

Using Gated Convolutional Selector to Improve Relation Extraction

Qian Yi

Institute of Automation, Chinese Academy of Sciences
University of Chinese Academy of Sciences
Beijing, China
Yiqian2016@ia.ac.cn

Guixuan Zhang

KLDRS

Beijing, China
guixuan.zhang@ia.ac.cn

Shuwu Zhang

BJDCRC

Beijing, China

Shuwu.zhang@ia.ac.cn

Abstract—Distant supervision is an effective way to collect large-scale training data for relation extraction. To better solve the wrong labeling problem accompanied by distant supervision, some methods have been proposed to remove noise sentences directly. However, these methods seldom consider the relation label when removing noise sentences, neglecting the fact that a sentence is regarded as noise because the relation it expresses is inconsistent with the relation label. In this paper, we propose a novel method to improve the performance of bag-level relation extractor via removing noise data with a sentence selector. Specifically, the gated convolutional unit of the sentence selector can selectively output features related to the given relation, and these features will be used to judge whether a sentence expresses the given relation. The sentence selector is trained with the data automatically labeled by the relation extractor, and the relation extractor improves its performance with the high-quality data selected by the sentence selector. These two modules are trained alternately, and both of them have achieved better performance. Experimental results show that our model significantly improves the performance of the relation extractor and outperforms competitive baseline methods.

Keywords-relation extraction; convolutional network

I. INTRODUCTION

Relation extraction aims to obtain the relationship between two entities from unstructured text. For example, given a sentence 'Donald Trump was born in America.' and two entities 'Donald Trump' and 'America', relation extraction intends to get the relation'place of birth' from them. Earlier works use manually labeled data to train the classifier in a supervised manner and have achieved good performance[1-5].However, the performances of these models are limited by the scale of the training data, and constructing a large-scale manually labeled dataset is labor consuming. In order to build large-scale dataset automatically, Mintz et al.[6] proposed distant supervision. Distant supervision is based on the idea that if an entity pair $(h; t)$ is contained by a triple $(h; t; r)$ of a given knowledge base, all sentences that contain the entity pair $(h; t)$ will be labeled as the relation r . The $h; t; r$ represent head entity, tail entity and relation, respectively. However, due to the existence of the multi-relational entity pairs, distant supervision suffers from the wrong labeling problem.

Various methods have been proposed to alleviate this issue. One common way among these studies is to employ Multi-Instance Learning(MIL) schema[7,8], in which sentences containing the same entity pair are divided into the same bag and the classification proceeds on bag-level. Zeng et al.[9] selected the most important sentence to represent the bag and trained the model with these selected sentences. Lin et al.[10] applied attention mechanism to give the important sentences lager weights and combined all sentences to obtain the bag representation. Jiang et al.[11] used cross-sentence max-pooling to find the most prominent features among all sentence representations. Recently, some researchers suggested that it was not enough to attenuate the effects of noise data through 'soft' means like attention mechanism. They tended to remove the noise data directly. Feng et al.[12] and Qin et al.[13] trained a sentence selector to distinguish between noise sentences and valid sentences through reinforcement learning(RL). Qin et al.[14] trained a generative adversarial network(GAN) and used the classifier to remove the noise data.

However, these 'hard' methods neglect the fact that when we consider a sentence as noise, it means that this sentence expresses a relation inconsistent with its label. For the three sentences 'Donald Trump is a presedent of America.', 'Donsld Trump was born in America.', 'Donsld Trump is the presedent of America.'and the entity pair 'Donald Trump' and 'America', if the labeling relation is 'place of birth', the first and the third sentence are noise sentences. But when the labeling relation is 'profession', the second sentence becomes the noise data. Therefore, we think it is crucial to consider the labeling relation when identifying noise data.

