Discovering Medical Entity Relations from Texts using Dependency Information

Ying Shen[1*] , **Jiyue Huang**[1] , **Jin Zhang**[1] , **Min Yang**[2] , **Kai Lei**[1,3,*]
[1]Shenzhen Key Lab for Information Centric Networking Blockchain Technology (ICNLAB), School of Electronics and Computer Engineering (SECE), Peking University
[2]SIAT, University of the Chinese Academy of Sciences
[3]PCL Research Center of Networks and Communications, Peng Cheng Laboratory
shenying@pkusz.edu.cn, {huangjiyue, zhangjin}@sz.pku.edu.cn, min.yang@siat.ac.cn, leik@pkusz.edu.cn

Abstract

Relations between medical concepts convey meaningful medical knowledge. Relation extraction on medical corpus is an important task of information extraction and is the key step of building medical knowledge graph. However, medical entity relation extraction generally has two major issues: (i) In the medical domain, the sentences containing entities are longer than most general domain sentences, which poses the difficulty in capturing the relations between far apart entities; (ii) It is expensive to collect a large amount of training data for the medical domain. To tackle these challenges, we propose RED, a neural network for Relation Extraction that fully explores Dependency information and incorporates such information into deep neural networks. The long-range relation between entities can be captured by organizing a sentence as a dependency tree, while the requirement for a large amount of training data can be reduced with the abstract-level features generated by dependency information. Experiments on real-world medical dataset in Chinese language demonstrate that the proposed method achieves the best performance compared with the state-of-the-art methods.

1 Introduction

Nowadays, Knowledge Graphs (KGs) are receiving increasing attention as constructing a large-scale KG is important and useful for many real-world applications such as medical question answering, precision medicine, and drug discovery. In KGs, relations between entities can be expressed in triplets, for example, (**Flu, disease_has_symptom, Fever**) means the relation between "**flu**" and "**fever**" is "**disease_has_symptom**". In the medical field, KGs can express relations between medical entities, where these relations are projected from the real-world facts.

Relation extraction (RE) is an important natural language processing task. It plays a vital role in robust knowledge extraction from unstructured texts in medical domain and serves as an intermediate step in the medical KG construction. To extract entity relations, distant supervision methods are widely adopted [Mintz *et al.*, 2009; Zeng *et al.*, 2015; Lin *et al.*, 2016; Shen *et al.*, 2015],where training data are automatically generated via aligning texts with KGs. It is assumed that if there is a relation between two entities in KGs, then the sentences or text containing both these two entities will express their relation. For example, the sentence "*Fever is a symptom of flu.*" indicates the relation "**disease_has_symptom**" between entities "**fever**" and "**flu**" .

Despite the effectiveness of previous studies, there are several challenges to be addressed for distantly supervised medical relation extraction: (1) In medical domain, the sentences are usually longer than those in the general domain. Thus the distance between the target entities in the same paragraph tends to be longer, which poses the difficulty in capturing the relations between the entities that are far apart from each other. (2) The labeled relation examples are often insufficient due to the high labeling cost.

To alleviate the aforementioned challenges, this paper proposes a neural network, RED, that fully explores dependency information and incorporates such information into deep neural networks for the medical relation extraction. Specifically, we first represent a given sentence by the embeddings of the words it consists of. The dependency features are also transformed into real-valued representation. Then, we use a Bidirectional Gated Recurrent Unit (BiGRU) network with word-level and sentence-level attention models combined as an encoder to read the source sentence via the embeddings of the words in the sentence. Finally, we integrate dependency information into the deep neural network to tackle the long-distance problem in medical domain.

To evaluate the performance of the proposed RED method, extensive experiments are conducted on a real-world medical dataset in Chinese language. Comparing with the state-of-the-art methods, the proposed method achieves the best performance by incorporating dependency information.

Our contributions can be summarized as follows:

- To improve the relation extraction from cross-sentence long-text, we incorporate dependency information into deep neural networks to reflects actual word order and explore the Chinese linguistic features to enhance the word semantic information representation.

