Information Processing and Management 57 (2020) 102373

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

A gated piecewise CNN with entity-aware enhancement for distantly supervised relation extraction

Haixu Wen[a], Xinhua Zhu[a,b,*], Lanfang Zhang[c], Fei Li[d]

[a] School of Computer Science and Information Engineering, Guangxi Normal University, Guilin 541004, China
[b] Guangxi Key Laboratory of Multi-Source Information Mining and Security, Guangxi Normal University, Guilin 541004, China
[c] Faculty of Education, Guangxi Normal University, Guilin 541004, China
[d] School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China

ARTICLE INFO

Keywords:
Distant supervision
Relation extraction
PCNN
Multi-head self-attention
Gate mechanism

ABSTRACT

The piecewise convolutional neural network (PCNN) is an important method for distant supervision relation extraction. However, the existing methods based on the PCNN still have the following shortcomings: these methods lack the consideration of the impacts of entity pairs and the sentence context on word encoding and do not distinguish the different contributions of the three segments in PCNN to relation classification. To solve these problems, we propose a novel gated piecewise CNN with entity-aware enhancement for distantly supervised relation extraction. First, we use a multi-head self-attention mechanism to combine the word embedding with the head/tail entity embedding and relative position embedding to generate an entity-aware enhanced word representation, which is capable of capturing the semantic dependency between each word and entity pair. Then we introduce a global gate to combine each entity-aware enhanced word representation with their average in the input sentence to form the final word representation of the PCNN input. Moreover, to determine the key segments where the most important information for relation classification appears, we design another gate mechanism to assign a different weight to each sentence segment to highlight the effects of key segments on the PCNN. Experiments on New York Times dataset demonstrate that our model significantly outperforms most of the state-of-the-art models.

1. Introduction

Relation extraction (RE) (Culotta & Sorensen, 2004; Zelenko, Aone, & Richardella, 2003), a subtask of information extraction (Wu & Weld, 2010), extracts the sematic relations between two given entities in natural language texts, which is a basic technology for some natural language processing (NLP) tasks such as knowledge base completion (De Sa et al., 2016; Li, Taheri, Tu, & Gimpel, 2016) and question answering (Li et al., 2019; Sun et al., 2015). However, there is a challenge for RE that obtaining large-scale manually labeled data is an extremely time-consuming job (Culotta & Sorensen, 2004; Mooney & Bunescu, 2006; Zelenko et al., 2003). Mintz, Bills, Snow, and Jurafsky (2009) proposes a distant supervision (DS) method to align a knowledge base (KB) such as Freebase (Bollacker, Evans, Paritosh, Sturge, & Taylor, 2008) with the unstructured texts to automatically generate large-scale training data. DS relies on a strong assumption that if two entities have a relation in the KB, all sentences containing the two entities will express the same relation. Apparently, the strong assumption ignores the cases that different sentences mentioning the same two entities express

[*] Corresponding author at: School of Computer Science and Information Engineering, Guangxi Normal University, Guilin 541004, China.
E-mail address: zxh429@263.net (X. Zhu).

https://doi.org/10.1016/j.ipm.2020.102373
Received 29 March 2020; Received in revised form 15 July 2020; Accepted 12 August 2020
Available online 18 August 2020

Table 1
An example of noisy labeling problem caused by distant supervision.

	Sentence	Label	Correct label
Bag	S1. **Donald Trump** was born in the **United States**. S2. **Donald Trump** is the 45th president of the **United States**. S3. **Donald Trump** was a businessman from New York, **United States**.	president_of	born_in president_of place_lived

different relations according to their contexts, which produces many noisy sentences with wrong labels. For example, as shown in Table 1, there is a relation fact (*Donald Trump, president_of, United States*) with three sentences in the knowledge base. All sentences are labeled as "*president_of*" by distant supervision, but the correct relations respectively expressed by S1 and S3 are "*born_in*" and "*place_lived*".

It is effective to adopt multi-instance learning (MIL) (Riedel, Yao, & McCallum, 2010) for distantly supervised relation extraction to reduce the effect of the noisy labeling problem. The MIL framework is built on at-least-one assumption that if all sentences containing the same entity pair are marked as a certain relation by DS, at least one sentence expresses this relation. In MIL, a relation label existing in a KB is assigned to a bag of sentences with the same entity pair rather than each single sentence. With deep learning (LeCun, Bengio, & Hinton, 2015) widely applied in NLP, many neural network methods (Han, Yu, Liu, Sun, & Li, 2018a; Lin, Shen, Liu, Luan, & Sun, 2016; Liu, Wang, Chang, & Sui, 2017; Ye & Ling, 2019; Zeng, Liu, Lai, Zhou, & Zhao, 2014) based on MIL have achieved a significant improvement in distantly supervised relation extraction. To alleviate the impact of the noisy sentences of a bag, a series of works apply an attention mechanism (Du, Han, Way, & Wan, 2018; Han et al., 2018a; Lin et al., 2016; Ye & Ling, 2019) to the MIL-based framework (Riedel et al., 2010), which achieves the state-of-the-art results. However, the above methods still have some shortcomings that need to be improved. For example, the existing methods lack the consideration of the impacts of the entity pairs and sentence context on word encoding, which may ignore some important semantic information. For the sentence "Other winners were Alain Mabanckou from Congo, **Nancy Huston** from **Canada** and Lonora Miano from Cameroon" with its label */people/person/nationality*, it is obvious that it is more easier to predict the relation of */people/person/nationality* by incorporating the features of the head entity "**Nancy Huston**" and the tail entity "**Canada**" into the word "**from**" than simply using the word embedding and relative position embedding to encode this word "**from**". Furthermore, the different contributions of the three segments to relation classification have not been explored further.

To solve the above problems, we propose a novel gated piecewise convolutional neural network with an entity-aware enhancement (EA-GPCNN). First, different from the word representation in the PCNN, which only combines the word embedding and the relative position with the entity pair, our framework uses two multi-head self-attention networks (Vaswani et al., 2017) to integrate the entity embedding and position embedding (Zeng et al., 2014) to generate an entity-aware enhanced word representation, aiming to capture the semantic dependency between each word and entity pair. Then, to achieve word encoding based on the sentence context, we add a global information integration into the word representations before the PCNN, that is, we introduce a gate structure (Xu, Ji, Huang, Deng, & Li, 2019) called the global gate in this paper to obtain the global information of the sentence and integrates it into each entity-aware enhanced word representation, which is then fed into the PCNN. Moreover, we believe that the three segments of the sentence divided according to the position of the two entities have different degrees of importance in relation classification. Therefore, we add a segment-level gate mechanism to assign different weights to the three segments in the output of the PCNN, which can emphasize the important segments and weaken the effects of irrelevant segments.

Our main contributions of this paper are as follows:

- We propose an entity-aware enhanced word representation with rich context, which enables the downstream modules to learn robust semantic features.
- We combine the global gate structure and the PCNN (Zeng, Liu, Chen, & Zhao, 2015) to better capture the global and local features of the sentence.
- We introduce a gate mechanism after the max-pooling layer of the PCNN model, which assigns different weights to the three segments and highlights the effects of crucial segments.
- Our model is evaluated on a widely used benchmark dataset and outperforms most of the state-of-the-art methods.

The remainder of the paper is organized as follows. Section 2 introduces the related work. Section 3 describes the research objective of the paper. Section 4 descibes our proposed model in detail. Section 5 reports the experimental results on New York Times dataset, ablation study and the extended datasets for other relation classification tasks. Section 6 gives the conclusion.