In this paper, we propose a novel method to improve the performance of bag-level relation extractor via removing noise data with a sentence selector. We design a gated convolutional network for the sentence selector. The gated convolutional network has two convolutional components. One acts as a feature extractor to extract the semantic features of the sentence. The other is a gate, which can select the features related to the given relation. Like the previous models, we encounter the problem of lacking training data for the sentence selector. To deal with this problem, we adopt an easy and reasonable method. We treat each sentence as a bag with only one sentence and use the

pre-trained bag-level relation classifier to classify it. A sentence will be labeled as a positive sample if the classification result is identical to its label. Otherwise, the sentence is labeled as a negative sample. This labeling method is consistent with the idea that the label of valid data is the same as the relation it conveys. As for bag-level relation extractor, we adopt the widely used architectures: piecewise convolutional neural networks(PCNN) [9] with attention mechanism. Moreover, because our model is a generic framework, the relation extractor here can be replaced by any other bag-level relation extractor with different structures. Then we train the bag-level relation extractor and the sentence selector alternately so that their performance can be improved jointly.

The main contributions of this paper can be summarized as follows:

- We design a novel relation-based gated convolutional sentence selector to select valid sentences for distantly supervised relation extraction.
- Experimental results show that our model significantly improves the performance of the relation extractor and outperforms competitive baseline methods.

II. RELATED WORK

The purpose of relation extraction is to obtain the relationship between two entities from unstructured text. Traditional methods leveraged syntactic information and adopted kernel-based classifier to build multi-class relation classifier[1,3]. Recently, more attention has been paid to neural networks methods. In order to extract relation features, previous neural networks models employed various structures to encode the sentence. Zeng et al.[4] adopted CNN to extract the semantic information of the sentence. Xu et al.[15] encoded sentence with Long Short-Term Memory(LSTM) along the shortest dependency path. Zhou et al.[16] combined attention mechanism and LSTM to encode the sentence. Zeng et al.[9] proposed PCNN to extract features from different parts of the sentence separately. Zhang et al.[17] adopted graph convolution over pruned dependency trees to improve the performance of relation extraction. Zhang et al.[18] used attention-based capsule networks to encode the sentence.

In order to solve the problem of lacking for manual annotation data, distant supervision was proposed [6]. To deal with the accompanying wrong labeling problem, researchers have proposed various methods. Zeng et al.[9] adopted the MIL framework. They collected all the sentences containing the same entity-pair as a bag and selected the most important sentence in each bag to train the network. Lin et al.[10] used attention mechanism to give each sentence an importance weight and combined all the sentences to represent the bag. Jiang et al.[11] used cross-sentence max-pooling to extract the features of a sentence bag. Liu et al.[19] softly revised incorrect bag labels with the posterior probability constraint. The above works focused on

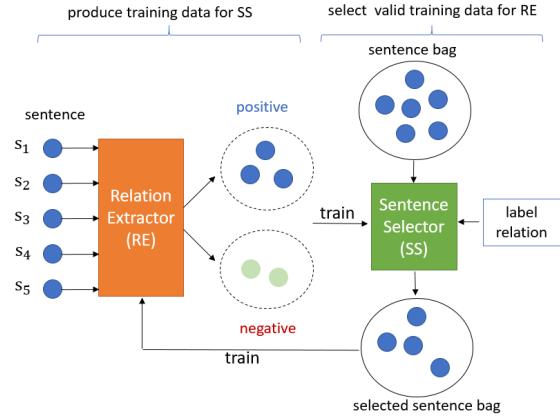


Figure 1. Overall framework.

highlighting the valid sentences of the sentence bag and reduce the effects of noise.

However, some researchers suggested that it was not enough to only weaken the effects of noise data by giving them a small weight, they tended to remove the noise data directly. Feng et al.[12] and Qin et al.[13] trained a sentence classifier to distinguish between noise sentences and other sentences through RL. Qin et al.[14] trained a generative adversarial network and used the classifier to remove the noise data.

III. MODEL

An overview of our framework is shown in Figure 1. The model consists of two parts: a relation extractor and a sentence selector. The sentence selector is trained with data automatically labeled by the relation extractor. And the relation-based selector selects data from each sentence bag according to the labeling relation. The relation extractor is then further trained with the selected high-quality sentence bags. These two modules help each other to obtain better training data and finally achieve better performance.