[*] Contact Author

- To relieve the wrong labeling problem in distant supervision, we leverage sentence-level and word-level attention to de-emphasize the noisy samples so that more attention is paid to the useful information.
- Experiments on real-world dataset demonstrate that our model consistently outperforms the state-of-the-art methods, and achieves comparable or better results over baseline methods with much less training data.

2 Preliminaries

2.1 Problem Definition

Given two medical entities (en_1 and en_2) and a set of sentences (noted as *Sent*) containing both of them, our model aims to predict the relation $r \in R$ of the given entities, where R is the relation set.

2.2 Linguistic Information Helps

Dependency parsing is based on linguistic information and dependency analysis [Fundel *et al.*, 2007] . A major advantage of dependency grammars is their ability to deal with languages that are morphologically rich and have a relatively long word distance. In the dependency parsing, by organizing the whole sentence into a dependency tree, dependency information can shorten the abstract distance between entities by organizing the whole sentence into a dependency tree [Tai *et al.*, 2015][Culotta, 2004]. Figure 1 gives an example of the structure of a dependency tree, which is generate by Standford parser. The number of hops from $entity_1$ to $entity_2$, which, between “flu” and “fever” is 8, where the distance in source text is 48 words. In this way, the linear sentence structure is transformed into a dense tree structure where the abstract distance between entities is shortened. Thus the long-distance relationship between two entities in a sentence can be better captured.

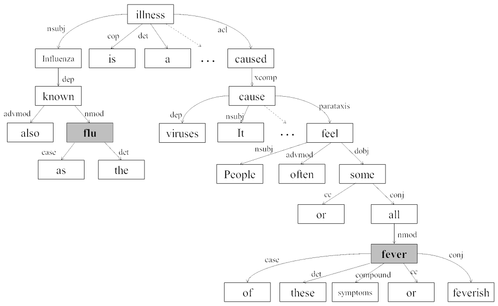

Figure 1: Dependency tree of the sentence containing “**flu**” (en_1) and “**fever**” (en_2). The linear sentence structure is transformed into a dense tree structure and the long-distance relationship between two entities in a sentence can be better captured. A part of the whole tree is shown here.

3 Methodology

In this section, we will develop our model RED in two main parts (see Figure 2). We first explore the entity features to learn the feature representation, so as to pre-train a set of given sentences. Then, we use a Bidirectional Gated Recurrent Unit (BiGRU) network with word-level and sentence-level attention mechanism to extract the relation between the target entities.

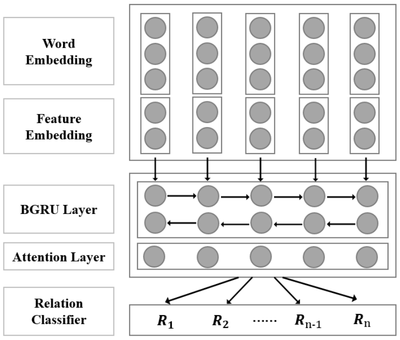

Figure 2: Overview of the RED architecture.

3.1 Input Representation

To feed training data to the neural network, the sentences we use are transformed into matrices. For a given sentence, it is represented by the embeddings of the words it consists of. As for the words, by looking up the pre-trained word embeddings, we represent them by real-valued vectors. Additional features used in the model are represented as vectors by looking up the corresponding embeddings.

- **Input representation**. For a given sentence, it is represented by the embeddings of the words it consists of. The dependency features are also transformed into real-valued representation.
- **BIGRU and attention models**. In this paper, we implement a Bidirectional Gated Recurrent Unit (BIGRU) network with word-level and sentence-level attention models combined.
- **Dependency Information**. Based on the deep neural network mentioned above, we add dependency information to it, so that more abstract information is provided.

Word Embeddings

The first layer of the network is word embedding layer which transforms words into representations that capture syntactic and semantic information about the words.

Given a sentence $Sent$ consisting of m words, each word w_i is converted into a real-valued vector. Therefore, the input to the next layer is a sequence of real-valued vectors. We choose Glove [Mikolov *et al.*, 2013] that is pre-trained on a medical corpus.

