Improving Distant Supervised Relation Extraction using shortest dependency path

Yang Luo¹, Lijun Zhang², Dawei Yin³, Dezhong Peng¹, Ke Yang⁴*
¹ College of Computer Science, Sichuan University, Chengdu, 610025, China
² Chengdu Ruibe Yingte Information Technology Limited Company
³ Sichuan Zhiqian Technology Co., Ltd
⁴ China National Tobacco Corporation Sichuan Province Co.
* Corresponding author’s e-mail: luoyangcsm@stu.scu.edu.cn

Abstract. Distant supervised relation extraction is the problem of classifying the semantic relation between a pair of entities in a given sentence, where data is created by aligning unlabelled text with knowledge base automatically. Prior researches indicated that most of the sentences in the distant supervised relation extraction setting are very long and benefit from the attention over words. We note that the shortest dependency path between the entity pair in the syntax analysis tree of a sentence can help word based attention. In this paper, we propose a novel distant supervised neural relation extraction method, which makes use of the shortest dependency path to supervise the learning of word attention. Through extensive experiments on benchmark datasets, we demonstrate effectiveness of our model.

1. Introduction

Relation extraction (RE) can be defined as a natural language processing task that classifying the relationship between a pair of entities in a sentence, which is a subtask of information extraction. In amount of natural language processing tasks including construction of information extraction, RE has been an important step. A lot of labelled data is essential to train an efficient RE model. However, obtaining large-scale annotation data requires expensive costs. In order to solve this problem, Mintz[1] proposed distant supervised relation extraction method, which aligned lots of unlabelled text with freebase[2] and generated labelled data automatically. Distant supervised relation extraction method is based on a hypothesis that a sentence can be marked as positive instance if it contains a pair of entities in knowledge base. However, there are much noisy data because of the strong hypothesis and incompleteness of the knowledge base. It mainly includes two kinds of errors, false positive and false negative instances. Therefore, distant supervised RE was regarded as a multi instance multi label learning problem by Surdeanu[3].

According to the observation of Zeng[4], 50% of the sentences in Riedel2010 Distant Supervised dataset[5], a popular distant supervised benchmark dataset, had 40 or more words in them. Sharmistha[6] proposed word based attention model (BGWA), which using word attention mechanism to make the RE model focus correct context information in sentence. We observed that shortest dependency path (SDP) of entities in dependency tree usually indicates the relation between entities. In other words, SDP reveals which words are significant for RE in a sentence.
A straightforward way to use SDP is to concatenate words with SDP as sentences inputs. We show in Section 3.2 that this approach doesn’t help the model. In our model, SDP is used to supervise the distribution of words attention in sentences, which shows better results compared to the baseline.

**Freebase**

| Relation            | Entity1          | Entity2          |
|---------------------|------------------|------------------|
| /business/company/founders | Microsoft       | Bill Gates       |
| ...                 | ...              | ...              |

**Mentions from free texts**

1. Bill Gates was the co-founder and CEO of Microsoft.
2. Bill Gates has officially retired from Microsoft, but remains chairman.

Figure 1: Training instances generated through distant supervision. Upper sentence: correct labelling; lower sentence: incorrect labelling.

The contributions of this paper are as follows:

- We propose a neural network model for improving distant supervision RE by combing SDP in the form of word-level supervision. The model is trained in a multi task learning setup, where the SDP is employed as supervision for attention weights.
- We demonstrate effectiveness of our model over the baseline through extensive experiments on benchmark datasets.

2. Methods

2.1. background

2.1.1. Relation Extraction.

A relation is defined as a semantic attribute of entities. In our task, only binary relations are considered. For example, Born_In(Yao Ming, Shanghai). Given a sentence set \( S = \{S_i\} \), each sentence in set contains the same pair of entities, then distant supervised RE can be formulated as \( F \) for a pair entities:

\[
F_r(S, (e_1, e_2)) = \begin{cases} 
1, & \text{if relation } r \text{ is true for pair } (e_1, e_2) \\
0, & \text{otherwise}
\end{cases}
\]

2.1.2. SDP.

Dependency parsing is a way to analyse sentences that a sentence is parsed by choosing for each word what other word is it a dependent of. Usually dependency parsing makes the dependencies a tree, which contains abundant syntactic information. SDP[7] is the shortest path between two entities on the tree. Early researches indicated that SDP contains some important information for RE.

![Figure 2: The SDP between the entity pair is highlighted in dependency parse tree.](image)

2.2. proposed method.

