Attention-based Gated Convolutional Neural Networks for Distant Supervised Relation Extraction

Xingya Li, Yufeng Chen, Jinan Xu and Yujie Zhang

Beijing Jiaotong University, Beijing, China
{17120394, chenyf, jaxu, yjzhang}@bjtu.edu.com

Abstract. Distant supervision is an effective method to generate large-scale labeled data for relation extraction without expensive manual annotation, but it inevitably suffers from the wrong labeling problem, which would make the corpus much noisy. However, the existing research work mainly focuses on sentence-level noise filtering, without considering noisy words which widely exist inside sentences. In this paper, we propose an attention-based gated piecewise convolutional neural networks (AGPCNNs) for distant supervised relation extraction, which can effectively reduce word-level noise by selecting the inner-sentence features. On the one hand, we construct a piecewise convolutional neural network with gate mechanism to extract features that are related to relations. On the other hand, we employ a soft-label strategy to enable model to select important features automatically. Furthermore, we adopt an attention mechanism after the piecewise pooling layer to obtain high-level positive features for relation predicting. Experimental results show that our method can effectively filter word-level noise and outperforms all baseline systems significantly.

Keywords: Relation Extraction, Distant Supervision, Gate Mechanism, Attention Mechanism.

1 Introduction

Relation extraction (RE) aims to identify the semantic relationship between two entities from natural language texts, and it is an important part of information extraction. One of the major challenges faced by RE is that its training requires large-scale labeled corpus, while manual annotation is too time-consuming and laborious. Thus, Mintz et al. (2009) proposed a distant supervised method for RE, which could annotate large-scale data automatically and heuristically with the existing knowledge bases. The labeling process is as follows: given a triplet in a knowledge base, also known as a relation fact, (*Syracuse, contains, Lake Onondaga*), all sentences containing the above two named entities will be labeled as relation *contains*.

Distant supervision is an effective method of automatically labeling training data, nevertheless, it is plagued by the wrong labeling problem (Riedel et al., 2010), since a sentence that mentions two entities does not necessarily express the relation contained in a known knowledge base. For example, in the sentence *[The Onondaga nation is an 11-square-mile parcel in a valley south of Syracuse and about eight miles from*

Onondaga Lake]. There is no *contains* relation between the entities *Syracuse* and *Onondaga Lake*, but it will still be regarded as a positive instance. In response, some research work adopted the Multi-instance Learning (MIL) method (Dietterich et al., 1997), and used a probabilistic graphical model to select sentences (Hoffmann et al., 2011; Suideanu et al., 2012). Zeng et al. (2015) combined MIL and piecewise convolutional neural networks (PCNNs) to select the most likely positive sentence. Lin et al (2016) introduced a sentence-level attention mechanism on the basis of PCNNs, and extracted effective features of all sentences by assigning different attention weights to sentences. Moreover, Ji et al (2017) proposed a model architecture called APCNNs, which also used attention mechanism and added entity description information, and achieved the best performance in such methods.

Although the above researches have achieved good results in distant supervised relation extraction, there still exists two problems. Specifically, (1) Not all words in a sentence contribute to judging relation labels. For example, considering a sentence *[The cultural appreciation was driven by literacy theories like Roland Barthes, not by Jean Baudrillard]*, where *Roland Barthes* and *France* are two corresponding entities. Obviously, the sentence describes the *national* relation between *Roland Barthes* and *France*, but the sub-sentence *[not by Jean Baudrillard]* has little effect on judging the relation *national*, regarded as noisy words or word-level noise, of which the features will cut down the precision of the relation extraction model. (2) The Hard-label method, in which the labels of entity pairs are immutable during model training, would enlarge the impact of wrong labels in distant supervision.

In this paper, we propose a novel word-level distant supervised method for relation extraction, named AGPCNNs (Attention-based Gated Piecewise Convolutional Neural Networks), which reduces word-level noise by selecting important features inside sentences. To settle the above first problem, a gate mechanism (Hochreiter et al., 1997) is used to screen out the features extracted by convolution layer. Besides, an attention mechanism is applied after piecewise pooling layer to gain the high-level positive features. To tackle the second problem, we introduce a soft-label strategy (Wang et al., 2018) in our model, by which the same instances may have different labels in different epochs of training. Finally, combined with a sentence-level noise filtering module, more positive correlation features are obtained. We tested on the public data sets, 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 contributions are summarized as follows:

- To handle the problem of noisy words, a gate mechanism is proposed to filter inner-sentence features uncorrelated to relation labels and an attention mechanism after piecewise pooling layer to obtain high-level features related to relation labels.
- A soft-label strategy is utilized to weaken the impact of hard labels on feature selection during training. Specifically, we use the bilinear transformation between entity pairs to conduct the feature selection process in the gate mechanism, which makes it more precise and suitable for reducing word-level noise.