2. Related work

Traditional RE methods (Culotta & Sorensen, 2004; Mooney & Bunescu, 2006; Zelenko et al., 2003) are mostly based on supervised learning and viewed as a multi-classification problem. These models lack enough labeled data, which are expensive to acquire, and this has become a bottleneck in improving the performance of RE. To solve this problem, distant supervision is proposed by Mintz et al. (2009) to automatically generate large-scale data through aligning knowledge base to plain texts. Since the assumption of distant supervision is too strong, it will produce some wrong labeled sentences. To reduce the impact of noisy data, Riedel et al. (2010) introduces a multi-instance learning framework to relax the strong assumption and assumes that if all sentences containing the same entity pair are marked as a certain relation by distant supervision, at least one sentence expresses this relation. Hoffmann, Zhang, Ling, Zettlemoyer, and Weld (2011) and Surdeanu, Tibshirani, Nallapati, and Manning (2012) use multi-instance multi-label learning to deal with the cases where there are multiple relations between a given entity pair. Owing to these methods relying on the accuracy of feature engineering, they cannot avoid the error propagation problem caused by NLP tools.

Recently, deep learning (LeCun et al., 2015) has been widely applied to distant supervision relation extraction and neural network models (Han et al., 2018a; Huang & Du, 2019; Ji, Liu, He, & Zhao, 2017; Lin et al., 2016; Liu et al., 2017; Ye & Ling, 2019; Zeng et al., 2015) have achieved remarkable improvements compared to statistical methods (Hoffmann et al., 2011; Mintz et al., 2009; Riedel et al., 2010; Surdeanu et al., 2012). Zeng et al. (2015) proposes a piecewise convolutional neural network (PCNN) to automatically extract sentence features and select the most likely sentence as a bag representation in the multi-instance learning framework. Lin et al. (2016) introduces an attention mechanism to the PCNN, which takes advantage of all the sentences information in a bag. Liu et al. (2017) demonstrates that a soft label at the entity level can alleviate the impact of mislabeling sentences during the training process. To enhance the ability of selecting valid instances, Du et al. (2018) employs a structured self-attention mechanism to better capture the contextual information of sentences. To address the noisy bags with wrong labels, Ye and Ling (2019) adopts both sentence-level and bag-level attention to reduce the noise in the relation extraction task. Huang and Du (2019) combines the CNN and self-attention mechanism to model the sentence and introduces curriculum learning (Bengio, Louradour, Collobert, & Weston, 2009) into their model to alleviate the effect of noisy data.

Some recent works (Han et al., 2018a; Hu et al., 2019; Ji et al., 2017; Vashishth, Joshi, Prayaga, Bhattacharyya, & Talukdar, 2018) attempt to use additional information to improve relation extraction. Ji et al. (2017) introduces entity description information to the RE task. Vashishth et al. (2018) uses additional information in the knowledge base to impose soft constraints on relation classification. Hu et al. (2019) utilizes the entity description and the structural information from the knowledge base to learning a label embedding for relation extraction. In addition, some works have tried to apply adversarial learning (Qin, Xu, & Wang, 2018a) and reinforcement learning (Feng, Huang, Zhao, Yang, & Zhu, 2018; Qin, Xu, & Wang, 2018b) to relation extraction models, which will filter noisy instances and improve the robustness of their models. Ru, Tang, Li, Xie, and Wang (2018) employs semantic similarity with word embedding to filter the wrong labels. Ye and Luo (2019) exploits the latent connections between relations and effectively avoids the problem of class imbalance in relation extraction.

Different from the above models, our proposed model further extracts the segment-level features of the PCNN output with a gate mechanism. Inspired by Du et al. (2018) and Xu et al. (2019), we use a self-attention mechanism (Vaswani et al., 2017) to obtain a more expressive word representation and introduce the global gate as defined in Section 3.2 to capture the global information of the sentences.

3. Research objective

The objective in this paper is to address the following shortcomings of the existing methods based on PCNN: (i) not considering the impacts of entity pairs and the sentence context on word encoding and (ii) not distinguishing the different contributions of the three segments in the PCNN to relation classification. To solve these problems, we try to propose a novel gated piecewise convolutional neural network with entity-aware enhancement. First, to capture the semantic dependency between each word and entity pair, our framework uses two multi-head self-attention networks (Vaswani et al., 2017) to integrate the entity embedding and position embedding (Zeng et al., 2014) to generate an entity-aware enhanced word representation. Furthermore, to achieve word encoding based on the sentence context, we introduce a gate structure (Xu et al., 2019) called the global gate in this paper to integrate the global feature into each entity-aware enhanced word representation to form the final word representation of the PCNN input. Finally, to emphasize the important segments and weaken the effects of irrelevant segments, we introduce a segment-level gate mechanism to assign different weights to the three segments in the output of the PCNN.

4. Methodology

In this section, we introduce a novel gated piecewise convolutional neural network with entity-aware enhancement (EA-GPCNN) for distantly supervised relation extraction. The framework of our model is shown in Fig. 1, which mainly contains three modules:

- **Entity-aware enhanced word representation**: We adopt two multi-head self-attention networks (Vaswani et al., 2017) to encode the sentence, which effectively leverages the information for two target entities and relative position features to produce an entity-aware enhanced word representation.
- **Global gate**: This module uses a gate structure (Xu et al., 2019) called the global gate in this paper to capture the global information of the entity-aware enhanced word representation and then integrates it into each entity-aware enhanced word representation.
- **PCNN with gate mechanism**: We employ a gating mechanism to allocate different weights to the three segments in the output of the PCNN (Zeng et al., 2015), highlighting the effects of the key segments.

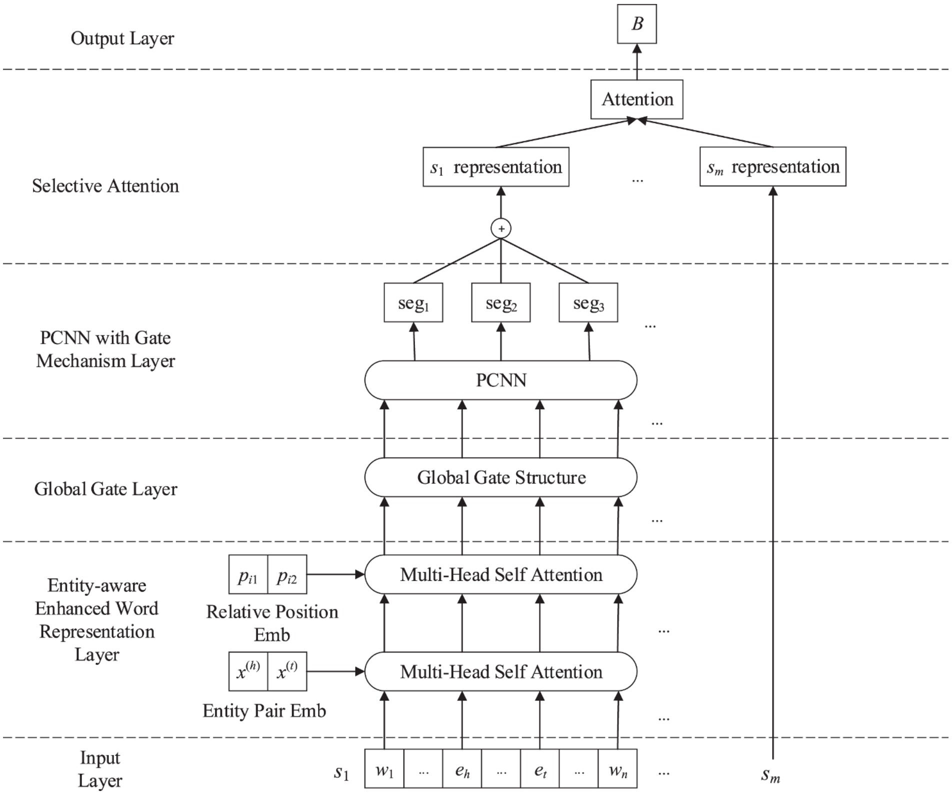

Fig. 1. The framework of our proposed approach with three main modules: the entity-aware enhanced word representation, the global gate and the PCNN with gate mechanism.

