Relation Extraction for Chinese Medicine Using ResCNN and Attention

Tianxiang Gao,[1] Xi Yang,[1] Zhuohao Zhang,[1] Linlin Li,[3] Zhou Zhao
[1]{gaotianxiang, 3150105002, zhuohaozhang}@zju.edu.cn, [3]linyan.lll@alibaba-inc.com
[1]College of Computer Science, Zhejiang University
[3]Alibaba Group
No.38 Zheda Road, Hangzhou, Zhejiang, China, 310027

Abstract

Chinese medicine text processing are important in medical research. Relation extraction in these texts helps accelerate related researches greatly. However, there is no model built specifically for these texts, and previously there is no large scale data set for training specified models. In this paper, we propose a solution to the Relation Extraction problem based on Chinese medicine dataset. We proposed an improved Residual Neural Network (ResCNN) with sentence-level attention over multiple instances to obtain better results, and constructed a large-scale dataset for NLP tasks rated to Chinese medicine texts. Our model is deployed in part of the knowledge base system at Alibaba.

Introduction

Relation Extraction(RE) is an important task in information extraction of natural language. RE identifies entities from unstructured text, and extracts semantic relations among these entities. Previous works have demonstrated that neural network models are powerful for RE tasks on general data sets, such as New York Times dataset.

Some recent research on relation extraction attempted distant supervision to integrate knowledge bases (KBs) with raw corpus to reduce noise. (Lin et al. 2016) proposed a sentence-level attention-based convolutional neural network(CNN) for relation extraction that can reduce the effect of wrong instances. (Huang and Wang 2017) designed ResCNN for distantly-supervised Relation Extraction and showed deeper CNN is more effective for noisy tasks. However, these models are designed for general corpus, and may give unsatisfying result for texts from certain domain.

Chinese medical text is very valuable in medical and pharmaceutical research. Extracting relation automatically from these texts can greatly accelerate researches in related fields. Chinese medical text typically comes with precise syntax and rigorous structure, which means they can require different model compared to generally purposed natural language corpus. However, previous relation extraction models for Chinese medical texts suffer from lack of large-scale available corpus. Such corpus are difficult to collect, because they are distributed in various not related sources. Besides, sometimes the corpus can contain unwanted noise such as mislabel or data disintegrity.

The main contribution of this paper is : 1).we propose a ResCNN model with sentence-level attention mechanism for RE task on Chinese medical text. Compared to CNN, our model is able to learn information within longer distance. Compared to ResCNN, our model introduced attention to tolerate noise. This model has been successfully deployed as a part of the Chinese medical knowledge base system in Alibaba Inc. 2). We constructed a large-scale Chinese medical dataset for training the model. To our knowledge, this is the first Chinese medical text dataset in the world.

Model Structure

Our model includes these components: embedding layer, residual convolutional layer, attention layer and utput layer. Figure 1 provides an overview of our model.

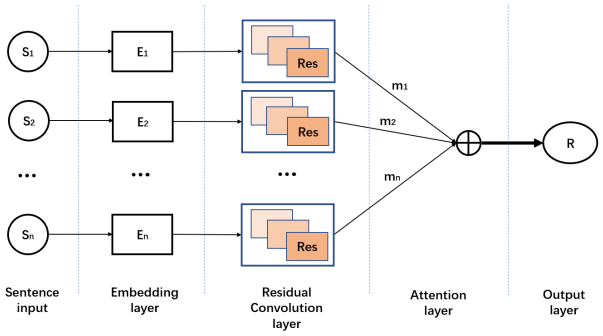

Figure 1: The overview of model structure

Embedding layer: This layer first divides Chinese sentence into word segments. Then, every word is embedded into a vector by an embedding matrix $E \in \mathbb{R}^{d^a \times |V|}$, where V is the vocabulary. We calculate the positional relationship between current word and entities in the sentence, then assign the corresponding vector with the size d_p to indicate the position of current word. Each sentence with s words will be represented as a vector sequence $S = \{w_1, w_2, ..., w_s\}$, where $w_i \in \mathbb{R}^d, d = d_w + d_p * 2$.