- The proposed model achieves significant results for distant supervised relation extraction. Furthermore, the gate mechanism could be adopted by other neural networks and enhance the performance of the corresponding tasks.

2 Methodology

Given a bag (a set of sentences containing the same entity pair) $B = \{s_1, s_2, \cdots, s_n\}$ and the corresponding two entities, our model will predict the probability of each relation label on *B*. The overall architecture of our proposed model is illustrated in Figure 1.

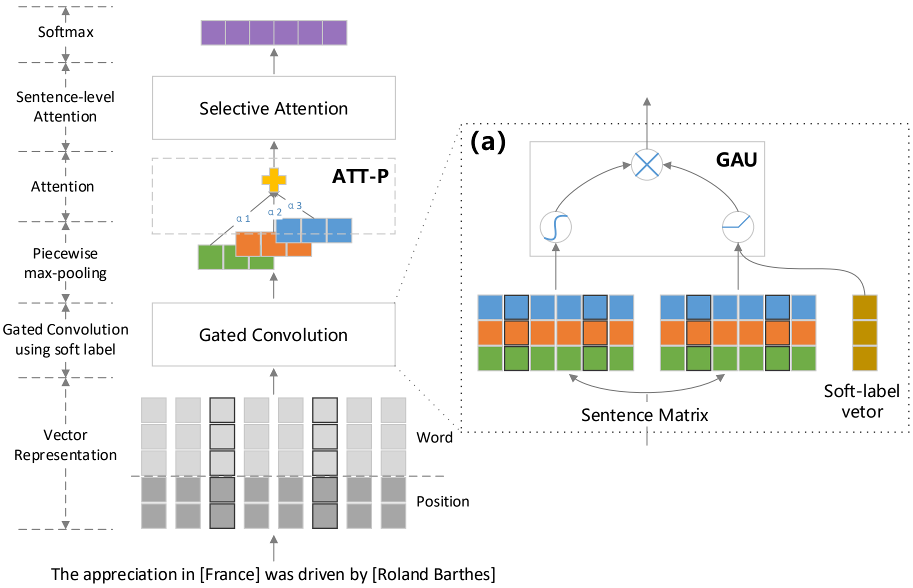

Fig. 1. The architecture of our model (AGPCNNs) used for distant supervised relation extraction. The right part (a) describes the Gated Convolution layer, which takes soft labels as the supervised information.

Compared with the traditional PCNNs model, we improve the convolution layer by adding a gate mechanism and the piecewise pooling layer by adding an attention mechanism to select important inner-sentence features. Furthermore, we adopt a soft-label strategy for the above two mechanisms, so that they can work better. We will describe the gate mechanism, the soft-label strategy and the attention mechanism in detail in Section 2.2-2.4.

2.1 Vector Representation

The input of AGPCNNs is represented by embeddings, which are composed of two parts: word embeddings and position 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 vector matrix **V** (or lookup table), where $\mathbf{V} \in \mathbb{R}^{|\mathbf{V}| \times d^{\mathrm{w}}}$, |**V**| is the size of the matrix **V** and d^{w} is the dimension of word embeddings.

Position Embeddings. Following Zeng et al. (2015), 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 *theorists* to *France* (e_1) and *Roland Barthes* (e_2) are 4 and -2 respectively. With two position matrices $\mathbf{PF}_1$ and $\mathbf{PF}_2$, which are initialized randomly, we can transfer relative distances to real-value vectors $\mathbf{E}_{\mathrm{p}} \in \mathbb{R}^{d^{\mathrm{p}}}$, where d^{p} is the dimension.