Given a set of sentences \( \{x_1, x_2, \ldots, x_n\} \), two corresponding entities and SDP of entities, our model computes the probability for each class relation. Our model is showed as in Fig.2.
2.2.1. Input Representation.
The inputs of the model are token sequence of raw sentences. Firstly, we transform word into low dimensional real-value vectors. And each word becomes a low-dimension vector by a word embedding matrix. Given a sentence $x$, which is consist of $n$ words $x = \{w_1, w_2, \ldots, w_n\}$, each word $w_i$ is transformed into a real-value vector. Word representations are encoded by vectors in an embedding matrix $V \in \mathbb{R}^{d \times |V|}$ where $V$ is the fixed-sized vocabulary. In addition, to specify the position of each pair of entities, we also use the positional embedding according to the relative distance between words and entity pair. For instance, in sentence "Yao Ming was born in Shanghai", the distance between born and Yao Ming is 2 and the distance between born and Shanghai is -2. The position information represents a kind of space relationship between words and entity pair. For each word, the position embedding is made of two parts, the first one is the embedding of the distance from the word to the head entity and the other one is the tail entity. The dimensions of the positional embedding and word embedding are $d_p$ and $d_w$ separately. Hence through the embedding layer, the sentence $x$ becomes a vector $w = \{w_1, w_2, \ldots, w_n\}$, where $w_i \in \mathbb{R}^{d_p + d_w \times 2}$.

2.2.2. Sentence Encoding.
In our model, Bi-GRU is used to get the representations of sentences. Firstly, a left-to-right GRU encodes inputs of a sentence to produce hidden output $h_\rightarrow \in \mathbb{R}^{n \times d_{\text{model}}}$, then as the same, a right-to-left GRU is used to get a hidden output $h_\leftarrow \in \mathbb{R}^{n \times d_{\text{model}}}$. Two hidden outputs is concatenated to get the output of Bi-GRU $h = [h_\rightarrow, h_\leftarrow] \in \mathbb{R}^{n \times 2d_{\text{model}}}$.

To maintain efficiency of proposed method, we apply the word attention mechanism to get the representation of sentences. Only a part of words of a sentence are helpful for discriminating the relation included [Sharmistha et al., 2018]. The degree of relevance of a word is calculated as an attention score in our model. We define $u_i$, the measure of relevance of the $i^{th}$ word in the sentence $x$ and it can be calculated as follows:

$$
\begin{align}
    u_i &= w_i \times A \times r; \\
    a_i &= \frac{\exp(u_i)}{\sum_{j=1}^{n} \exp(u_j)}; \\
    \hat{w_i} &= a_i \times w_i \\
\end{align}
$$

$A \in \mathbb{R}^{d_{\text{model}} \times d_{\text{model}}}$ is a matrix and $r \in \mathbb{R}^{d_{\text{model}} \times 1}$ is a relation query vector. Attention score $a_i$ is represented by taking the softmax operation over $\{u_i\}$. Hence, the input $w$ is transformed into the sentence representation $\hat{w} \in \mathbb{R}^{n \times d_{\text{model}}}$. Similar to Zeng[4], we apply the piecewise max pooling on $\hat{w}$ before, between, and after the entity pair. Finally, the output is $w_0 \in \mathbb{R}^{1 \times 3 \times d_{\text{model}}}$.

Figure 3: The framework of our approach.
2.2.3. Bag Aggregation.
According to the assumption of distant supervised RE, some sentences in bag is hardly related to the relation label. To solve this problem, Lin[8] proposed attention based sentences selecting method by assigning different weights to instances in a bag. We also follow this thought. Similar to the way in Sentence Encoding, we finally get the bag representation \( o_b \in \mathbb{R}^{1 \times 3 \times d_{model}} \). Then, a dense layer and a softmax layer are used to yield the probabilities scores of classes.

2.2.4. SDP Supervision.
In our model, SDP is applied to supervise the learning of word based attention mechanism. The score of each token represents the importance in a sentence, we think that words in SDP show apparent semantic relation and naturally it’s a thought to help the learning of weights with SDP. Specifically, the total loss of our model is made of two parts, classification loss and SDP supervision loss. The last one is calculated by the difference between SDP and weight scores.

3. Experiments and Result

3.1. Datasets.
We validate effectiveness of our models on different datasets summarized in Table 1. Riedel developed Riedel dataset by referring to Freebase knowledge base with New York Times articles. The training set are created with sentences from year 2005 and 2006, the test set is created from year 2007. GIDS is created by Jat[6] by the distant supervised assumption and makes sure the last one assumption of multi instance learning.

| Dataset       | relation      | Sentences    | Entity-pairs |
|---------------|---------------|--------------|--------------|
| Riedel2010-b  | Dataset with development set |              |              |
| Train         | 53            | 455,771      | 233,064      |
| Dev           | 53            | 114,317      | 58,635       |
| Test          | 53            | 172,448      | 96,678       |
| GIDS Dataset  |               |              |              |
| Train         | 5             | 11297        | 6498         |
| Dev           | 5             | 1864         | 1082         |
| Test          | 5             | 5663         | 3247         |