4.1. Entity-aware enhanced word representation

4.1.1. Word and position representation

For a given sentence $s = \{w_1, w_2, ..., w_n\}$, which contains head entity e_h and tail entity e_t, we use the pretrained word2vec (Mikolov, Chen, Corrado, & Dean, 2013) to map each word of the sentence to a low-dimension dense vector. Then, we get a sequence $[x_1, x_2, ..., x_n] \in \mathbb{R}^{n \times d_w}$, where d_w denotes the dimension of the word embedding and n is the length of the sentence. The vector representations of head entity e_h and tail entity e_t in a bag are $x^{(h)}$ and $x^{(t)}$, respectively.

In addition, the relative positions (Zeng et al., 2014) are the important features of relation extraction, which help the RE models pay more attention to those words close to the entity pair. They are denoted by the distances between each word w_i and two target entities e_h and e_t. For instance, for "[*Donald Trump*]$_{e_1}$ was born in the [*United States*]$_{e_2}$", the relative positions of word "*born*" to entity e_1 "*Donald Trump*" and entity e_2 "*United States*" are 2 and -3 respectively. For a word w_i, its position values will be projected as two low-dimensional vectors with randomly initialized weights denoted as $p_{i1} \in \mathbb{R}^{d_p}$ and $p_{i2} \in \mathbb{R}^{d_p}$, where d_p denotes the dimension of the relative position embedding.

4.1.2. Integration of entity and position embedding

Different from previous PCNN-based models (Han et al., 2018a; Liu et al., 2017; Ye & Ling, 2019; Zeng et al., 2015), our model uses two multi-head self-attention networks (Vaswani et al., 2017) to model the sentence to get the input representation of the PCNN instead of simply concatenating the word embedding and position embedding (Zeng et al., 2014), which is able to capture the semantic dependency between each word and entity pair. Fig. 2 illustrates the structure of the multi-head self-attention mechanism. First, we suppose there is a single head for self attention. The queries Q, keys K and values V are calculated by the linear transformation of the input sequence $X \in \mathbb{R}^{n \times d}$ (d denotes the dimension of the word embedding). Then, the inputs Q, K and V are fed into the scaled dot-product attention module to compute the output of the attention. The scaled dot-product attention is calculated as follows:

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d}})V \tag{1}$$

In multi-head self-attention, the scaled dot-product attention is performed on h heads in parallel using the projected vectors, which helps the model capture the information from different subspaces (Vaswani et al., 2017). The outputs of all heads are concatenated and finally projected to the original representation space. Specifically, multi-head self-attention can be formulated as

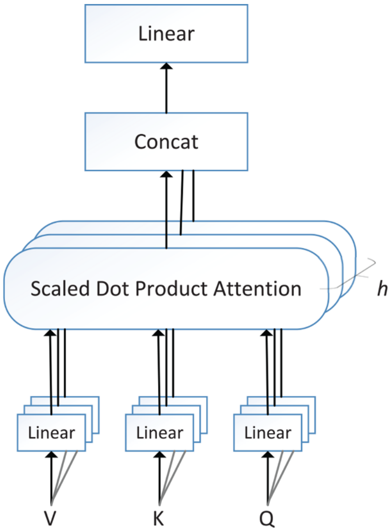

Fig. 2. Multi-head self-attention. The inputs Q (queries), K (keys), and V (values) should be the same vectors. In our work, they are equivalent to the representation vectors of the input sequence.

below:

$$MultiHead(Q, K, V) = [head_1; head_2; ..., head_h]W^R \tag{2}$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \tag{3}$$

where $W_i^Q \in \mathbb{R}^{d\times d_k}$, $W_i^K \in \mathbb{R}^{d\times d_k}$, $W_i^V \in \mathbb{R}^{d\times d_v}$ and $W^R \in \mathbb{R}^{d\times d}$ are the projection matrices, which are learnable during the training process. h is the number of attention heads and $d_k = d_v = d/h$. W^R is a linear transformation for the multi-head attention outputs and W_i^Q, W_i^K and W_i^V are respectively used to project the inputs queries, keys and values to the ith subspace.

In the following, we adopt two multi-head self-attention layers (Vaswani et al., 2017) described as above to combine the head/tail entity and relative position information to generate an entity-aware enhanced word representation. Firstly, we concatenate each word embedding of the input sentence with the head entity embedding $x^{(h)}$ and tail entity embedding $x^{(t)}$ to form the input sequence $X^{(e)} = [x_1^{(e)}, x_2^{(e)}, ..., x_n^{(e)}] \in \mathbb{R}^{n\times 3d_w}$, where $x_i^{(e)} = [x_i; x^{(h)}; x^{(t)}]$, and x_i is the ith token of the input sentence. The sequence $X^{(e)}$ is fed into the first multi-head self attention layer ($Q = K = V = X^{(e)}$), and we obtain the hidden representation $X^{(h)} = [x_1^{(h)}, x_2^{(h)}, ..., x_n^{(h)}] \in \mathbb{R}^{n\times 3d_w}$:

$$X^{(h)} = MultiHead(X^{(e)}, X^{(e)}, X^{(e)})) \tag{4}$$

Then, to further highlight the correlation between each word and two target entities, we concatenate the two position embeddings p_{i1} and p_{i2} to the hidden representation to form the sequence $X^{(p)} = [x_1^{(p)}, x_2^{(p)}, ..., x_n^{(p)}] \in \mathbb{R}^{n\times(3d_w+2d_p)}$, where $x_i^{(p)} = [x_i^{(h)}; p_{i1}; p_{i2}] \in \mathbb{R}^{3d_w+2d_p}$, and $x_i^{(h)}$ is the ith token of the hidden representatiion $X^{(h)}$. Then, $X^{(p)}$ is fed into the second multi-head self-attention layer, and we obtain an entity-aware enhanced word representation:

$$X^{(ep)} = MultiHead(X^{(h)}, X^{(h)}, X^{(h)})) \tag{5}$$

where $X^{(ep)} = [x_1^{(ep)}, x_2^{(ep)}, ..., x_n^{(ep)}] \in \mathbb{R}^{n\times(3d_w+2d_p)}$ denotes an entity-aware enhanced word representation, which can provide a more expressive input representation for downstream modules.

4.2. Global gate

Inspired by the work of Xu et al. (2019), we apply a gate structure that we call the global gate in this paper to capture the global information of the entity-aware enhanced word representation. The global gate can integrate the global information into each entity-aware enhanced word representation, which produces a more expressive representation as an input of the PCNN. We first compute the mean of the entity-aware enhanced word representation $X^{(ep)}$ and we obtain the result $\bar{x}$, which is regarded as the context vector of sequence $X^{(ep)}$. For each token $x_i^{(ep)}$ in $X^{(ep)}$, the global gate module multiplies the context vector $\bar{x}$ element-wise with input token representation $x_i^{(ep)}$, whose result is fed into a gating function to obtain the corresponding gated vector g_i. The gated vector g_i indicates whether there is a correlation between the current word $x_i^{(ep)}$ and the context vector $\bar{x}$. Finally, gated vector g_i is multiplied word $x_i^{(ep)}$ and we get the gated output containing global information. The details of the global gate are shown in Fig. 3. The global gate is calculated as follow:

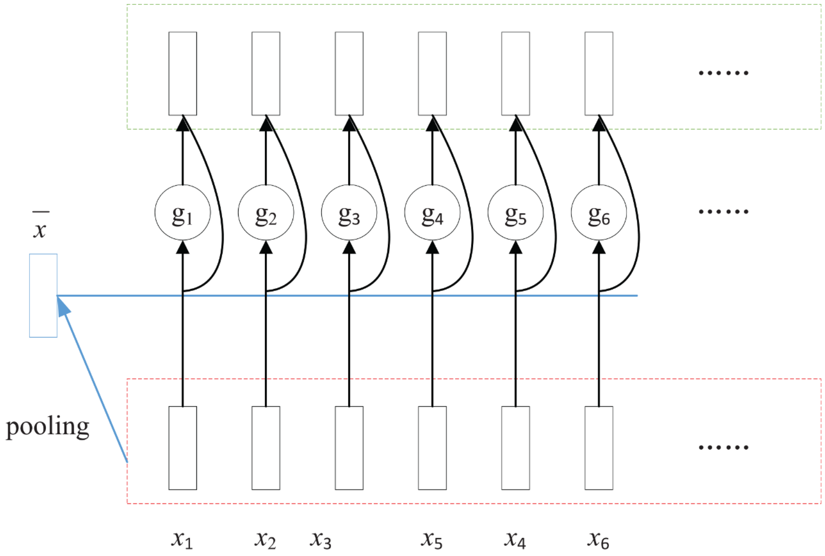

Fig. 3. The structure of the global gate.