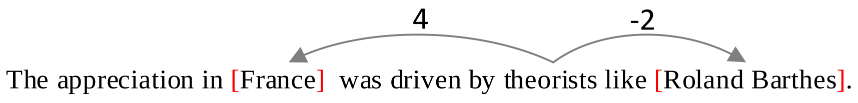
4
-2
The appreciation in [France] was driven by theorists like [Roland Barthes].

Fig. 2. An example of relative distances

Finally, the input representation of a word is a vector concatenated by word embeddings and position embeddings. With the vector representation of words, we transfer the sentence *s* into a matrix $\mathbf{S} \in \mathbb{R}^{|s| \times d}$, where |*s*| is the length of *s*, $d = d^{\mathrm{w}} + d^{\mathrm{p}}$.

2.2 Convolution with Gate Mechanism

Convolution. Convolutional neural networks (CNNs) can effectively extract all local features of the input and perform global predictions.

Convolution is an operation between a weight matrix (also known as filter) $\boldsymbol{w}$ and the vector matrix **S** of a sentence *s*. The result of convolution operation is $\boldsymbol{c} = \{c_1, c_2, \cdots, c_{|s|-w+1}\}$, where w is the window size, and $c_j = f(\boldsymbol{w} \otimes \mathbf{S}_{(j-w+1):j} + b_{\mathrm{c}})$, where $1 \leq j \leq |s| - w + 1$, f is a nonlinear activation function and $b_{\mathrm{c}} \in \mathbb{R}$ is bias.

Gate Mechanism. Considering the impact of inner-sentence noise on the model performance, we use a gate mechanism to select positive features at word level. Gate mechanism have shown effectiveness of gate mechanisms in language modeling (Kalchbrenner et al, 2016; Gehring et al, 2017). 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:

$$c_{\mathrm{GAU}} = \tanh(\boldsymbol{w}_{\mathrm{c}} \otimes \mathbf{S} + b_1) \times \mathrm{relu}(\boldsymbol{v} \otimes \mathbf{S} + b_2) \tag{1}$$

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.

2.3 Soft-label Strategy

Generally, the relation labels of entity pairs are unchangeable during training, no matter whether they are correct 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 inner-sentence 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 $\boldsymbol{l}_{\mathrm{relation}}=\boldsymbol{e}_1\mathbf{W}_{\mathrm{B}}\boldsymbol{e}_2$ as the soft label between the two entities $(\boldsymbol{e}_1,\boldsymbol{e}_2)$ to help model to select important features. Specifically, we obtain the feature c_j by:

$$m_j = \mathrm{relu}(\boldsymbol{w}_m \otimes \mathbf{S}_{(j-w+1):j} + \boldsymbol{l}_{\mathrm{relation}} + b_m) \tag{2}$$

$$n_j = \tanh(\boldsymbol{w}_n \otimes \mathbf{S}_{(j-w+1):j} + b_n) \tag{3}$$

$$c_j = m_j \times n_j \tag{4}$$

The ability to capture different features typically requires the use of multiple filters in the convolution, so we use n filters $\mathbf{W} = \{\boldsymbol{w}_1, \boldsymbol{w}_2, \cdots, \boldsymbol{w}_n\}$. The convolution result is:

$$c_{ij} = m_{i,j} \times n_{i,j} \tag{5}$$

where $1 \le i \le n$, $1 \le j \le |s| - w + 1$. The overall output result of the gated convolution layer is $\mathbf{C} = \{\boldsymbol{c}_1, \boldsymbol{c}_2, \cdots, \boldsymbol{c}_n\}$

2.4 Pooling

Piecewise Max Pooling. Max pooling operation is usually used to extract the most dominant features in feature maps, but ignores the structure information and fine-grained information. Thus, PCNNs divides an instance into three segments according to the given entity pair and does max pooling operation on each segment. For each convolutional result $\mathbf{C}_i$, it can be divided into $\mathbf{C}_i = \{\boldsymbol{c}_{i,1}, \boldsymbol{c}_{i,2}, \boldsymbol{c}_{i,3}\}$, then the piecewise max pooling process is defined as $p_{ij} = \max(\boldsymbol{c}_{i,j})$, where $1 \le i \le n$, $j = 1,2,3$. We concatenate all vectors $\boldsymbol{p}_i = [p_{i,1}, p_{i,2}, p_{i,3}]$ as $\mathbf{P} \in \mathbb{R}^{3n}$.

