Entity Relation Extraction Based on CNN and Attention Mechanism

Anonymous ACL submission

Abstract

Distant supervision combined with neural network model has been widely used in entity relation extraction. However, there are often a lot of noisy data in the labeled dataset obtained by distant supervision, which seriously hurts the performance of the extraction model. In this paper, we propose a model based on improved sentence-level attention (IATT) incorporating with piecewise convolutional neural networks (PCNN). Our model can reduce the impact of noisy instances maximally and make full use of the semantic information of the positive instances by combining sentence feature vectors, which contains positive instances as much as possible, and abandons possible noisy sentences. Experiments show that our model achieves higher precision on relation extraction than the baseline methods without compromising recall rate.

1 Introduction

Entity relation extraction is defined as the task of extracting binary relations between two entities from plain texts. Supervised methods are widely used for this task due to their relatively high performance, but they often suffer from lacking of sufficient and accurate labeled dataset. To alleviate this issue, distant supervision(Mintz et al., 2009) was proposed to generate labeled data automatically by aligning relation facts(two entities with some relation e.g., (Beijing, capital, China)) in a knowledge base (KB) with sentences mentioning these relations. However, there are still some deficiencies in distant supervision.

First, distant supervision inevitably accompanies with the wrong labeled data, for it assumes that if two entities have a relation in a known knowledge base, then all sentences that mention these two entities will express this relation in some way. In fact, a sentence that mentions two entities does not necessarily express the relation in a knowledge base. For example, (Apple, founder, Steve Jobs) is a relation fact in KB, "Steve Jobs had experienced decades of ups and downs of Apple" is a sentence of plain texts. Obviously, this sentence does not express the relation "founder", but it will be labeled as "founder", which leads to wrong labeled data. Although Riedel et al.(2010) relaxed this assumption afterwards, there are still a lot of noisy data in the labeled dataset.

Second, classical distant supervision methods (Mintz et al., 2009; Riedel et al., 2010; Hoffmann et al., 2011) have applied supervised models to capture lexical and syntactic features of the labeled data obtained through distant supervision. These features are often derived from existing Natural Language Processing (NLP) tools, which inevitably lead to error propagation or accumulation since the errors exist in NLP tools. To alleviate this issue, some recent studies have utilized deep neural networks to extract sentence features automatically. Zeng et al.(2015) proposed a model incorporating multi-instance learning with PCNN, which can build relation extractor based on distant supervision data. Although the method achieves significant improvement in relation extraction, it is still far from reasonable result. This method assumes that at least one sentence that mentions the same entity pair will express their relation in KB, and only selects the most likely one sentence for each entity pair in training and prediction. It's apparent that the method will lose a large amount of rich information contained in neglected sentences. Therefore, Lin et al.(2016) proposed a sentence-