$$\bar{x} = \frac{\sum_i^n x_i^{(ep)}}{n} \tag{6}$$

$$g_i = \sigma(W^g \bullet (x_i^{(ep)} \odot \bar{x}) + b^g) \tag{7}$$

$$x_i^{(g)} = x_i^{(ep)} \odot g_i \tag{8}$$

where σ denotes the sigmoid activation function and $\odot$ denotes element-wise multiplication. $W^g \in \mathbb{R}^{d_h \times d_h}(d_h = 3d_w + 2d_p)$ is a learnable parameter and b^g is a bias. g_i is the ith column of the global gate matrix $G \in \mathbb{R}^{n \times d_h}$. At the end, the gated output $X^{(g)} = [x_1^{(g)}, x_2^{(g)}, ..., x_n^{(g)}] \in \mathbb{R}^{n d_h}$ is fed into the PCNN.

4.3. PCNN with gate mechanism

We select the PCNN (Zeng et al., 2015) as the sentence encoder to capture the semantic features of the input sequence. However, unlike the previous methods (Huang & Du, 2019; Lin et al., 2016; Liu et al., 2017; Ye & Ling, 2019; Zeng et al., 2015), we adopt a gate mechanism to assign different weights to the three segments of the PCNN output, which highlights the key segments and softens the role of unrelated segments, aiming to further explore the different contributions of the three segments to relation classification. We use $X^{(g)} = [x_1^{(g)}, x_2^{(g)}, ..., x_n^{(g)}] \in \mathbb{R}^{n \times d_h}$ as the input of the convolution layer and define the weight matrix of the filter $W^c \in \mathbb{R}^{l \times d_h}$, where l is the size of the filter. After the convolution layer, we get a vector $C \in \mathbb{R}^{n-l+1}$ and its ith element c_i is calculated as follows:

$$c_i = W^c \otimes \ x_{i:i+l-1}^{(g)} + b^c \qquad 1 \le i \le n \tag{9}$$

where $x_{i:i+l-1}^{(g)}$ denotes the concatenation of l words in the sequence $X^{(g)}$. $\otimes$ denotes the convolution operator and b^c is a bias. To extract different features of the sentence, we use a set of filters denoted as $[W_1^c, W_2^c, ..., W_{d_c}^c]$ in the convolution layer, where d_c is the number of filters.

Then, piecewise max-pooling (Zeng et al., 2015) is used for each filter, which is divided into three segments according to the positions of the head and tail entities. Three segments of C_i are $\{C_i^1, C_i^2, C_i^3\}$ and we can define the output of the ith filter C_i as follows:

$$q_i = [q_{i,1}, q_{i,2}, q_{i,3}] = [maxpool(C_i^1), maxpool(C_i^2), maxpool(C_i^3)] \tag{10}$$

At the end, all vectors are concatenated to obtain $q = [q^1, q^2, q^3] \in \mathbb{R}^{3 \times d_c}$, where $q^1, q^2, q^3 \in \mathbb{R}^{d_c}$.

After obtaining the output of PCNN q, we apply a segment-level gate mechanism to measure the different contributions of the three segments to relation classification. The sentence representation is calculated as follows:

$$g^i{}_{seg} = \sigma(W^s q^i + b^s) \tag{11}$$

$$P^{(i)} = g^i{}_{seg} \odot q^i \tag{12}$$

$$s = \tanh([P^{(1)}; P^{(2)}; P^{(3)}]) \tag{13}$$

where σ denotes the sigmoid activation fucntion. $W^s \in \mathbb{R}^{d_c \times d_c}$ and b^s is a bias, which are learnable parameters. $g^i{}_{seg}$ is the gate vector of q^i $(i = 1{,}2{,}3)$ and $g^i{}_{seg}$ is multiplied by the i-th output q^i element-wise to obtain the weighted segment output $P^{(i)}$. Then, the three segments $P^{(1)}$, $P^{(2)}$, $P^{(3)}$ are concatenated together to get the final representation of the sentence s through a hyperbolic tangent function.

4.4. Bag representation

Due to the problem of noisy labels, it is necessary for the model to have a good strategy of selecting those correctly labeled sentences in a bag. We use selective attention (Lin et al., 2016) to calculate the attention weight for each sentence in a bag according to the correlation between the sentence and the predicted relation. For a bag $B = \{s_1, s_2, ..., s_m\}$ containing m sentences, which is marked as relation r, the representation of bag B can be calculated as follows:

$$B = \sum_i \alpha_i s_i \tag{14}$$

$$\alpha_i = \frac{exp(s_i A v_r)}{\sum_j exp(s_j A v_r)} \tag{15}$$

where α_i is the attention weight of sentence s_i, A is a weighted diagonal matrix and $v_r \in \mathbb{R}^{3d_c}$ is the relation vector.

Then, the bag representation B is used to compute the scores of all relations through a linear transformation:

$$o = MB + b^o \tag{16}$$

where $b^o \in \mathbb{R}^{d_r}$ is the bias vector, $M \in \mathbb{R}^{d_r \times 3d_c}$ is the representation matrix of all relations and d_r is the number of relations.

Finally, for a bag B with its relation r, we calculate the probability $p(r|B, \theta)$ as follows:

$$p(r|B, \theta) = \frac{exp(o_r)}{\sum_k^{d_r} exp(o_k)} \tag{17}$$

where o_r indicates the score of the ith relation and θ is all learnable parameters in this paper.

4.5. Loss function

In the distantly supervised relation extraction task, the training set is split into many bags with a labeled relation and the training and test process are built on the bag level. We suppose that there are K bags in the training set $B = \{B_1, B_2, ..., B_K\}$, and their labels are $\{r_1, r_2, ..., r_K\}$. We use cross-entropy to define the objective function as follows:

$$J(\theta) = \sum_{i=1}^{K} \log p(r_i|B_i, \theta) \tag{18}$$

where θ indicates all parameters of our model. We adopt stochastic gradient descent (SGD) to minimize the objective function, which iterates by randomly selecting a mini-batch from the training set.

5. Experiments

5.1. Dataset and evaluation criteria

We evaluate our model on a widely-used dataset, developed by Riedel et al. (2010), which is generated by aligning Freebase (Bollacker et al., 2008) relations with the New York Times corpus[1]. Sentences from 2005 to 2006 are used as the training set and the sentences from 2007 are used as the test set. In the training set, there are 522,611 sentences, 281,270 entity pairs, and 18,252 relation triples. The test set contains 172,448 sentences, 96,678 entity pairs, and 1950 relation triples. The NYT dataset has 53 relations including an NA relation, which indicates that there is no relation between two entities in a sentence.

Following the previous works (Du et al., 2018; Lin et al., 2016; Liu et al., 2017; Ye & Ling, 2019), we use the held-out evaluation methods (Lin et al., 2016) to evaluate the models. We report the experimental results using Precision-Recall curves (PR curves), Precision@N (P@N) metric. In addition, we also use the area under the curve (AUC) in the ablation study.

5.2. Parameter setting

For most of the hyperparameters, we follow the settings in Lin et al. (2016). Grid search is employed to determine some optimal values on training set. We select the batch size among {30,50,100,200} and the number of heads for the two multi-head self-attention layers among {3,4,5,6}. In the training process, we employed mini-batch SGD with the initial learning rate of 0.1, which is decayed by one-tenth every 100,000 steps. We choose the number of heads $h_2 = 5$ for the first multi-head self-attention layer, $h_2 = 5$ for the second multi-head self-attention layer. This paper uses the 50-dimensional word embedding and the 5-dimension position embedding released by Lin et al. (2016) to initialize the word representation. In addition, we use dropout (Srivastava, Hinton, Krizhevsky,

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

Table 2
Hyper-parameters of the model in our experiments.