Attention Mechanism. To select the positive feature more precisely, we propose an attention mechanism to help our model focus on positive features at high level. Inspired by Wu et al. (2018), we adopt the attention mechanism after the piecewise max pooling layer (denoted as ATT-P) to obtain more positive high-level features. For getting more positive related features, we also adopt the soft-label strategy here:

$$\alpha_{\mathrm{H}}{}^{j} = \frac{\exp(\eta^{j})}{\sum_k \exp(\eta^{k})} \tag{6}$$

$$\eta^{j} = \boldsymbol{p}_i^{j} \mathbf{A}(e_1 \mathbf{W}_{\mathrm{B}} e_2) \tag{7}$$

where $\mathbf{A}$ and $\mathbf{W}_{\mathrm{B}}$ is weighted matrices. Then the representation $\boldsymbol{\gamma} \in \mathbb{R}^{3n}$ of the sentence s is $\boldsymbol{\gamma} = \sum_{j=1}^{3} \alpha_{\mathrm{H}}{}^{j} \boldsymbol{p}_i^{j}$.

Finally, we obtain the feature vector $\boldsymbol{b}_s \in \mathbb{R}^{3n}$ of the sentence s from:

$$\boldsymbol{b}_s = \tanh(\boldsymbol{\gamma}) \tag{8}$$

2.5 Sentence-level Attention and Output

In previous studies, bilinear and nonlinear attention mechanisms have been proved helpful to model performance. Considering the computational efficiency and effectiveness, we adopt the non-linear form in our method.

We also use $\boldsymbol{l}_{\mathrm{relation}} = \boldsymbol{e}_1 \mathbf{W}_{\mathrm{B}} \boldsymbol{e}_2$ as relation labels between two entities. For the feature vector $\boldsymbol{b}_i^{j}$ of the j-th sentence in the i-th bag B_i, the corresponding attention weight α^{j} is calculated as follows:

$$\alpha^{j} = \frac{\exp(\omega^{j})}{\sum_k \exp(\omega^{k})} \tag{9}$$

$$\omega^{j} = \mathbf{W}_a^{T}(\tanh[\boldsymbol{b}_i^{j}; \boldsymbol{l}_{relation}]) + b_a \tag{10}$$

The final feature vector of B_i is expressed as $\boldsymbol{r}_i = \sum_{j=1}^{|B_i|} \alpha^{j} \boldsymbol{b}_i^{j}$.

The vector representation of the bag is then fed to the softmax classifier to predict the final relation labels and calculate the cross-entropy objective function on all training bags (T):

$$J(\theta) = -\sum_{i=1}^{T} \log p(y_i \mid \boldsymbol{r}_i; \theta) \tag{11}$$

3 Experiments

3.1 Datasets and Evaluation Metrics

We evaluate our approach on a widely used dataset which is developed by Riedel et al. (2010). 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 522,611 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 show the aggregated P-R curve (Precision/recall Curve) and Precision@N (precision at top n predictions) in the experiments.

3.2 Parameter Settings

In our experiments, we use the word2vec tool (Mikolov et al., 2013) to pre-train 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}, batch size among {50, 100, 160}, the learning rate λ among {0.001, 0.01, 0.1, 0.5}. The best configurations are: $d^{w} = 50$, $d^{p} = 5$, $w = 3$, n = 300, $\lambda = 0.1$, the batch size is 50. We use dropout strategy and Adadelta to train our models. According to experience, the dropout rate is fixed to 0.5.

3.3 Performance Evaluation

We compare our method with three previous works: PCNNs+MIL(Zeng et al., 2015) selects the sentence with the highest score as the representation of a bag; PCNNs+ATT (Lin et al., 2016) and APCNNs (Ji et al., 2017) use sentence-level bilinear and nonlinear form attention to synthesize all sentences' information in a bag as it's representation. Moreover, we add GAU module to PCNNs+MIL (denoted as PCNNs+MIL+GAU) to verify the effectiveness of GAU. In addition, we remove ATT-P module (denoted as GPCNNs) to testify the contribution of ATT-P to model. At the same time, we replace GAU with GTU (denoted as PCNNs + GTU) to prove the validity of our soft-label strategy. Figure 3 shows the aggregated P-R curves, and Table 1 shows the Precision@N with *N* = {100, 200, 300} of our approach and all the baselines.