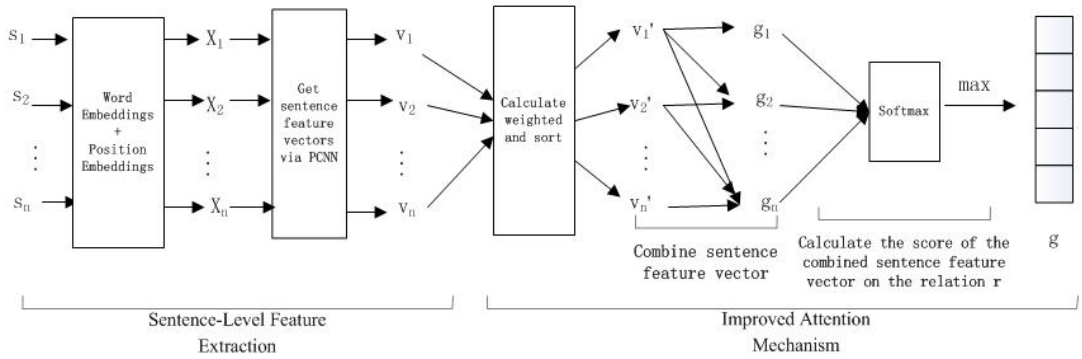

Figure 1:Convolution neural network model based on improved attention mechanism

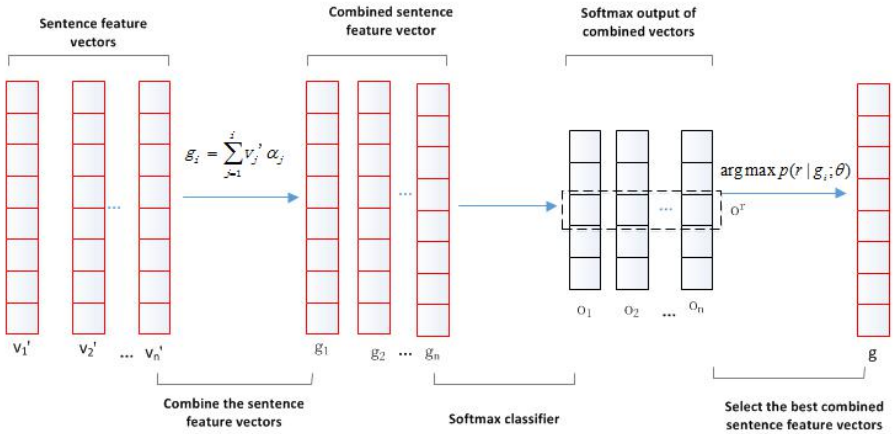

Figure 2: Improved attention mechanism

level attention-based CNN model for distant supervised relation extraction. The sentence-level attention can dynamically reduce the weights of those noisy instances, then extract relation with the feature vector weighted by sentence-level attention. Although the method achieves significant improvement, the noisy instances with low weight still affect the performance of the model, and the impact will be more serious as the noisy instances increase.

In this paper, we propose a model based on improved sentence-level attention(IATT) incorporating with PCNN. Our model can reduce the impact of noisy instances maximally and make full use of the semantic information of the positive instances. Experiments show that the precision of our method has increased by 5% to 11% without compromising the recall as compared with the baseline methods.

2 Our approach

Suppose a set S contains n sentences for the same entity pair (e_1, e_2), S ={s_1, s_2, ..., s_n}. We introduce our model in two main parts:

Sentence-Level Feature Extraction. Get the feature vector representation of each sentence (i.e., {v_1,v_2,...,v_n}) via word embedding, convolution operation and pooling operation of CNN.

Improved Attention Mechanism. Based on the difference of support degree for relation r of each sentence feature vector, we utilize the improved attention mechanism to construct combined feature vector set {g_1, g_2, ..., g_n}, and calculate the score of g_i ($1 \leqslant i \leqslant n$) on relation r, the combined feature vector g with the highest score is selected, and we believe that g contains the most positive instance information and the least noisy information. At last , we use g to train CNN. As shown in Figure 1.

2.1 Sentence-Level Feature Extraction

We transform a sentence s_i ($s_i \in S$) into its distributed representation by a PCNN. First, words in the sentence are transformed into low-dimensional real-valued feature vectors, the vector representation set {X_1,X_2,...,X_n} of sentences is generated. X_i represents the vector of sentence s_i ($1 \leqslant i \leqslant n$), where semantic information and location information are considered. After that, the

convolutional layer and piecewise max-pooling layer are used to construct feature vector set $\{v_1,v_2,...,v_n\}$, as shown in Figure 1. Details can be referred to Zeng et al.(2015).

2.2 Improved Attention Mechanism

We utilize the improved attention mechanism to construct combined feature vector which contains the most positive information and the least noisy information.

Order Sentence Vectors by Weights. For a sentence set $S=\{s_1,s_2,...,s_n\}$, we can get the feature vector representation $\{v_1,v_2,...,v_n\}$ for each sentence according to section 2.1 and Figure 1. Since there are differences on the degree of expressing the relation r in sentence vectors, we assign weight for each sentence vector according to its support degree for the relation r (r is relation fact in KB, which contains the same entity pairs with the sentence set S). We calculate the weight β_1 as follows:

$$\beta_i = \frac{\exp(e_i)}{\sum_{k=1}^{n}\exp(e_k)} \quad , 1\leq i\leq n \tag{1}$$

where e_i is referred as a query-based function which describes the matching score of the input sentence feature vector v_i and the predicted relation r. Its calculation method is as follows:

$$e_i = v_i A r \quad , \quad 1\leq i\leq n \tag{2}$$

where A is a weighted diagonal matrix, and r is the query vector associated with relation r which indicates the representation of relation r.