Feature Embeddings

Considering the intrinsic language difference, based on experimental comparison [Miyao *et al.*, 2008] we utilize HanLP[1] in the medical domain to capture the dependency and part-of-speech information. For experiments, the adopted dependency parsing component can automatically process

[1] https://github.com/hankcs/HanLP

multiple sentences without manually combining their dependency trees.

All features mentioned below are transformed into vectors by looking up the pre-trained embeddings. To transform dependency features into real-valued representations, we first count the possible number of a dependency feature. Then we use x-bit to encode. The select of x needs to be satisfied 2^x ¿ the possible number of dependency features.

All aforementioned features associate with word embedding following the format:

$$w_i = [r_i^w, r_i^{f1}, r_i^{f2}, ..., r_i^{fn}], \tag{1}$$

where r_i^w is the word embedding of $word_i$, r_i^{fj} refers to the j^{th} feature embeddings. They are concatenated to become optimized input of BiGRU.

Position Embeddings In a sentence, the words close to the target entities are usually informative for determining the relation between the two entities.

In our work, we utilize the relative word position. For each word in a sentence, the position feature is derived from the relative distances of the current word to the target entities en_1 and en_2. For instance, for the description mentioned above, the relative position of the word "influenza" to the target entity "**flu**" (en_1) is -6.

Part of Speech Embeddings Parts-of-speech (POS) aims to classify words based on grammatical properties (e.g., syntactic functions and morphological changes).

Given a sentence $Sent$, we employ Stanford parser to capture the POS information pos of each word, including numeral, cardinal, adverb, superlative, list item marker, modal auxiliary, etc. Then the word-related POS information pos is transformed into a vector via word2vec.

Dependency Embeddings To tackle the challenges in medical domain, dependency information is added to deep neural networks. Xu, Mou, Li, Chen, Peng and Jin [2015] use shortest dependency path (SDP) in long short term memory networks (LSTM) to classify relations, which focuses on the most relevant information for entities relation. However, due to the long-distance-relationship problem in the medical domain, some hidden contents may be ignored by the SDP method. So in our method, the complete dependency information is used to capture the long-distance relationship.

In the dependency tree, a sentence is represented as $Sentence = [word_0, word_1, ..., word_m]$, where $word_i$ represents the i^{th} word of the sentence, $word_0$ is actually a virtual node (also called root node, as *illness* in Figure 1), which is introduced artificially to emphasize the key word of dependency tree. An edge in dependency tree can be given by:

$$d = (h, c, l); 0hn, 0cn, l \in L. \tag{2}$$

where (h, c, l) stands for the dependency edge from $word_h$ to $word_c$, with the relation l out of dependency relation set L.

Stanford dependency parser is used in our model to extract dependency features based on paths in the dependency tree. By incorporating the abstract-level dependency information, the proposed method can capture the key information for recognizing a relation between entities quickly. For example, the dependency information âEntity-A is the CEO of Entity-Bâ can directly teach the proposed method to recognize the CEO relation quickly, while such abstract dependency information only can be learnt by seeing many training instances if we donât incorporate dependency information. Thus, the proposed method requires less training data by using dependency information.

Dependency information is obtained from the hierarchical tree structure, including relative dependency features and dependency tags. Relative dependency features show the relation between the current word to the root of the tree or the entities. Dependency tags imply the relation between the current word and its parent node in the dependency tree. Thus, the sentence and the distance can be shortened by dependency information.

Relative dependency features: Relative root feature implies the relation between current node and the root node. There are three types of relations here: the child node of the root, the root node itself, and others.

Relative entity feature implies the relation between current node and $entity_1$ and $entity_2$. There are four types of relations: the child node of $entity_1$/$entity_2$, the parent node of $entity_1$/$entity_2$, entity node itself, and others.

Dependency tags: the tag of the current word to its parent node on the dependency tree.

Finally, the word embeddings and feature embeddings are concatenated to form the input for the next layer as:

$$w_i = [r_i^w, r_i^{posi}, r_i^{root}, r_i^{e1}, r_i^{e2}, r_i^{dep}, r_i^{pos}] \tag{3}$$

For a word w_i, r_i^w is its word embedding, r_i^{posi} is the position embedding, r_i^{pos} is the POS embedding, r_i^{root} is the relative root feature embedding, r_i^{e1} is the relative en_1 feature embedding, r_i^{e2} is the relative en_2 feature embedding, and r_i^{dep} is the dependency tag embedding .