In this section, we will first describe these two parts in detail, and then introduce the specific details of training and test.

A. Input Layer

Given a sentence s , the input layer transforms the sentence into an embedding matrix, which contains both semantic information and positional information of each word, and feed it to the subsequent networks.

Embedding Word embeddings are low dimensional, continuous and real-valued vectors, which can capture semantic meanings of words. Each word in the vocabulary corresponds to a word embedding vector $c v_w \subseteq \mathbb{R}^{d_w}$. In this paper, we use word embeddings pre-trained on the New York Times(NYT) corpus with Skip-Gram[20].

Position Embedding Position embeddings are vectors that embed the relative distances of each token to the two target entities. For example, in the sentence "SteveJobs was the co-founder and CEO of Apple and...", the relative position from token co-founder to entity SteveJobs and Apple is 3 and -4, respectively. Each relative position value corresponds to a position embedding vector $v_p \subseteq \mathbb{R}^{d_p}$.

For each word w , we concatenate its word embedding vw and two position embeddings (each corresponds to the relative distance from one entity) $vpen1$ and $vpen2$ as its representation $v \subseteq \Re^{d_w+2*d_p}$. Then for each sentence with n words $s = \{w_1, \dots, w_n\}$, we obtain an embedding matrix $S = \{v_1; \dots; v_n\}$ by concatenating all words representations.

B. Relation Extractor

Given an entity pair $(h; t)$ and its sentence bag $S_{h,t} = \{s_1, s_2, \dots\}$, the relation extractor intends to obtain the relation of the bag. Because our model is a generic framework, the relation extractor module can use any bag-level relation extractor. In this paper, we adopt the widely used model: PCNN with attention mechanism.

Sentence Encoder We use PCNN as sentence encoder to encode the sentence embedding matrix into a representation vector.

Convolution A filter $W_r \in \Re^{k_h \times d_v \times m}$ is applied to extract local features of a sentence, where $d_v = d_w + 2 * d_p$ is the dimension of the word vector, m is the width of the filter and k_h is the dimension of the output channel. By sliding W_r along the sentence embedding matrix S_i , we can get the k_h -dimensional feature vector:

$$h_i = \text{ReLU}([v_{i-(m-1)/2}, \dots, v_{i+(m+1)/2}] \otimes W_r + b_r) \quad (1)$$

Then all the feature vectors are concatenated to form a feature map $H = \{h_1; \dots; h_n\}$.

Piece-wise Max-pooling Piece-wise Max-pooling operation is then applied over the feature map H to get the final sentence representation:

$$\begin{aligned} q_j^{(1)} &= \max_{1 \leq i \leq i_{en1}} \{h_{i,j}\} \\ q_j^{(2)} &= \max_{i_{en1} < i \leq i_{en2}} \{h_{i,j}\} \\ q_j^{(3)} &= \max_{i_{en2} < i \leq n} \{h_{i,j}\} \end{aligned} \quad (2)$$

where the subscript j represents the j -th value of a vector, i_{en1} and i_{en2} are the positions of two entities. Then we concatenate the three pooling vectors to get the final sentence representation:

$$q = \{q^{(1)}, q^{(2)}, q^{(3)}\} \quad (3)$$

Sentence Selective Attention After obtaining sentence representation, we apply selective attention to compute the attention score i for each sentence. Then the bag embedding u is computed as a weighted sum of sentence representations:

$$u = \sum_i^{|S_{h,t}|} \alpha_i q_i \quad (4)$$

where the weight i indicates the degree of correlation between sentence and the relation, and $|S_{h,t}|$ is the number

of sentences in the bag. We assign a query vector q_r for each relation r . The attention score is computed as:

$$\begin{aligned} e_i &= q_r^T W_a q_i \\ \alpha_i &= \frac{\exp(e_i)}{\sum_j^N \exp(e_j)} \end{aligned} \quad (5)$$

where W_a is the weight matrix and N is the number of relations.