Parameters	Description	Value
d_w	Dimension of word embedding	50
d_p	Dimension of position embedding	5
l	window size	3
d_c	filter number	230
batch	batch size	50
lr	initial learning rate	0.1
dropout	dropout rate	0.5
gradient clip	gradient clip	5

Sutskever, & Salakhutdinov, 2014) in the output layer with a probability of 0.5. We list the hyperparameters of our model in Table 2.

5.3. Baseline models

We compare the proposed model EA-GPCNN with previous methods which are described as follows:

- **Mintz** (Mintz et al., 2009) is a traditional multi-class logistic regression model.
- **MultiR** (Hoffmann et al., 2011) is a probabilistic graphical model of multi-instance learning for solving overlapping relations.
- **MIMLRE** (Surdeanu et al., 2012) proposes a graphical model that can simultaneously model multiple instances and multiple labels.
- **PCNN** (Zeng et al., 2014) is a CNN-based model with piecewise max-pooling to generate sentence representations for relation extraction.
- **PCNN + ATT** (Lin et al., 2016) adopts selective attention on multiple instances to alleviate the wrongly labeled problem.
- **PCNN + ATT + SL** (Liu et al., 2017) introduces a label-level denoising method to reduce the impact of noisy labels.
- **RESIDE** (Vashishth et al., 2018) adopts a graph convolutional neural network (Kipf & Welling, 2017) to encode the syntactic features of the sentences and introduces side information such as entity type information and relation aliases to improve the performance of distantly supervised relation extraction.

5.4. Experiment results and discussion

5.4.1. Precision-Recall curve

We compare our model with baselines to obtain the Precision-Recall curves and Precision@N, which are shown in Fig. 4 and Table 3 respectively. As shown by the PR curves in Fig. 4, our model achieves the best performance over the entire recall range,

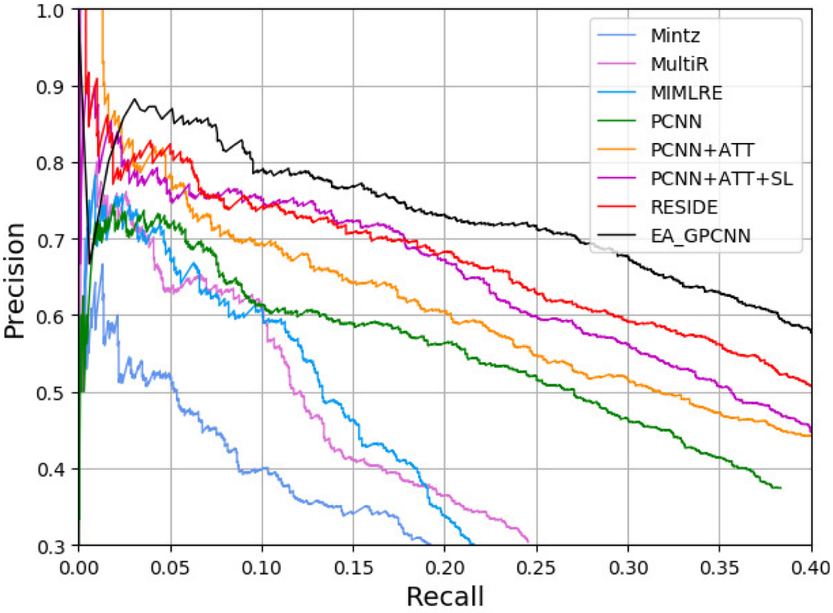

Fig. 4. Precision-recall curves for our proposed model and previous baselines.

Table 3
P@N values of the entity pairs with different number of test instances.

Approach	ONE				TWO				ALL			
P@N(%)	100	200	300	mean	100	200	300	mean	100	200	300	mean
PCNN	73.3	64.8	56.8	65.0	70.3	67.2	63.1	66.9	72.3	69.7	64.1	68.7
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
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
RESIDE	80.0	75.5	69.3	74.9	83.0	73.5	70.6	75.7	84.0	78.5	75.6	79.3
EA-GPCNN	**90.0**	**79.0**	**77.3**	**82.1**	**90.0**	**84.5**	**78.6**	**84.3**	**91.0**	**87.5**	**82.0**	**86.8**

especially for the recall values from 0.05 to 0.40. In general, we observe the following: (1) The performances of the statistical baselines Mintz, MultiR and MIMLRE begin to decline rapidly when the recall rate is greater than 0.1. This indicates that feature-based methods cannot accurately capture the semantic information of the sentence and the error propagation problem of the NLP tools affects the performance of RE. However, neural models can automatically learn features without relying on handcrafted features. (2) The EA-GPCNN significantly outperforms the PCNN, PCNN + ATT, and PCNN + ATT + SL, which demonstrates the effectiveness of our model in solving the problems of noisy labels that appears in distantly supervised relation extraction. It is proved that our model can express the semantic dependency between each word and entity pair through the entity-aware enhanced word representation and improve the performance of RE by applying the segment-level gate mechanism to the output of the PCNN. (3) The performance of RESIDE is better than PCNN + ATT, which shows that rich background knowledge can help the model to alleviate the noise. Moreover, the EA-GPCNN achieves the best performance in Fig. 4, which indicates that our model can make full use of the sentence context without any additional information.

5.4.2. P@N evaluation

We also perform Precision@N tests on entity pairs with different numbers of instances (Lin et al., 2016) and the three test settings employed in the experiment are as follows: ONE randomly selects one instance in the bag, TWO randomly selects two instances in the bag, and ALL uses all instances in the bag for the evaluation. Table 2 shows the results on the NYT dataset regarding P@100, P@200, P@300 and the mean of the three settings for each model. From the results, we observe the following: (1) As the number of instances increases from ONE to ALL, the performances of all methods are improved. The results show that the number of instances in the bags increases, the information that is utilized in relation classification increases. (2) Our EA-GPCNN gets the highest precision values in all three test settings. In comparison with the PCNN + ATT and PCNN + ATT + SL, our proposed method can improve the mean P@N of all sentences by 14.6% and 4.0%, respectively. Compared to RESIDE, our model achieves higher precision by a large margin over all recall values in the PR curves and attains an improvement of 7.5%. These results prove that our model does well in dealing with the problem of noisy labels in distantly supervised relation extraction.

5.4.3. Ablation study

To demonstrate the effectiveness of each module of our model, we perform a series of ablation studies. The P@N, P-R Curves and AUC are used as the evaluation metrics in this section. We report three different ablation models, which are described below:

- EA-GPCNN w/o gate means that the segment-level gate mechanism that is applied to the output of the PCNN (Zeng et al., 2015) is removed.
- EA-GPCNN w/o global gate means that the global gate structure is removed.
- EA-GPCNN w/o ent means that the entity-aware enhanced word representation module is removed. It uses the word representation formed by the word embedding and position embedding used in Zeng et al. (2015).

The results of P@N are shown in Table 4 whose corresponding P-R Curves and AUC are shown in Fig. 5 and Table 5, respectively. These results show that our model significantly outperforms all the variant models. Without the segment-level gate mechanism, the performance of the EA-GPCNN w/o gate decreases by 3.4%, and its AUC value changes from 0.44 to 0.41. This demonstrates the effectiveness of our segment-level gate mechanism on highlighting the key segments that help our model improve the performance of relation classification. When we remove the global gate structure, the performance of the EA-GPCNN w/o global gate decreases by 3.0%, and its AUC value is 0.42. This means that the global gate structure enables the model to better obtain the global features of the sentence and integrates them into each entity-aware enhanced word representation. Compared to the original EA-GPCNN model, the

Table 4
P@N results of ablation study.