Then we descending order the sentence vector representations $\{v_1,v_2,...,v_n\}$ as $\{v_1',v_2',...,v_n'\}$ according to the weights calculated by formula (1).

Combine Sentence Feature Vectors and Select the Best One. As shown in Figure 2, we combine these ordered feature vectors $\{v_1',v_2',...,v_n'\}$ one by one from which has the highest weight, and the combined sentence feature vectors $\{g_1,g_2,...g_n\}$ is generated.

$$g_i = \sum_{j=1}^{i}\alpha_j v_j \quad , \quad 1\leq i\leq n \tag{3}$$

where α_j is the weight of v_j', which is calculated as follows:

$$\alpha_j = \frac{\exp(e_j)}{\sum_{k=1}^{i}\exp(e_k)} \quad , \quad 1\leq j\leq i \tag{4}$$

The parameters in formula (4) have the same meaning as formula (1). Then we calculate the conditional probability for each combined feature vector as follows:

$$p(r\mid g_i;\theta) = \frac{\exp(o_i^r)}{\sum_{k=1}^{n_1}\exp(o_i^k)} \quad , \quad 1\leqslant i\leqslant n \tag{5}$$

where n_1 is the total number of relations and o^r_i is the score associated to relation r, which is defined as follows:

$$o_i^r = W_0(g_i \circ f) + b \quad , \quad 1\leq i\leq n \tag{6}$$

where $b\in R^{n1}$ is a bias vector , $W_0\in R^{n1\times 3n}$ is a weight matrix and f is a vector of Bernoulli random variables with probability p. The operation (g_i o f) is called dropout (Srivastava et al., 2014) which can prevent overfitting.

Finally, we select the best combined sentence feature vector as follows:

$$g = \arg\max p(r\mid g_i;\theta) \ , 1\leqslant j\leqslant n \tag{7}$$

Then we define the objective function using cross-entropy at the set level as follows:

$$J(\theta) = \sum_{j=1}^{N}\log p(r_j\mid S_j^i,\theta) \tag{8}$$

where N is the number of sentence sets and S^i_j represents the best combined sentence feature vector in the j-th sentence set, which is obtained according to formula (7).

3 Experiments

3.1 Dataset , Evaluation Metrics and Experimental Settings

We use the dataset developed by (Riedel et al., 2010), which was generated by aligning Freebase relations with the New York Times (NYT) corpus. Similar to previous work (Mintz et al., 2009), we evaluate our model in the held-out evaluation and report the precision/recall curves in our experiments.

In this paper, we use the word2vec tool to train the word embeddings on the NYT corpus.

All the parameters we used in the experiments are as follows:

Window size	w=3	Feature maps	n=230
Word dimension	d_w=50	Position dimension	d_p=5
Batch size	B=160	Adadelta parameter	ρ=0.95,ε=1 e^{-6}
Dropout probability	p=0.5	Learning rate	λ=0.01

Table 1: Parameters used in our experiments

3.2 Experimental Result

Figure 3 is the comparison of our method with other neural extraction methods. We have the following observation from Figure 3: (1) the PCNN+ONE method (Zeng et al., 2015) brings better performance as compared to PCNN+AVE (Lin et al., 2016, Lin mentioned it but did not recommend it). The reason is that the PCNN+AVE method treats each sentence in the set S equally, which leads to obtain richer semantic information as well as noisy information. Obviously, the damages of noisy information are far more serious than the advantage of rich semantic information. The experiment indicates that the proportion of noisy data in the dataset obtained by distant supervision is very large. (2) PCNN+ATT method (Lin et al., 2016) brings better performance as compared to PCNN+AVE method and PCNN+ONE method, which indicates that the sentence-level attention is useful for reducing the influence of noisy data by setting weights for sentences. (3)The PCNN+IATT method brings better performance as compared to PCNN+ATT method, which indicates that the improved attention mechanism can take full use of the positive instances and abandon the noisy instances efficiently. (4)The PCNN+IATT method achieves the highest precision over the entire range of recall compared to other methods, which indicates that the proposed improved attention mechanism is beneficial. This mechanism can take full use of the semantic information and effectively filter out noisy sentences, which alleviates the wrong labeled problem greatly in distant supervision relation extraction.