Residual Convolutional layer: We use residual convolutional neural network (ResCNN) to encode embedded vec-

Entity1	Entity2	Relation	Training	Testing
drug	disease	treat	17716	1044
drug	symptom	treat	19328	897
Any	Any	NA	5506	664

Table 1: Relationship in the dataset

tors. Residual neural network is composed from multiple convolutional layers. Each convolutional layer is:

$$\mathbf{p_i} = f(\mathbf{w} \cdot \mathbf{x}_i + \mathbf{b})$$

where $f()$ is the non-liner activation function. The input of first convolution layer is the sentence vector $\mathbf{x}_{0_i} = \{w_i, w_{i+1}, ..., w_{i+l-1}\}$. Each convolutional layer of it is followed by a Relu layer, and each residual layer is constructed by two convolutional layers. Residual layers shortcut the connection:

$$\mathbf{P}(\mathbf{x_i}) = \mathbf{x_i} + \mathbf{F}(\mathbf{x_i})$$

The input of each residual layer is the output p of its previous layer. After the last residual layer, a max-pooling layer $\widehat{p} = max(\mathbf{p})$ gives the maximum value in the output of residual layers. The final output $q = \{\widehat{p_1}, \widehat{p_2}, ..., \widehat{p_m}\}$ is merged from the output of all filters.

Attention layer: Our model integrated attention mechanism to assign weight to each symbol, and implicitly drop noisy symbols. For input sentences $T = \{s_1, s_2, ..., s_n\}$, we get a set of max value vector $Q = \{q_1, q_2, ..., q_n\}$ after convolutional layer, which is the input of the attention layer. After all the attention layers, the output are aggregated to give the final output of entity relations. In the back propagation stage, our model adopts cross-entropy loss as the loss function.

Experiments

To evaluate our model, we constructed a comprehensive Chinese medical corpus derived from open texts, including drug instructions, electronic medical records, medical QAs and some other related sources. Our proposed dataset involves corpus from multiple types of sources, but all follows the typical features of Chinese medical text. We correlated the relations in our data set according to knowledge bases, and added some negative correlation coefficients. The training set contains 42, 661 statements with an average length of 28.3 words and, and each statement includes exactly one relationship. The test set contains 2605 statements with an average length of 27.9 words. Each sentence in both training and testing set includes exactly two entities. Table1 shows the statistics of relationships in our datasets.

We compared our method with four baselines:

- **CNN:** CNN in (Zeng et al. 2014).
- **RNN:** recurrent neural network encoder.
- **CNN + ATT:** CNN with attention in (Lin et al. 2016)
- **ResCNN:** Residual convolution network with 9 layers in (Huang and Wang 2017)
- **PCNN + ATT:** Piecewise CNN with attention over instance learning in (Lin et al. 2016)
- **RNN + ATT:** RNN baseline with attention over instance learning implemented by ourselves.

Methods	Auc
CNN	58.36%
RNN	72.49%
CNN+ATT	92.89%
ResCNN	88.12%
PCNN+ATT	93.77%
RNN+ATT	86.13%
ResCNN+ATT	95.52%

Table 2: Experimental results on Accuracy

Table2 summarized our experiment result. The result shows that our model achieved state-of-the-art performance in Chinese medical relation extract task. We also observed that models with attention are always better than models without attention. Among all the models without attention, ResCNN has the best performance. Attempting various types of attention may further improve the performance of our model.

Conclusion

We proposed a novel neural network model with ResCNN and attention mechanism for relation extraction task on Chinese medical datasets. To train and evaluate the model, we constructed the first large-scale Chinese medicine text data set in the world. Our model achieved state-of-the-art performance in our data set. We hope that our Chinese medical data set can not only be used for our model, but also benefit other related researches. The successful deployment of our model in Alibaba demonstrated that our model, as well as the data set we constructed, are practical and valuable for industrial level application.

References

Huang, Y. Y., and Wang, W. Y. 2017. Deep residual learning for weakly-supervised relation extraction.

Lin, Y.; Shen, S.; Liu, Z.; Luan, H.; and Sun, M. 2016. Neural relation extraction with selective attention over instances. In *Meeting of the Association for Computational Linguistics*, 2124–2133.

Zeng, D.; Liu, K.; Lai, S.; Zhou, G.; and Zhao, J. 2014. Relation classification via convolutional deep neural network.