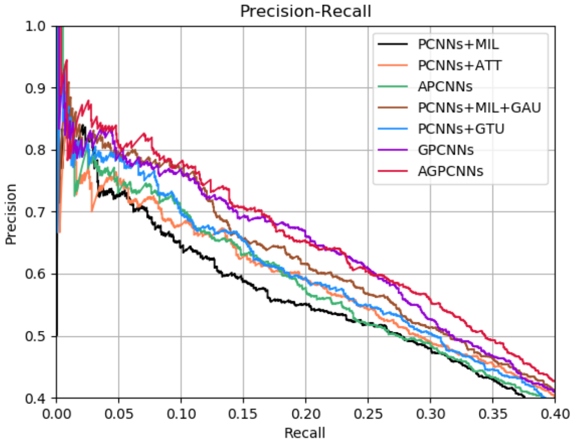

Fig. 3. Aggregate precision / recall curves for AGPCNNs and all the baseline models. For the sake of clarity, we show all the curves with different colors and bold lines.

Table 1. Precision@N of our model and all the baseline models.

Precision@N(%)	Top 100	Top 200	Top 300	Average
PCNNs+MIL	72.89	69.23	64.05	68.72
PCNNs+ATT	74.26	72.14	68.44	72.61
APCNNs	76.24	74.13	69.44	73.27
PCNNs+MIL+GAU	81.19	77.61	74.42	77.74
PCNNs+GTU	78.22	76.62	68.77	74.54
GPCNNs	83.16	77.11	74.09	78.12
AGPCNNs	**83.17**	**79.10**	**75.08**	**79.12**

From Fig 3 and Table 1 we have the following observations:

1) AGPCNNs achieves the best P-R curve over baselines. And its Precision@N values are the highest, which are about 5% higher than baselines on average. This indicates that AGPCNNs is effective because GAU and ATT-P module can select more important inner-sentence features at fine-grained level.
2) All the models with GAU outperform that without GAU. It demonstrates that sentence-level noise filtering methods combined with the gate mechanism can obtain more positive features than those denoising at sentence level only, which verifies the effectiveness of the GAU module.
3) AGPCNNs performs much better than PCNNs+GTU. It means that the gate mechanism can select important features more precisely with the help of the soft-label strategy, and also verifies the bilinear transformation between entity pairs can map their relation effectively.
4) Integrated with ATT-P module, AGPCNNs achieves more improvements than GPCNNs. It shows that ATT-P can obtain helpful high-level global features and resist noise further.

3.4 Effectiveness of GAU Module

In order to verify the effectiveness of the gate mechanism in word-level feature selection more intuitively, we 1) remove all the attention mechanisms of AGPCNNs (denoted as PCNNs+GAU); 2) apply GAU to CNNs (we name it as CNNs+GAU). CNNs and PCNNs are used as baselines in this experiment. Note that these methods are sentence-level extraction, different from bag-level extraction in section 3.1. The experimental results are shown in Figure 4 below:

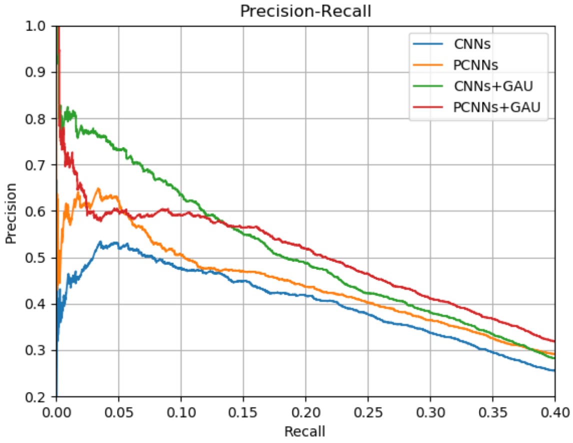

Fig. 4. Aggregate precision/recall curves for a variety of models on sentence-level extraction.

From Fig 4 we can see:

1) Compared with PCNNs and CNNs, PCNNs+GAU and CNNs+GAU have achieved significant improvement, indicating that GAU module can effectively improve the performance of the model with different pooling mechanisms, verifying the effectiveness of the gate mechanism and reflecting the robustness of the GAU module.
2) The performance of PCNNs+GAU is significantly improved compared with PCNNs, indicating that the word-level noise has a great impact on the performance of the RE models, and also demonstrating the GAU module can effectively filter word-level noise by obtaining more important features.