3.2 Attention-based Bidirectional GRU

A Gated Recurrent Unit (GRU) is proposed by Cho, Van Merriënboer, Gulcehre, Bahdanau, Bougares, Schwenk and Bengio [2014] to enable each recurrent unit to capture dependencies of different time scales. The GRU has gating units that modulate the flow of information inside the unit, without having separate memory cells [Chung *et al.*, 2014]. In GRU, we have:

-1ex

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]), \tag{4}$$

-1ex

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]), \tag{5}$$

-1ex

$$\widetilde{h}_t = tanh(W \cdot [r_t * h_{t-1}, x_t]), \tag{6}$$

0ex

$$h_t = (1 - z_t) * h_{t-1} + z_t * \widetilde{h}_t. \tag{7}$$

where σ is the sigmoid function, and $\widetilde{h}_t$ is the candidate activation at time t. z_t is an update gate which decides how much the unit updates the activation.

In principle, GRU are suitable for capturing relationships among sequential data. BiGRU introduces a second layer to the unidirectional GRU networks. The hidden to hidden connections flow in opposite temporal order. The model is therefore able to exploit information both from the past and the future.

We use BiGRU as an encoder to read the source sentence via the embeddings of the words in the sentence. Referring to the selective attention mechanism [Lin *et al.*, 2016], we apply the attention into the pooling layer to treat source representations as a memory and model the interaction between the decoder and the memory. The attention mechanism aims at recognizing which source sentences best represent the relation from the ones labeled through distant supervision.

In distant supervision, there are inevitably some wrong labeling errors. So in our model, sentence-level and word-level attention are complemented to de-emphasize the noisy samples. The sentence-level attention α_i can be computed by:

-1ex

$$\alpha_i = \frac{exp(e_i)}{\sum_j exp(e_j)}, \tag{8}$$

0ex

$$Sent_i = \sum_i a_i s, \tag{9}$$

where e_i scores the relativity between the sentence and the predicted relation. α_i is the weight of a set of sentences s containing a pair of entities.

In addition, we apply attention mechanism in the word level to obtain the attention weight and pay more attention to the useful information. The word-level attention α_w is given by: -1ex

$$M_w = tanh(H), \tag{10}$$

-1ex

$$a_w = softmax(W_w^T M_w), \tag{11}$$

0ex

$$Sent_w = Ha_w^T, \tag{12}$$

where $W_w \in R^{1\times 1}$, is a weight parameter vector, W_w^T is the transpose vector, and $a_w \in R^{g\times 1}$ is the normalized attention.

Then the conditional probability can be calculated through a softmax layer: -1ex

$$Sent = Sent_i a_i + Sent_w a_w. \tag{13}$$

0ex

$$p(i|Sent;\theta) = \frac{exp(Output_i)}{\sum_{j=1}^{|\mathbf{R}|}, exp(Output_j)}, \tag{14}$$

where **R** is the relation set and *Output* is the result from output layer, θ is the parameters of the model.

To train the neural network, cross-entropy is used to calculate the cost function: 0ex

$$J(\theta) = \sum_{j=0}^{|Sent|} log(r_i|Sent_i;\theta), \tag{15}$$

where $Sent$ is the sentence set, r is the relation, and θ stands for the parameters.

4 Experiments and Results

Experiments are conducted to demonstrate that adopting dependency information can tackle the long-sentence and large-entity-distance challenges met in the medical domain. With dependency information, our model can achieve better performance, and less data is required.

4.1 Dataset

For distant supervision, KGs are used to be aligned with sentences, so that they are automatically labeled.

A medical KG built by our research team is used for the medical text labeling [Shen *et al.*, 2018]. The KG is built based on data from online wikis, authoritative medical websites and medical industry knowledge bases, which contains around 80k entities and 600k relations. The relations include"**disease_has_symptom**" , "**disease_has_complication**" , "**symptom's_location**" , "**operation's_location**" , etc. Together with the relation "**Others**" (which implies there is no relation between two entities), there are 27 different types of relations.

