computer information from Hokkaido University, Sapporo, Japan, in 2003 and 2006, respectively. His research focuses on natural language processing, machine translation, information retrieve, text mining, and machine learning. He is a member of CCF, CIPSC, ACL and the ACM.

Jian Liu received his B.S. and M.S. degrees in Computer Science from Northeast University in 2013 and 2015, and Ph.D degree in Pattern Recognition and Intelligent System from National Lab of Pattern Recognition (NLPR), Institution of Automation, Chinese Academy of Sciences, in 2020. He is currently a lecturer in the School of Computer and Information Technology, Beijing Jiaotong University. His research interests include natural language processing, information extraction, and machine learning. He is currently working on projects for extracting event information from large unstructured texts.