3.5 Case Study

To explicitly illustrate the adverse impact of word-level noise on feature selection and the effectiveness of our proposed model, we show an example of attention weights in a bag during testing. As shown in Table 2, all sentences contain the phrase *[president emeritus of]*, which clearly indicates the */company* relation between the entities *John Brademas* and *New York University*, so they are all positive instances.

Table 2. An example of attention weights. The bold strings are head/tail entities and the underlined strings are keywords to predict the relation. The relation */company* corresponds the */business/person/company* in Freebase.

Triplet	Instances	APCNNs	AGPCNNs
（John Brademas, New York University, /company）	1. Thirty-five years ago, President vetoed the legislation, refusing to encourage "the family-centered child rearing". "I don't think we've ever recovered from that veto message." said **John Brademas**, president emeritus of **New York University**, and, as a former Democratic congressman from Illinois, a sponsor of that legislation.	0.085	0.354
	2. An article on Jan.11 about a conference in New York misidentified the home state of **John Brademas**, president emeritus of **New York University**.	0.665	0.324
	3. Correction: January 25, 2006, Wednesday An article on Jan.11 about a conference in New York misidentified the home state of **John Brademas**, president emeritus of **New York University**.	0.249	0.322

It can be seen that, except the phrase *[president emeritus of]*, other words do not have direct or indirect connection with the entity pair relation */company*, which means these words are noisy. Compared with the other two sentences, the first sentence contains the most amount of noisy information, the third sentence contains the second, and the second sentence the least. Due to the lack of effective inner-sentence feature selection mechanisms, the APCNNs model assigns the higher weight to the second sentence (0.665) and the lesser weight to the first sentence and the third sentence (0.085, 0.249). However, AGPCNNs model is barely affected by noisy information, and assigns similar weights to the three sentences (0.354, 0.324, 0.322). The attention weights verify that our proposed model can effectively select more important word-level features no matter how much the noisy information sentences contains and make full use of the supervision information in a bag.

4 Related Work

Due to the high costs of manual annotation, distant supervision plays an increasingly important role in RE. However, this method faces the challenge brought by the wrong labeling problem, resulting in being prone to generating lots of noise instances. For this, Riedel et al. (2010) modeled distant supervised RE as a single labeling problem by using multi-instance learning. The following research work (Hoffmann et al., 2011; Surdeanu et al., 2012) adopted 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 troubled by the problem of error propagation.

As neural networks have been widely used and achieved good results in many tasks, Zeng et al. (2015) proposed PCNNs with MIL to select the most likely positive sentences. Lin et al. (2016) use selective attention over instance with PCNNs to select valid sentences. Ji et al. (2017) assign more precise attention weights by making use of entity descriptions. Focused on the imbalance of datasets, a label-free method has been proposed by Wang et al. (2018). Besides, reinforcement learning has been used to select the valid instances before training for relation extraction (Feng et al., 2018; Qin et al., 2018). However, all the above approaches filter noise at the sentence level, ignoring the word-level (inner-sentence) noise, resulting in the insufficient use of the supervision information in a bag. On the other hand, the fixed relation labels (hard labels) of entity pairs during training also enlarge the influence of the wrong labels.

Different from the existing researches, we propose an AGPCNNs model, which uses a gate mechanism in convolution layer and an attention mechanism after piecewise pooling layer to reduce word-level noise by selecting more important inner-sentence features. Furthermore, we also introduce a soft-label strategy (Wang et al., 2018; Kalchbrenner et al., 2016; Gehring et al., 2017) in our model by adopting bilinear transformation results of entity pairs as relation labels, making feature selection more precise. Experimental results verify the effectiveness of the AGPCNNs model. In addition, Liu et al. (2018) also proposed a word-level noise filtering method. The differences between our model and theirs lie in: 1) Liu et al. (2018) used NLP tools to build dependency subtrees, and introduced external knowledge to filter word-level noise through transfer learning. While our model only uses two mechanisms (gate mechanism and attention mechanism), and does not use external tools and knowledge; 2) The soft-label strategy is introduced into our model, which weakens the impact of wrong labels on feature selection.