Loss Function Finally, we obtain the conditional probability $p(r | S_{h,t}, \theta)$ through feeding the bag representation u to a fully connected layer:

$$\begin{aligned} p(r | S_{h,t}, \theta) &= \frac{\exp(o_r)}{\sum_k \exp(o_k)} \\ o &= W_c u + b_c \end{aligned} \quad (6)$$

where W_c is a weight matrix and b_c is a bias vector.

Given the collection of sentence bag $\Omega = \{S_{h_1, t_1}, S_{h_2, t_2}, \dots\}$ and corresponding labeling relation $\{r_1, r_2, \dots\}$, the loss function is defined as follows:

$$J_R = -\frac{1}{|\Omega|} \sum_{i=1}^{|\Omega|} \log p(r_i | S_{h_i, t_i}, \theta) \quad (7)$$

where $|\Omega|$ is the number of bags.

C. Sentence Selector

The sentence selector is a binary classifier which can judge whether a sentence expresses the given relation.

Obtaining Training Data Since there is no training data for the sentence selector, we propose a method to label data automatically. We transform the relation extractor introduced in Section B into a sentence-level relation extractor by regarding each sentence as a bag with only one sentence. Those sentences whose classification result is consistent with their labeling relation are labeled as positive. Otherwise, they are labeled as negative.

After that, we get a data set $D = \{s, e, y, r\}$, in which s represents the sentence, e is the entity pair, y is the two-category label and r is the labeling relation obtained by distant supervision.

Gated Convolutional Network Given a sentence s_i and its corresponding relation r_i , the input to the gated convolutional network is the same as the input embedding matrix in Section A. Specifically, each token is embedded into a word embedding and two position embeddings, so that we get the input embedding matrix $S = \{v_1; \dots; v_n\}$.

Then we feed the input embedding matrix S_i to the gated convolutional unit. The gated convolutional unit

contains two convolutional components. One is a plain convolution operation:

$$h_i^s = \tanh([v_{i-(m-1)/2}, \dots, v_{i-(m+1)/2}] \otimes W_s + b_s) \quad (8)$$

The other convolution operation is a convolutional gate:

$$h_i = \sigma([v_{i-(m-1)/2}, \dots, v_{i-(m+1)/2}] \otimes W_g + b_g) \quad (9)$$

Then we compute an element-wise multiplication between the feature vector h^s and the relation gate vector h :

$$h_i^g = h_i^s \circ h \quad (10)$$

where the symbol \circ represents the element-wise multiplication. The max pooling procedure is then performed over the feature maps to obtain the sentence embedding:

$$q_j^g = \max_{1 \leq i \leq n} \{h_{i,j}^g\} \quad (11)$$

Loss Function We feed the sentence embedding q_i^g to a fully connected layer to compute the posterior probability $p(y'|s, r, \phi)$:

$$p(y'|s, r, \phi) = \frac{\exp(o_{y'}^g)}{\sum_k \exp(o_k^g)} \quad (12)$$

$$o^g = W_o q^g = b_o$$

where y' is the two-category label.

Given the collection of sentences $\Lambda = \{s_1, s_2, \dots\}$, its label relation $\{r_1, r_2, \dots\}$ and the corresponding label $\{y_1, y_2, \dots\}$, the loss function is defined as follows:

$$J_s = -\frac{1}{|\Lambda|} \sum_{i=1}^{|\Lambda|} \log p(y_i | s_i, r_i, \phi) \quad (13)$$

where $|\Lambda|$ is the number of bags.

D. Training and Test

Because the performance of the relation extractor and the sentence selector influence each other. We train the two modules alternately.

During training, we first adopt Adam algorithm to minimize the loss function Eq. 7 with the original dataset. After using relation extractor to generate training data for the sentence selector, we then optimize the sentence selector by minimizing the Eq. 13 with Adam. Next, we utilize the sentence selector to select the sentences to further train the relation extractor. The relation extractor and the instance selector are trained alternately as described above until convergence.