Medical texts are derived from medical textbooks and online encyclopedia and no specific medical domain is restricted [2][3][4][5]. 207,480 sentences are used for the experiments. We randomly choose 20% of the data as testing data and treat the remaining as training data.

4.2 Implementation Details

The length of the sentences is limited to 100. For those whose length is less than 100, "BLANK" tokens are added to the tail of the sentences and make them up to 100. HanLP is utilized to extract the dependency features together with the POS features in the medical domain. Stochastic gradient descent (SGD) is used to minimize the cost function. We employ dropout to prevent overfitting. The hidden size of BiGRU is 280. Dimension of word embeddings is 100 and all other feature embeddings are 10-dimensional. Other hyperparameters include: learning rate 0.001, dropout probability 0.5, batch size 100.

Held-out evaluation is employed in our model. It compares the relation instances extracted from the testing sentences with those in the medical KG we use. Instead of time-consuming human evaluation, the held-out evaluation provides an approximate measure of precision under the assumption that the tested systems have similar performances in relation facts inside and outside the medical KG.

4.3 Performance Comparison

To demonstrate the effectiveness of our method, we empirically compare different distant supervision methods. The selected base models include:

(1) **CNN**, a convolutional deep neural network to extract lexical and sentence level features [Zeng *et al.*, 2014].

[2] http://www.xywy.com/

[3] http://www.a-hospital.com/

[4] https://baike.baidu.com/

[5] http://www.baike.com/

(2) **PCNN**, reducing the impact of noise and wrong label problems by employing Piecewise Max Pooling in convolutional neural network [Zeng *et al.*, 2015].

(3) **SDP-LSTM**, a novel neural network to classify the relation of two entities in a sentence by leveraging the shortest dependency path (SDP) with long short term memory (LSTM) units [Xu *et al.*, 2015].

All these baseline models are implemented with the sentence-level attention (-Att) proposed in the work of Lin et al. [2016] and the At-least-one multi-instance learning (-One) used in [Zeng *et al.*, 2014]. Also, a Bidirectional GRU network with attention mechanism (BIGRU-Att) [6] is taken for comparison. For these base models, we follow exactly the same parameter settings as those in their original papers.

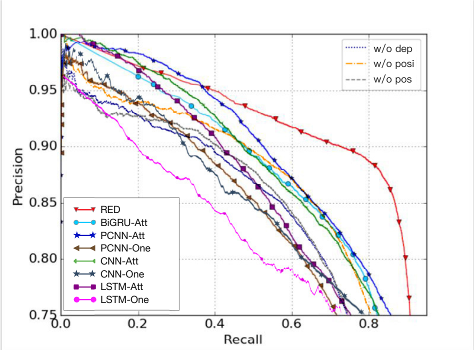

Figure 3: Aggregate precision/recall curves of CNN-One, CNN-Att, PCNN-One, PCNN-Att, LSTM-One, LSTM-Att, BIGRU-Att, RED in the medical domain.

We mark our attention-aware dependency-based model as RED. Held-out evaluation is conducted and the results are shown in Figure 3. In addition, in order to analyze the effectiveness of our method, we also report the ablation test in terms of discarding dependency information (w/o dependency), position embeddings (w/o posi), and POS tags embeddings (w/o pos), respectively. From the results, we can see that:

(1) In general, our model achieves the best performance. Especially when recall is larger than 0.4, our model outperforms other models by a large margin. We can also observe that in the medical domain, our model doesn't suffer from the sharp decline in the precision-recall curves occurring to CNN and PCNN models at very low recall. (2) The ablation results have demonstrated that all three factors contribute, with dependency embedding contributing most. This is within our expectation since dependency information shortens the abstract distance (hops) in the dependency tree between source and target entities, as well as introduces structural and syntactic information to enrich overall sentence representation. The dependency embedding can reduce the semantic ambiguity thus alleviate the difficulty of relation extraction from cross-sentence long-text.