5 Conclusions and Future Work

Aiming at tackling the low-quality corpus problem, we propose a novel distant supervised approach for relation extraction, named AGPCNNs, which uses gate mechanism, attention mechanism and soft-label strategy to cut down word-level noise by valid inner-sentence feature selection. The Gate mechanism can effectively select word-level features extracted by convolution layer. The Attention mechanism is adopted after piecewise pooling layer to obtain high-level global feature relevant to relation labels. The Soft-label strategy is introduced to improve the accuracy of feature selection by using bilinear transformation results between entity pairs to conduct the selection process. The experimental results show that our model is 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 authors are supported by the National Nature Science Foundation of China (Nos. 61473294, 61370130 and 61876198), the Fundamental Research Funds for the Central Universities (Nos. 2015JBM033), and the International Science and Technology Cooperation Program of China (No. K11F100010)

References

1. Mintz, M., Bills, S., Snow, R., Jurafsky, D.: Distant supervision for relation extraction without labeled data. In: 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, pp. 1003-1011 (2009).
2. Riedel, S., Yao, L., McCallum, A.: Modeling relations and their mentions without labeled text. In: Machine Learning and Knowledge Discovery in Databases, pp. 148-163 (2010).
3. Dietterich, T. G., Lathrop, R. H., Lozano-Pérez, T.: Solving the multiple instance problem with axis-parallel rectangles. Artificial intelligence 89(1-2), 31-71 (1997).
4. Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., Weld, D. S.: Knowledge-based weak supervision for information extraction of overlapping relations. In: Proceedings of ACL, pp. 541-550. Association for Computational Linguistics (2011).
5. Surdeanu, M., Tibshirani, J., Nallapati, R., Manning, C. D.: Multi-instance multi-label learning for relation extraction. In: Proceedings of the 2012 joint conference on empirical methods in natural language processing and computational natural language learning, pp. 455-465. Association for Computational Linguistics (2012).
6. Zeng, D., Liu, K., Chen, Y., Zhao, J.: Distant supervision for relation extraction via piecewise convolutional neural networks. In: Proceedings of EMNLP, pp. 1753-1762 (2015).
7. Lin, Y., Shen, S., Liu, Z., Luan, H., Sun, M.: Neural relation extraction with selective attention over instances. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 2124-2133 (2016).
8. Ji, G., Liu, K., He, S., Xu, L., Zhao, J.: Distant supervision for relation extraction with sentence-level attention and entity descriptions. In: AAAI, pp. 3060-3066 (2017).
9. Hochreiter, S., Schmidhuber, J.: Long short-term memory. Neural computation 9(8), 1735-1780 (1997).
10. Wu, W., Chen, Y., Xu, J., Zhang, Y.: Attention-Based Convolutional Neural Networks for Chinese Relation Extraction. In: Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data, pp. 147-158. Springer, Cham (2018).
11. Wang, G., Zhang, W., Wang, R., Zhou, Y., Chen, X., Zhang, W., Zhu, H., Chen, H.: Label-free distant supervision for relation extraction via knowledge graph embedding. In: Proceedings of EMNLP, pp. 2246-2255 (2018).
12. Kalchbrenner, N., Espeholt, L., Simonyan, K., Oord, A., Graves, A., Kavukcuoglu, K.: Neural machine translation in linear time. arXiv preprint arXiv, 1610.10099 (2016).
13. Gehring, J., Auli, M., Grangier, D., Yarats, D., Dauphin, Y.: Convolutional sequence to sequence learning. In: Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, pp. 1243-1252 (2017).
14. Mikolov, T., Chen, K., Corrado, G., Dean, J.: Efficient estimation of word representations in vector space. arXiv preprint arXiv, 1301.3781 (2013).
15. Liu, T., Zhang, X., Zhou, W., Jia, W.: Neural relation extraction via inner-sentence noise reduction and transfer learning. In: Proceedings of EMNLP, pp. 2195-2204 (2018).
16. Feng, J., Huang, M., Zhao, L., Yang, Y., Zhu X.: Reinforcement learning for relation classification from noisy data. In: AAAI, pp. 5779-5786 (2018).
17. Qin, P., Xu, W., Wang, W. Y.: Robust distant supervision relation extraction via deep reinforcement learning. In: Proceedings of ACL, pp. 2137-2147 (2018).