During testing, we first select the test data with the sentence selector. Then we feed selected sentence bags into the relation extractor.

What needs to be mentioned is that, in order to increase the recall, we set a threshold $u < 0.5$ when selecting data, and only sentences whose selecting probability is smaller than u will be removed. This operation increases the tolerance for classification errors of sentence selector and enhances the recall of the relation extractor.

IV. EXPERIMENTS

A. Dataset and Evaluation

In this paper, we evaluate our model on the widely used New York Times(NYT) dataset developed by [7]. This dataset is constructed by aligning Freebase with New York Times(NYT) corpus through distant supervision. There are 522611 sentences in the training set and 172448 sentences in the test set, and these sentences are labeled by 53 candidate relations. Among the 53 relations, there is a label NA, which represents there is no relation between the two entities in a sentence. During training, We randomly extract ten percent of the sentences from the training data as the validation data and the rest as the training data.

We evaluate all methods with the held-out evaluation. The held-out evaluation compares the relational facts extracted from the test set by models with all the facts existing in the test sentences(which is labeled by Freebase through distant supervision). For evaluation, we present precision-recall curves for all models.

B. Implementation Detail

In our experiment, our parameter settings are as follows: the dimension of word embedding d_w and position embedding d_p are 50 and 5; the width of the convolution kernel m is 3 and the dimension of the output channel k_h of the convolution filter is 230; the max sentence length is 120; the batch size is fixed to 50 and dropout probability is fixed to 0.5. We adopt Adam to update the parameters, and the learning rate for training relation extractor and sentence selector are set to 0.001 and 0.0005. As for the threshold u for the sentence selector during selecting, we tune it on the validation dataset and pick $u = 0.3$ in the candidate set $\{0.1, 0.2, 0.3, 0.4, 0.5\}$.

C. Comparison with Baseline Models

To evaluate the performance of our proposed model, we compare our model with various baseline models. PCNN+MIL [9] proposed piecewise CNN to encode the sentence and adopted the MIL framework. CNN + ATT and PCNN+ATT[10] employed attention mechanism to reduce the influence of noise data and used CNN and PCNN as sentence encoder respectively. APCNNS+D [21] used

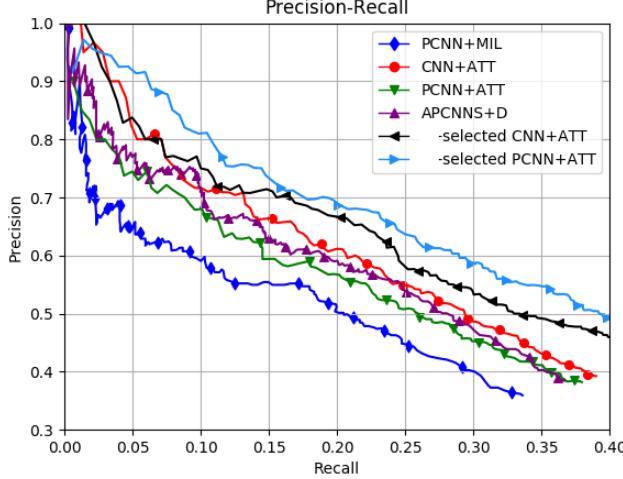


Figure 2. Comparison with baseline models.

external entity descriptions and attention mechanism to obtain better bag representation.

As shown in Figure 2, our models, which are denoted as selected+PCNN+ATT and selected+CNN+ATT, have a significant improvement on all baseline models. For a more detailed comparison, we show the precision@N(P@N) of our models(reselected+ PCNN+ATT and re-selected+CNN+ATT) and the corresponding baseline models(CNN+ATT and PCNN+ATT) in Table 1. The results demonstrate the effectiveness of the sentence selector for distant supervision relation extraction. The re-selected CNN+ATT model and the re-selected PCNN+ATT both achieve higher values for P@100, P@200, P@500 compared to the original baseline models. Moreover, the mean value of selected+CNN+ATT is 6% higher than CNN+ATT, and selected+PCNN+ATT is 5.3% higher than PCNN+ATT.