P@N(%)	100	200	300	mean
EA-GPCNN	**91.0**	**87.5**	**82.0**	**86.8**
EA-GPCNN w/o gate	89.0	82.0	79.3	83.4
EA-GPCNN w/o global gate	90.0	81.5	80.0	83.8
EA-GPCNN w/o ent	79.0	76.5	71.3	75.6

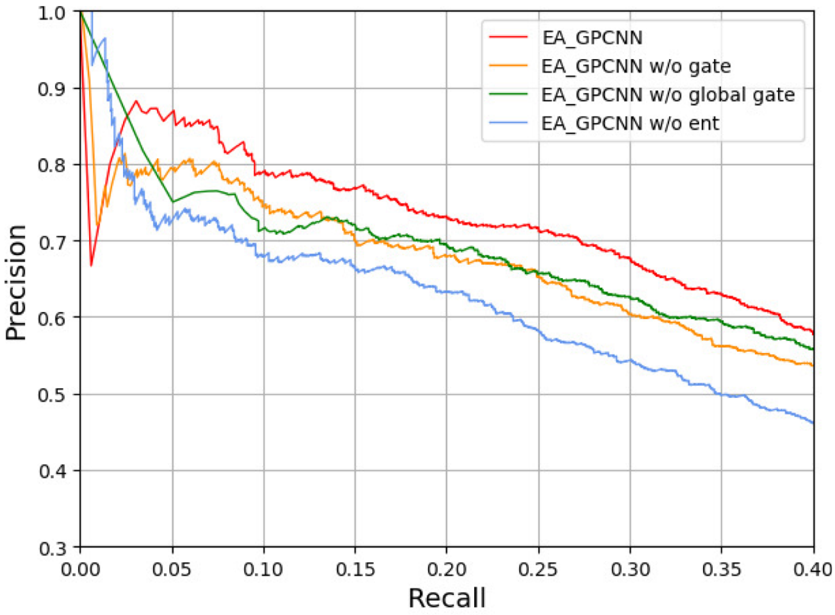

Fig. 5. Precision-recall curves for variant models of ablation study.

Table 5
Ablation study regarding precision-recall AUC value.

Approach	AUC
EA-GPCNN	0.44
EA-GPCNN w/o gate	0.41
EA-GPCNN w/o global gate	0.42
EA-GPCNN w/o ent	0.38

EA-GPCNN w/o ent adopts the traditional word representation in Zeng et al. (2014) to encode the input sentence. The mean P@N of this model decreases by 11.2%, and its AUC decreases from 0.44 to 0.38. This proves that the effective integration strategies of entity embedding and position embedding proposed in our model are able to capture the semantic dependency between each word and entity pair, which can improve the performance of distantly supervised relation extraction.

5.4.4. Case study

Table 6 shows two examples to analyze the effects of the segment-level gate mechanism. We randomly select two bags containing only one instance to compare our proposed EA-GPCNN with its variant which removes the segment-level gate mechanism that is utilized to the output of PCNN. Red texts indicate that this part contains some important information which is relevant to the predicted relation. For example, as shown in Bag 1, the segment between the head entity "**Washington Heights**" and the tail entity "**New York City**" contains the most important information of the relation */location/neighborhood/neighborhood_of*. We find that without the segment-level gate mechanism, the model will misclassify the Bag 1 into *NA*. A similar situation also appears in Bag 2. This result demonstrates that the module of the segment-level gate mechanism can improve the performance of RE.

5.5. Extended experiments

To prove the effectiveness of our proposed model, we carry out experiments on two datasets of sentence-level supervised RE tasks:

Table 6
A case study of gate mechanism on the three segments of the PCNN.

Bag	Sentence	Relation	EA-GPCNN	EA-GPCNN w/o gate
B1	Most of them also grew up in **New York City** neighborhoods like Hell 's Kitchen, Forest Hills, **Washington Heights** and ...	*/location/neighborhood/neighborhood_of*	Correct	Wrong
B2	... as a promising Alpine racer at the same **Idaho** mountains where the Olympic medalists **Picabo Street** and Christin Cooper had been groomed.	*/people/person/place_lived*	Correct	Wrong

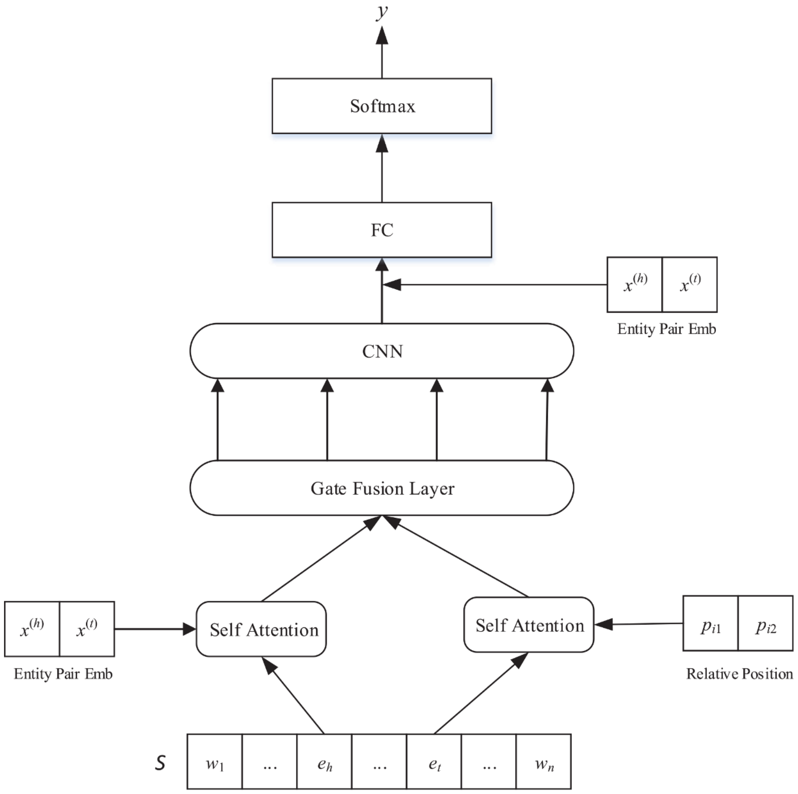

Fig. 6. The architecture of our model used for SemEval-2010 Task 8.

SemEval-2010 Task 8 (Hendrickx et al., 2010) and Wiki80 (Han et al., 2019).

5.5.1. Extended experiment 1 on SemEval-2010

SemEval-2010 Task 8 is a well-known public dataset for supervised relation extraction. It contains 10,717 annotated sentences including 8000 training instances and 2717 test instances. The dataset has 19 relations, which consists of 9 relations with two directions and undirected *Other* class. We adopt the official evaluation metric of SemEval-2010 Task 8, which is based on the macro-averaged F1-score (excluding *Other*), and takes into consideration the directionality.

We modify the entity-aware enhanced word representation module described in Section 4 with a novel gated fusion layer. Fig. 6 shows the architecture in detail. More specifically, for each word in a sentence $s = \{w_1, w_2, ..., w_n\}$, we concatenate its word embedding and position embedding to get its first representation $x_i^{(wp)} = [x_i; p_{i1}; p_{i2}]$. Then, we also concatenate its word embedding and head entity and tail entity to get its second representation $x_i^{(we)} = [x_i; x^{(h)}; x^{(t)}]$. The input sequences $X^{(we)} = [x_1^{(we)}, x_2^{(we)},...,x_n^{(we)}]$ and $X^{(wp)} = [x_1^{(wp)}, x_2^{(wp)},...,x_n^{(wp)}]$ are respectively fed into a self-attention layer (multi-head self-attention with one head) to obtain two self-attention representations S_E and S_P. In order to integrate the entity information into each word embedding, we propose a gate mechanism to fuse the information of S_E and S_P and we call it gated fusion layer. The gated fusion layer is calculated as follows:

$$R = sigmoid(W^E S_E + W^P S_P + b^R) \tag{19}$$

$$U = \tanh(W^{ue} S_E + W^{up} S_P + b^U) \tag{20}$$

Table 7
Comparison with previous results on SemEval-2010 Task 8.