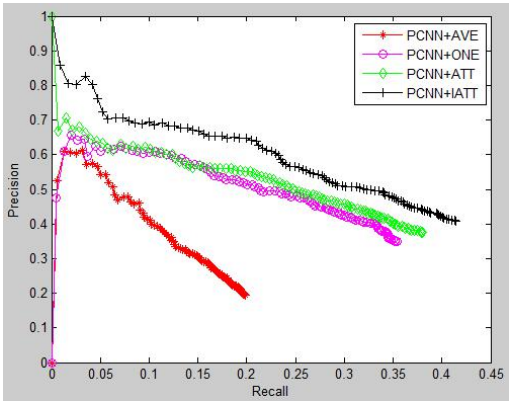

Figure 3: Comparison of comparison of our method with other neural extraction methods

Fig.4 is the comparison of our method with traditional methods. From Fig.4 we can observe that: PCNN+IATT model significantly outperforms the main traditional methods (Mintz et al., 2009; Surdeanu et al., 2012) over the entire range of recall. It demonstrates that the human-designed feature cannot precisely express the semantic meaning of the sentences, and the inevitable error brought by NLP tools will degrade the performance of relation extraction. In contrast, PCNN+IATT which learns the representation of each sentences automatically can express each sentence well and can filter out noisy sentences effectively.

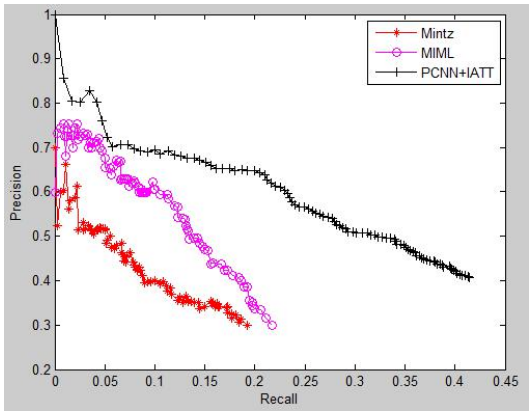

Figure 4:Comparison of our method with traditional methods

4 Conclusions

In this paper, we propose a model based on improved sentence-level attention (IATT) incorporating with PCNN. Our model can reduce the impact of noisy instances maximally and make full use of the semantic information of all the positive instances. The experimental results show that our model achieves significant and consistent improvements in relation extraction as compared with the state-of-the-art methods.

References

Mike Mintz, Steven Bills, Rion Snow,and Dan Jurafsky.2009. *Distant supervision for relation extraction without labeled data.* In Proceedings of ACL-IJCNLP,pages 1003-1011.

Sebastian Riedel, Limin Yao, and Andrew McCallum.2010. *Modeling relations and their mentions without labeled text.* In Proceedings of ECML-PKDD,pages 148-163.

Raphael Hoffmann, Congle Zhang, Xiao Ling, Luke Zettlemoyer, and Daniel S Weld. 2011. *Knowledge based weak supervision for information extraction of overlapping relations.* In Proceedings of ACLHLT, pages 541-550.

Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati,and Christopher D Manning. 2012. *Multi-instance multi-label learning for relation extraction.* In Proceedings of EMNLP, pages 455-465.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky,Ilya Sutskever, and Ruslan Salakhutdinov. 2014. *Dropout: A simple way to prevent neural networks from overfitting.* JMLR, 15(1):1929-1958.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou,and Jun Zhao. 2014. *Relation classification via convolutional deep neural network.* In Proceedings of COLING, pages 2335‑2344.

Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao.2015. *Distant supervision for relation extraction via piecewise convolutional neural networks.* In Proceedings of EMNLP.

Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan,and Maosong Sun.2016. *Neural Relation Extraction with Selective Attention over Instances.*In proceeding of ACL.