D. Comparison with Selector-Based Models

We also compare the performance of our model with other selector-based models to further assess the sentence selector. PCNN+ATT+DSGAN [14] trained a generative adversarial network and used the classifier to remove the noise data. PCNN+ATT+RL [22] trained a sentence selector through reinforcement learning.

As Figure 2(b) shows, when compared to other selector-based models, PCNN+ATT with relation-based gated selector also achieves better performance on both precision and recall. Moreover, compared to the RL and GAN, our model is more stable and easier to converge when training the sentence selector.

V. CONCLUSION

In this paper, we propose a novel method to improve the performance of bag-level relation extractor via removing noise data with a sentence selector for neural relation

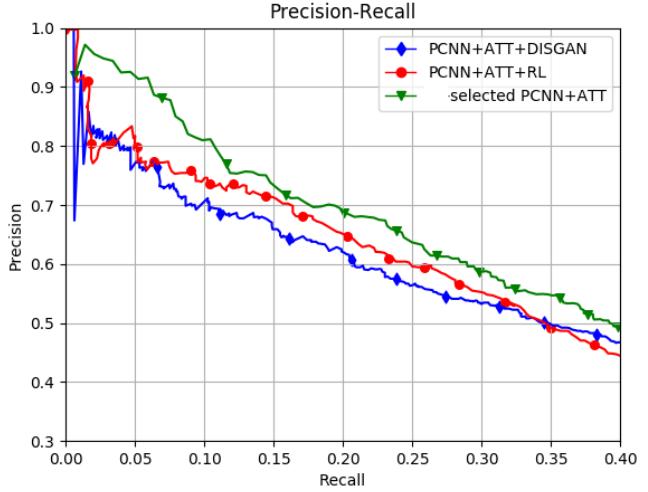


Figure 3. Comparison with selector-based models.

extraction under the distant supervision scenario. The whole model contains a relation extractor and a sentence selector composed of a gated convolutional network. We train the sentence selector without manually labeled data and employ the selector to select high-quality data for training the relation extractor. We conduct experiments on a widely used dataset. The experimental results confirm the effectiveness of the gated convolutional unit and our framework significantly improves the performance of the original bag-level relation extractor.

ACKNOWLEDGMENT

This work was supported by the National Key R&D Program of China (2018YFB1403900) and the Science and Technology Program of Beijing (Z201100001820002). It was also the research achievement of the Key Laboratory of Digital Rights Services, which is one of the National Science and Standardization Key Labs for Press and Publication Industry.

REFERENCES

- [1] D. Zelenko, C. Aone, and A. Richardella, “Kernel methods for relation extraction,” *Journal of machine learning research*, vol. 3, Feb. 2003, pp. 1083-1106, doi: 10.3115/1118693.1118703.
- [2] G. Zhou, J. Sun, J. Zhang, and M. Zhang, “Exploring various knowledge in relation extraction,” *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics*, Jun. 2005, pp. 427-434, . doi:10.3115/1219840.1219893.
- [3] R. Mooney, and R. Bunescu, “Subsequence kernels for relation extraction,” *Advances in neural information processing systems*, Dec. 2006, pp. 171-178.
- [4] D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao, “Relation classification via convolutional deep neural network,” *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics*, Aug. 2014, pp. 2335-2344.
- [5] C. Santos, B. Xiang, and B. Zhou, “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, Association for Computational Linguistics*, May. 2015, pp. 626-634, doi: . 10.3115/v1/P15-1061.