Model	F1
SVM (Rink & Harabagiu, 2010)	82.2
CNN (Zeng et al., 2014)	82.7
RNN (Zhang & Wang, 2015)	82.5
SDP-LSTM (Xu et al., 2015)	83.7
Att-BLSTM (Zhou et al., 2016)	84.0
CR-CNN (dos Santos et al., 2015)	84.1
Our Model	**84.4**

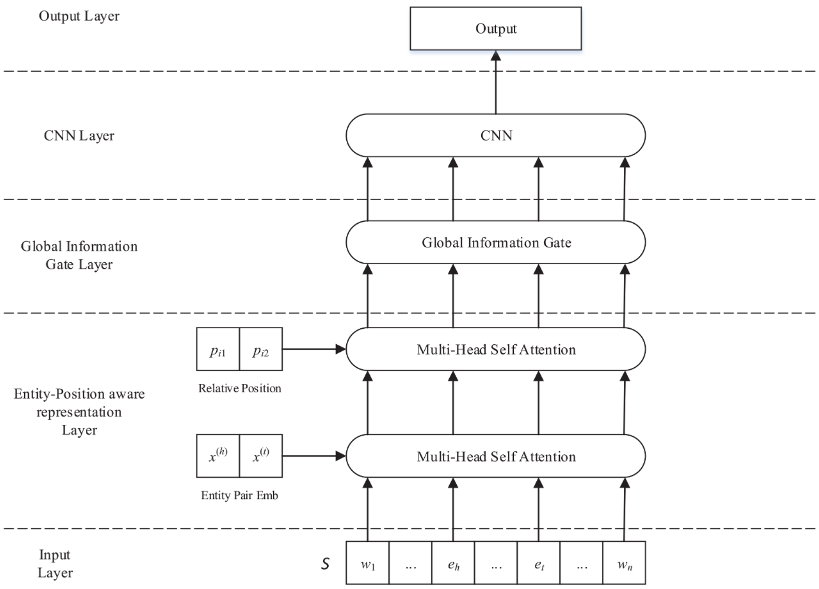

Fig. 7. The architecture of our model used for Wiki80.

$$H_g = S_P*(1 - R) + U*R \tag{21}$$

where W^E, W^P, W^{ue}, W^{up}, b^R and b^U are trainable parameters. H_g is the entity-aware enhanced word representation and is fed into the CNN to obtain the sentence representation S_R. Then, we concatenate S_R and the entities embedding $x^{(h)}$, $x^{(t)}$, which is fed into a softmax classifier to get the prediction result.

Table 7 compares our model with some state-of-the-art models on SemEval-2010 Task 10 dataset. Our model achieves an F1-score of 84.4% which outperforms SVM (Rink & Harabagiu, 2010), CNN (Zeng et al., 2014), RNN (Zhang & Wang, 2015), SDP-LSTM (Xu et al., 2015), Att-BLSTM (Zhou et al., 2016) and CR-CNN (dos Santos, Xiang, & Zhou, 2015).

5.5.2. Extended experiment 2 on Wiki80

Wiki80 is a new dataset which is derived from a large scale few-shot dataset, FewRel (Han et al., 2018b). It contains 80 relations and 56,000 instances. Since Wiki80 is not an official benchmark, we directly report the results on the validation set following (Han et al., 2019) and take accuracy as the evaluation metric of Wiki80.

In this task, we adopt the entity-aware enhanced word representation module, the global gate structure module described in Section 4 and combine them with a convolutional neural network to form our model for SemEval-2010 Task 8. The detail of the model is shown in Fig. 7. We compare our model with the CNN-based model (Zeng et al., 2014), Multi-Window CNN (Nguyen & Grishman, 2015) and Att-BLSTM model (Zhou et al., 2016). As shown in Table 8, our model has the highest accuracy 79.33% among the above models and improves performance of 4.33% than the CNN-based model, 0.85% than Multi-Window CNN model and 7.19% than Att-BLSTM model. These results demonstrate that the entity-aware enhanced word representation module and the global gate structure module can also applied to other supervised RE tasks and achieve a better result.

6. Conclusion

In this paper, we propose a gated piecewise convolutional neural network with entity-aware enhancement for distantly supervised relation extraction. First, it combines entity embedding and relative position embedding through two multi-head self-attention layers, enabling the model to capture the semantic dependency between each word and entity pair, which produces a robust word representation for downstream modules. Then, we use the global gate structure to obtain the global features of the input sentence and integrate them into each entity-aware enhanced word representation. Finally, the framework introduces another gate mechanism to

Table 8
Accuracies of various models on Wiki80.

Model	Accuracy
CNN (Zeng et al., 2014)	75.00
Multi-Window CNN (Nguyen & Grishman, 2015)	78.48
Att-BLSTM (Zhou et al., 2016)	72.14
Our Model	**79.33**

assign different weights to the three segments and highlights the effects of crucial segments. Experiments are conducted on the NYT dataset and prove that our model outperforms most of the state-of-the-art models.

In the future, we will explore more efficient and lightweight encoding methods that dynamically integrate entity and position information to produce a more expressive word representation. In addition, we will introduce some external information such as entity type and entity description information to our model to reduce the influence of the noisy labeling problem.

CRediT authorship contribution statement

Haixu Wen: Investigation, Visualization, Methodology, Writing - original draft, Software. **Xinhua Zhu:** Conceptualization, Data curation, Writing - review & editing, Formal analysis. **Lanfang Zhang:** Funding acquisition, Project administration, Supervision. **Fei Li:** Software, Validation.

Acknowledgements

This work was supported in part by the National Natural Science Foundation of China under the contract number 61967003, the Natural Science Foundation of Guangxi of China under the contract number 2020GXNSFAA159082, Research Fund of Guangxi Key Lab of Multi-source Information Mining & Security (No. 20-A-03-02) and the Innovation Project of School of Computer Science and Information Engineering, Guangxi Normal University under the contract number JXXYYJSCXXM-005.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.ipm.2020.102373.

References

Bengio, Y., Louradour, J., Collobert, R., & Weston, J. (2009). *Curriculum learning. International conference on machine learning, ICML*.

Bollacker, K., Evans, C., Paritosh, P., Sturge, T., & Taylor, J. (2008). *Freebase: a collaboratively created graph database for structuring human knowledge. Proceedings of the 2008 ACM SIGMOD international conference on management of data*1247–1250.

Culotta, A., & Sorensen, J. (2004). *Dependency tree kernels for relation extraction. Proceedings of the 42nd annual meeting of the association for computational linguistics (ACL-04), Barcelona, Spain*423–429. https://doi.org/10.3115/1218955.1219009.

De Sa, C., Ratner, A., Ré, C., Shin, J., Wang, F., Wu, S., & Zhang, C. (2016). Deepdive: Declarative knowledge base construction. *ACM SIGMOD Record, 45*(1), 60–67.

dos Santos, C., Xiang, B., & Zhou, B. (2015). *Classifying relations by ranking with convolutional neural networks. Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing (volume 1: Long papers)*. Beijing, China: Association for Computational Linguistics626–634. https://doi.org/10.3115/v1/P15-1061.

Du, J., Han, J., Way, A., & Wan, D. (2018). *Multi-level structured self-attentions for distantly supervised relation extraction. Proceedings of the 2018 conference on empirical methods in natural language processing*. Brussels, Belgium: Association for Computational Linguistics2216–2225. https://doi.org/10.18653/v1/D18-1245.

Feng, J., Huang, M., Zhao, L., Yang, Y., & Zhu, X. (2018). *Reinforcement learning for relation classification from noisy data. Thirty-second AAAI conference on artificial intelligence*.

Han, X., Gao, T., Yao, Y., Ye, D., Liu, Z., & Sun, M. (2019). *OpenNRE: An open and extensible toolkit for neural relation extraction. Proceedings of EMNLP-IJCNLP: System demonstrations*169–174. https://doi.org/10.18653/v1/D19-3029.

