Combined Networks with Multi-level Attention for Distantly-Supervised Relation Extraction

Hongyang Yuan
School of Computer Science, Chongqing University, Chongqing, 400000, China
yuanhongyang@cqu.edu.cn

Abstract. Distantly-supervised relation extraction is expected to extract relational facts from very large corpora. However, it is inevitably accompanied by the problem of mislabeling, which affects the performance of relation extraction. To obtain richer information from sentences and make full use of the information of entity pairs, we propose the combined networks containing piecewise convolutional neural networks (PCNN) and bidirectional gated recurrent units (BiGRU) with multi-level attention. In this model, PCNN and BiGRU are utilized to obtain local features and sequence information. Then the multi-level attention is proposed to extract the most correlated information, which can affect the assignment of sentence attention weight. And it reduces the influence of wrong labeled instances indirectly. The experimental results show that our method can enhance the representation power of the network and prove the effectiveness of our model compared with several baseline methods.

1. Introduction
In this paper, we apply the combined networks with multi-level attention to bring better performance in the relation classification task. First, PCNN [1] introduces the piecewise max-pooling to expand effects on the location feature of the word and the entity pair. And BiGRU is applied to make up for the shortcomings that PCNN can’t focus on the hidden state by every time step. Meanwhile, we get multi-level attention to focus on word and sentence which is related to the relation label. The contribution of this paper can be summarized as follows:

- The semantic and syntactic information are fully learned via the combined networks.
- Multi-level attention can learn the word-relation relevance deeply and get a better representation of sentence which affects the performance of relation classification.
- Through extensive experiments on datasets, we demonstrate that our model is effective.

2. Related Work
Distant supervision [2] is proposed to solve the time-consuming problem caused by manual tagging. However, it introduces the noise issues, because the sentence mentioning entity pair may not express the relation in the knowledge base. Many works have been done to improve the performance of the relation classification task. Multi instance learning (MIL) [3] is adopted in relation extraction [4-5]. PCNN [1] with MIL is proposed to capture structural information between two entities and get informative features. To make full use of information, Lin et al. [6] introduce sentence-level attention to reduce noise and strengthen positive instances. It can effectively filter a substantial number of noisy instances.
3. Methodology
In this section, we introduce our model in detail. And figure 1 shows the framework of our model.

![Combined Networks Diagram](image)

Figure 1. The architecture of combined networks with multi-level attention.

3.1. Look Up
Similar to Zeng et al. [7], each word is represented by the combination of the word embeddings and position embeddings. We get the matrix which is the representation of the sentence. We consider \( S \) to be a sequence \( \{q_1, q_2, \ldots, q_s\} \), where \( q_i \in \mathbb{R}^{d_w} \), \( d = d_w + 2d_p \). \( d_w \) and \( d_p \) are the dimension of the word embeddings and position embeddings, respectively.

3.2. Combined Networks
We adopt PCNN [1] to learn sentence-level features that consider the structure information of entity positions. And we utilize a BiGRU layer to obtain the detail from the past and the future. The hidden size of BiGRU and window size of PCNN are both \( n \). Then the sentence representation \( f \in \mathbb{R}^{5n} \) after the operation of concatenation can be expressed by the following formula:

\[
g = PCNN(S) \tag{1}
\]
\[
h = BiGRU(S) \tag{2}
\]
\[
f = [g; h] \tag{3}
\]

3.3. Multi-level Attention
Word attention is introduced to measure the contribution of each word in the sentence. Motivated by TransE [8], the relation \( r \) between the head entity \( e_1 \) and the tail entity \( e_2 \) is expressed as \( e_1 - e_2 \).

The word attention weight \( \alpha_i \) is computed as follows:

\[
\alpha_i = \text{soft max}(W_w^T (\tanh[q_i; e_1 - e_2]) + b_w) \tag{4}
\]
\[
y = \sum_{i=1}^{l} \alpha_i q_i \tag{5}
\]
\[
z = My + b \tag{6}
\]
\[
f = [f \oplus z] \tag{7}
\]
where \( ; \) denotes the vertical concatenation of \( q_i \) and \( e_1 - e_2 \), \( W_u \in \mathbb{R}^{d+d_u} \) is an intermediate matrix and \( b_u \) is an offset value. The sentence can be represented by the vector \( y \in \mathbb{R}^d \). After a non-linear layer, we add \( z \) to the output of the combined networks as the final representation of the sentence.

Here we use sentence attention [6] to represent each sentence \( x \) in the bag \( X \). We add a neural network layer, where \( W_x \in \mathbb{R}^{3n \times n} \), then define the conditional probability through a softmax classifier to predict the label of relation. Finally, we compute the cross-entropy function

\[
J(\theta) = \sum_{k=1}^{N} \log p(r_k | X_k, \theta)
\]

and \( N \) is the number of the bags in the training set.

\[
o = W_x x + b_x
\]

\[
p(r | X, \theta) = \frac{\exp(o_r)}{\sum_{j=1}^{n} \exp(o_j)}
\]

4. Experiments

We evaluate our model on the widely used dataset NYT [3]. Similar to Lin et al. [6], we report the aggregate curves precision/recall curves and Precision@N (P@N) in our experiments. To demonstrate the effects of our model, we compare our methods with baseline models[2, 4-6]. We denote CN+ATT as the combined works with sentence attention, PCNN+2ATT as PCNN with multi-level attention, and CN+2ATT as the whole framework proposed in this paper.

From table 1 and figure 2 we can observe that:

- Both CN+ATT and PCNN+2ATT performs better than PCNN+ATT. It indicates the combined networks can improve the representational capacity of a network. Meanwhile, multi-level attention avoids wrong labeling sentences by giving low attention.
- For all baselines, our model performs best in all settings.

| P@N(%) | 100 | 200 | 300 | Average |
|--------|-----|-----|-----|---------|
| PCNN+ATT | 76.2 | 73.1 | 67.4 | 72.2     |
5. Conclusion

In this paper, we use combined networks to fully extract the semantic information of the sentence and thus obtain the local features of the sentence and the global features of the sequence. Meanwhile, we use multi-level attention to enhance sentence representations to influence the results of relation classification and improve the ability of the overall model to recognize noise. Compared with baselines, our method can further improve the performance of relation extraction.

References

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

[2] Mintz, M., Bills, S., Snow, R., & Jurafsky, D. (2009, August). 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: Volume 2-Volume 2 (pp. 1003-1011).

[3] Riedel, S., Yao, L., & McCallum, A. (2010, September). Modeling relations and their mentions without labeled text. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases (pp. 148-163). Springer, Berlin, Heidelberg.

[4] Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L., & Weld, D. S. (2011, June). Knowledge-based weak supervision for information extraction of overlapping relations. In Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1 (pp. 541-550).

[5] Surdeanu, M., Tibshirani, J., Nallapati, R., & Manning, C. D. (2012, July). 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).

[6] Lin, Y., Shen, S., Liu, Z., Luan, H., & Sun, M. (2016, August). 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).

[7] Zeng, D., Liu, K., Lai, S., Zhou, G., & Zhao, J. (2014). Relation classification via convolutional deep neural network.

[8] Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., & Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In Advances in neural information processing systems (pp. 2787-2795).