Baselines.
- Mintz: A logistic regression model by Mintz[1] for RE.
- MultiR: A probabilistic graphical model for distant supervision RE by Hoffmann[9]
- MIMLRE: A multi instance multi label model for distant supervision RE by Surdeanu.
- PCNN: A convolutional neural network RE model by Zeng[4].
- PCNN+ATT: A CNN based RE model by Lin[8].
- BGWA: Bi-GRU based distant supervision RE model by Jat[6].
3.2. Evaluation Criteria.
According to previous researches[6], held-out evaluation method is used to evaluate all models. And in our experiments, we use top-N precision (P@N) metric and Precision-Recall curve to evaluate all models.

3.3. Results.
For showing the performance of our model, all baselines are used to compare with it. The P-R curves on benchmarks are presented in Figure 4 and the top-N precision results are listed in table 2. We found that our method better results compared to the baselines on both the datasets and it shows that the SDP based word supervision attention mechanism is useful to improve distant supervised RE. At the same time, the better results of piecewise convolutional neural networks with attention over piecewise convolutional neural networks show attention mechanism matters in distant supervision relation extraction.

Table 2. P@N in Riedel dataset

|       | One          |          |          | Two          |          |          | All         |          |
|-------|--------------|----------|----------|--------------|----------|----------|-------------|----------|
|       | P@100        | P@200    | P@300    | P@100        | P@200    | P@300    | P@100       | P@200    |
| PCNN  | 73.30        | 64.80    | 56.80    | 70.30        | 67.20    | 63.10    | 72.30       | 69.70    |
| PCNN+ATT | 73.30      | 69.20    | 60.80    | 77.20        | 71.60    | 66.10    | 76.20       | 73.10    |
| BGWA  | 78.00        | 71.00    | 63.30    | 81.00        | 73.00    | 64.00    | 82.00       | 75.00    |
| BGWA+SDP | 77.00      | 72.00    | 67.41    | 77.03        | 75.00    | 70.03    | 77.02       | 76.00    |

4. Conclusion & Future Work
In this paper, we propose a distant supervision RE method, a novel neural network model which takes advantage of SDP, for improving distant supervised relation extraction. In our model, SDP is applied to supervise the learning of word attention weights. Through amounts of experiments on benchmark, we validate performance of our method over the baselines. The more usages of SDP in distant supervised relation extraction are left as future work.

Acknowledgments
This work is supported by Sichuan Science and Technology Planning Project(Grant No.2019YFG0495).

References
[1] Mintz, M., Bills, S., Snow, R., Jurafsky, D. (2009) 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. 1003-11011.

[2] Bollacker, K., Evans, C., Paritosh, P., Sturge, T., Taylor, J. (2008) Freebase: A collaboratively created graph database for structuring human knowledge. In: Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data. New York, NY, UAS. 1247-1250.

[3] Surdeanu, M., Tibshirani, J., Nallapati, R., Manning, C.D. (2012) 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. 455-465.

[4] Zeng, D.j., Liu, K., Lai, S.W., Zhou, G.Y., Zhao, J. (2014) Relation classification via convolutional deep neural network. In Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2335–2344.

[5] Riedel, S., Yao, L.M., McCallum, A. (2010) Modeling relations and their mentions without labeled text. In: Machine Learning and Knowledge Discovery in Databases. 148-163.

[6] Jat, S., Khandelwal, S., Talukdar, P. (2018) Improving Distantly Supervised Relation Extraction using Word and Entity Based Attention. ArXiv preprint.

[7] Bunescu, R., Mooney, R. (2005) A shortest path dependency kernel for relation extraction. In: Proceedings of the conference on human language technology and empirical methods in natural language processing, 724-731

[8] Lin, Y.K., Shen, S.Q., Liu, Z.Y. et al. (2016) Neural relation extraction with selective attention over instances. In: Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany. 1:2121-2133.

[9] Hoffmann, R., Zhang, C., Ling, X., Zettlemoyer, L. (2011) 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: 541-550.

[10] Bahdanau, D., Cho, K., Bengio, Y. (2014) Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv:1409.0473.

[11] Han, X., Sun, L. (2016) Global distant supervision for relation extraction. In: Thirtieth AAAI Conference on Artificial Intelligence.

[12] Vrandečić, D., Krotzsch, M. (2014) Wiki-data: A free collaborative knowledgebase. Commun. ACM, 57(10):78–85.

[13] Cho, K., Merrienboer, B.v., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y. (2014) Learning phrase representations using rnn encoder–decoder for statistical machine translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. 1724-1734.