Han, X., Yu, P., Liu, Z., Sun, M., & Li, P. (Yu, Liu, Sun, Li, 2018a). *Hierarchical relation extraction with coarse-to-fine grained attention. Proceedings of the 2018 conference on empirical methods in natural language processing*2236–2245.

Han, X., Zhu, H., Yu, P., Wang, Z., Yao, Y., Liu, Z., & Sun, M. (Zhu, Yu, Wang, Yao, Liu, Sun, 2018b). *FewRel: A large-scale supervised few-shot relation classification dataset with state-of-the-art evaluation. Proceedings of the 2018 conference on empirical methods in natural language processing*. Brussels, Belgium: Association for Computational Linguistics4803–4809. https://doi.org/10.18653/v1/D18-1514.

Hendrickx, I., Kim, S. N., Kozareva, Z., Nakov, P., Ó Séaghdha, D., Padó, S., ... Szpakowicz, S. (2010). *SemEval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. Proceedings of the 5th international workshop on semantic evaluation*. Uppsala, Sweden: Association for Computational Linguistics33–38 URL https://www.aclweb.org/anthology/S10-1006

Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., & Weld, D. S. (2011). *Knowledge-based weak supervision for information extraction of overlapping relations. Proceedings of the 49th annual meeting of the association for computational linguistics: Human language technologies-volume 1*. Association for Computational Linguistics541–550.

Hu, L., Zhang, L., Shi, C., Nie, L., Guan, W., & Yang, C. (2019). *Improving distantly-supervised relation extraction with joint label embedding. Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics3821–3829. https://doi.org/10.18653/v1/D19-1395.

Huang, Y., & Du, J. (2019). *Self-attention enhanced CNNs and collaborative curriculum learning for distantly supervised relation extraction. Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics389–398.

Ji, G., Liu, K., He, S., & Zhao, J. (2017). *Distant supervision for relation extraction with sentence-level attention and entity descriptions. Thirty-first AAAI conference on artificial intelligence*.

Kipf, T. N., & Welling, M. (2017). *Semi-supervised classification with graph convolutional networks. International conference on learning representations (ICLR)*.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature, 521*(7553), 436–444.

Li, X., Taheri, A., Tu, L., & Gimpel, K. (2016). *Commonsense knowledge base completion. Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)*1445–1455.

Li, X., Yin, F., Sun, Z., Li, X., Yuan, A., Chai, D., ... Li, J. (2019). *Entity-relation extraction as multi-turn question answering. Proceedings of the 57th annual meeting of the association for computational linguistics*. Florence, Italy: Association for Computational Linguistics1340–1350. https://doi.org/10.18653/v1/P19-1129.

Lin, Y., Shen, S., Liu, Z., Luan, H., & Sun, M. (2016). *Neural relation extraction with selective attention over instances. Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)*2124–2133.

Liu, T., Wang, K., Chang, B., & Sui, Z. (2017). *A soft-label method for noise-tolerant distantly supervised relation extraction. Proceedings of the 2017 conference on empirical methods in natural language processing*1790–1795.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv:1301.3781.

Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009). *Distant supervision for relation extraction without labeled data. Proceedings of the joint conference of the 47th annual meeting of the acl and the 4th international joint conference on natural language processing of the AFNLP: Volume 2*. Association for Computational

Linguistics1003–1011.

Mooney, R. J., & Bunescu, R. C. (2006). Subsequence kernels for relation extraction. In Y. Weiss, B. Schölkopf, & J. C. Platt (Eds.). *Advances in neural information processing systems 18* (pp. 171–178). MIT Press.

Nguyen, T. H., & Grishman, R. (2015). *Relation extraction: Perspective from convolutional neural networks. Proceedings of the 1st workshop on vector space modeling for natural language processing*. Denver, Colorado: Association for Computational Linguistics39–48. https://doi.org/10.3115/v1/W15-1506.

Qin, P., Xu, W., & Wang, W. Y. (Xu, Wang, 2018a). *DSGAN: Generative adversarial training for distant supervision relation extraction. Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*. Melbourne, Australia: Association for Computational Linguistics496–505. https://doi.org/10.18653/v1/P18-1046.

Qin, P., Xu, W., & Wang, W. Y. (Xu, Wang, 2018b). *Robust distant supervision relation extraction via deep reinforcement learning. Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*. Melbourne, Australia: Association for Computational Linguistics2137–2147. https://doi.org/10.18653/v1/P18-1199.

Riedel, S., Yao, L., & McCallum, A. (2010). *Modeling relations and their mentions without labeled text. Joint European conference on machine learning and knowledge discovery in databases*. Springer148–163.

Rink, B., & Harabagiu, S. (2010). *Utd: Classifying semantic relations by combining lexical and semantic resources. Proceedings of the 5th international workshop on semantic evaluation*. Association for Computational Linguistics256–259 URL http://dblp.uni-trier.de/db/conf/semeval/semeval2010.html#RinkH10

Ru, C., Tang, J., Li, S., Xie, S., & Wang, T. (2018). Using semantic similarity to reduce wrong labels in distant supervision for relation extraction. *Information Processing & Management, 54*(4), 593–608.

Srivastava, N., Hinton, G., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. (2014). Dropout: A simple way to prevent neural networks from overfitting. *The Journal of Machine Learning Research, 15*(1), 1929–1958.

Sun, H., Ma, H., Yih, W.-t., Tsai, C.-T., Liu, J., & Chang, M.-W. (2015). *Open domain question answering via semantic enrichment. Proceedings of the 24th international conference on world wide web*1045–1055.

Surdeanu, M., Tibshirani, J., Nallapati, R., & Manning, C. D. (2012). *Multi-instance multi-label learning for relation extraction. Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning*. Association for Computational Linguistics455–465.

Vashishth, S., Joshi, R., Prayaga, S. S., Bhattacharyya, C., & Talukdar, P. (2018). *RESIDE: Improving distantly-supervised neural relation extraction using side information. Proceedings of the 2018 conference on empirical methods in natural language processing*. Brussels, Belgium: Association for Computational Linguistics1257–1266. https://doi.org/10.18653/v1/D18-1157.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). *Attention is all you need. Advances in neural information processing systems*5998–6008.

Wu, F., & Weld, D. S. (2010). *Open information extraction using wikipedia. Proceedings of the 48th annual meeting of the association for computational linguistics*. Association for Computational Linguistics118–127.

Xu, D., Ji, J., Huang, H., Deng, H., & Li, W.-J. (2019). Gated group self-attention for answer selection. arXiv:1905.10720.

Xu, Y., Mou, L., Li, G., Chen, Y., Peng, H., & Jin, Z. (2015). *Classifying relations via long short term memory networks along shortest dependency paths. Proceedings of the 2015 conference on empirical methods in natural language processing*. Lisbon, Portugal: Association for Computational Linguistics1785–1794. https://doi.org/10.18653/v1/D15-1206.

Ye, H., & Luo, Z. (2019). Deep ranking based cost-sensitive multi-label learning for distant supervision relation extraction. *Information Processing & Management,* 102096.

Ye, Z.-X., & Ling, Z.-H. (2019). Distant supervision relation extraction with intra-bag and inter-bag attentions. arXiv:1904.00143.

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

Zeng, D., Liu, K., Chen, Y., & Zhao, J. (2015). *Distant supervision for relation extraction via piecewise convolutional neural networks. Proceedings of the 2015 conference on empirical methods in natural language processing*1753–1762.

Zeng, D., Liu, K., Lai, S., Zhou, G., & Zhao, J. (2014). *Relation classification via convolutional deep neural network. Proceedings of COLING 2014, the 25th international conference on computational linguistics: Technical papers*. Dublin, Ireland: Dublin City University and Association for Computational Linguistics2335–2344.

Zhang, D., & Wang, D. (2015). Relation classification via recurrent neural network. abs/1508.01006.

Zhou, P., Shi, W., Tian, J., Qi, Z., Li, B., Hao, H., & Xu, B. (2016). *Attention-based bidirectional long short-term memory networks for relation classification. Proceedings of the 54th annual meeting of the association for computational linguistics (volume 2: Short papers)*207–212.