- [6] M. Mintz, S. Bills, R. Snow, and D. Jurafsky, “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, ACL and AFNLP, Aug. 2009, pp. 1003-1011, doi: doi:10.3115/1690219.1690287.
- [7] S. Riedel, L. Yao, and A. McCallum, “Modeling relations and their mentions without labeled text,” Joint European Conference on Machine Learning and Knowledge Discovery in Databases, Springer-Verlag Berlin Heidelberg, Sep. 2010, pp. 148-163, doi: doi:10.1007/978-3-642-15939-8_10.
- [8] R. Hoffmann, C. Zhang, X. Ling, L. Zettlemoyer, and D. Weld, “Knowledge-based weak supervision for information extraction of overlapping relations,” Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Jun. 2011, pp. 541-550.
- [9] D. Zeng, K. Liu, Y. Chen, and J. Zhao, “Distant supervision for relation extraction via piecewise convolutional neural networks,” Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Sep. 2015, pp. 1753-1762, doi: 10.18653/v1/D15-1203.
- [10] Y. Lin, S. Shen, Z. Liu, H. Luan, and M. Sun, “Neural relation extraction with selective attention over instances,” Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Dec. 2016, pp. 2124-2133, doi: 10.18653/v1/P16-1200.
- [11] X. Jiang, Q. Wang, P. Li, and B. Wang, “Relation extraction with multi-instance multi-label convolutional neural networks,” Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics, Dec. 2016, pp. 1471-1480.
- [12] J. Feng, M. Huang, L. Zhao, Y. Yang, and X. Zhu, “Reinforcement learning for relation classification from noisy data,” Thirty-Second AAAI Conference on Artificial Intelligence, Apr. 2018, pp. 5779-5786.
- [13] P. Qin, W. Xu, Wang, and W. Wang, “Robust Distant Supervision Relation Extraction via Deep Reinforcement Learning,” Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Jul. 2018, pp. 2137-2147, doi: 10.18653/v1/P18-1199.
- [14] P. Qin, W. Xu, Wang, and W. Wang, “Dsgan: Generative Adversarial Training for Distant Supervision Relation Extraction,” Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Jul. 2018, pp. 496-505, doi: 10.18653/v1/P18-1046.
- [15] Y. Xu, L. Mou, G. Li, Y. Chen, H. Peng, and Z. Jin, “Classifying relations via long short term memory networks along shortest dependency paths,” Proceedings of the 2015 conference on empirical methods in natural language processing, Association for Computational Linguistics, Sep. 2015, pp. 1785-1794, doi: 10.18653/v1/D15-1206.
- [16] P. Zhou, W. Shi, J. Tian, Z. Qi, B. Li, H. Hao, and B. Xu, “Attention-based bidirectional long short-term memory networks for relation classification,” Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Aug. 2016, pp. 207-212, doi: 10.18653/v1/P16-2034.
- [17] Y. Zhang, P. Qi, and C. Manning, “Graph convolution over pruned dependency trees improves relation extraction,” Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Sep. 2018, pp. 2205-2215, doi: 10.18653/v1/D18-1244.
- [18] N. Zhang, S. Deng, Z. Sun, X. Chen, W. Zhang, and H. Chen, “Attention-based capsule networks with dynamic routing for relation extraction,” Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Oct. 2018, pp. 985-992, doi: 10.18653/v1/D18-1120.
- [19] T. Liu, K. Wang, B. Chang, and Z. Sui, “A soft-label method for noise-tolerant distantly supervised relation extraction,” Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Sep. 2017, pp. 1790-1795, doi: 10.18653/v1/D17-1189.
- [20] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean, “Distributed representations of words and phrases and their compositionality,” Advances in Neural Information Processing Systems, Dec. 2013, pp. 3111-3119.
- [21] G. Ji, K. Liu, S. He, and J. Zhao, “Distant supervision for relation extraction with sentence-level attention and entity descriptions,” Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, Association for the Advancement of Artificial Intelligence, Jun. 2017, pp. 1-7.
- [22] Y. Zhang, V. Zhong, D. Chen, G. Angeli, and C. Manning, “Position-aware attention and supervised data improve slot filling,” Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Sep. 2017, pp. 35-45, doi: 10.18653/v1/D17-1004.