4.4 Less Training Data

This part is to demonstrate the idea that by adopting dependency information, less training data is required for training the model. We conduct a series of experiments by gradually reducing the scale of the training data. Meanwhile, the scale of the testing data stays put. For comparison, the same testing data is used in the experiments.

The scales of training data are marked as $s_i \in \mathbf{S}$, and $\mathbf{S}$ = $\{100\%, 90\%, 80\%, ...\ 30\%, 20\%, 10\%\}$. It means that s_i of the whole training data is used. We take the BIGRU-Att model as comparison here, which is an evolved version of Lin et al. [2016]. From the results shown in Figure 4, we can see that: (1) Of all the $s_i \in \mathbf{S}$, our model achieves higher precision rate than BIGRU-Att, which proves the effect of dependency information; (2) Less data is required to get the same or even better result compared to methods without dependency information. For instance, the performance (0.896) of s_i = 20% in RED is higher than s_i = 100% in BIGRU-Att (0.891). It means that with only one fifth of the training data, our model can achieve the performance as good as what BIGRU-Att can do with the whole training data.

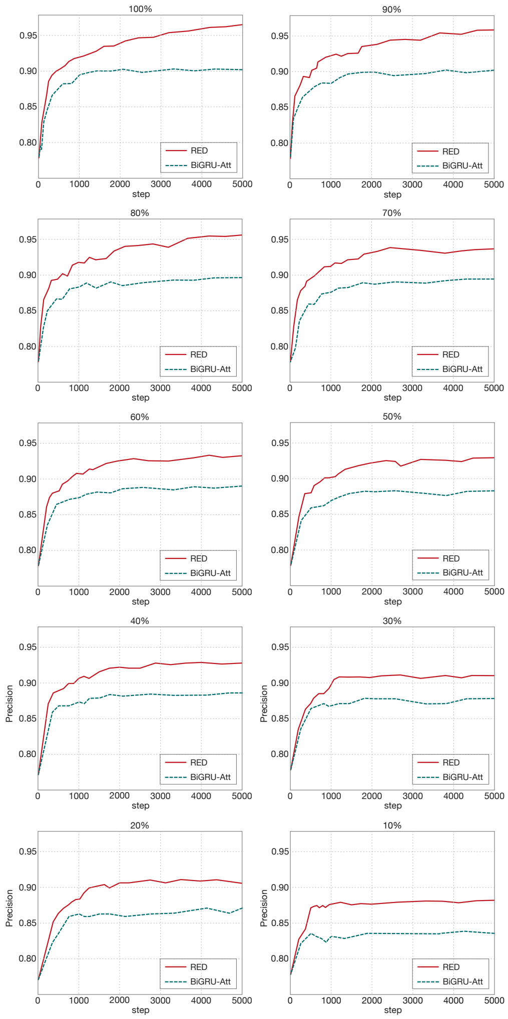

Figure 4: Precision over steps of each s_i in $\mathbf{S}$ = $\{100\%, 90\%, 80\%,$ ... , $30\%, 20\%, 10\%\}$

[6] https://github.com/thunlp/TensorFlow-NRE

These results indicate that dependency features can provide abstract-level features which can alleviate the requirement for a large amount of training data.

4.5 Case Study

Several representative examples are selected to demonstrate the improvement of relation extraction in the medical domain (see Table 1).

[Example 1]

"Lower extremity ***arteriosclerosis*** *can cause insufficient blood supply to the leg* ***muscles*** *and intermittent claudication. In severe cases, amputation may be required..."*

The probability of entity pair (atherosclerosis, muscle) belongs to (disease_site) and (symptom_site) are 0.708541 and 0.172634, respectively. Under different scenarios, arteriosclerosis can be either a disease or a symptom. In this sentence, our method can correctly identify the relation type between this pair-wise entity, which demonstrates the effectiveness of incorporating linguistic information into RE tasks.

[Example 2]

"To evaluate the occurrence, treatment, and outcome of ***hydrocephalus*** *complicating community-acquired* ***acute bacterial meningitis*** *in adults..."*

The probability of entity pair (acute bacterial meningitis, hydrocephalus) belongs to (disease_complication) and (disease_symptom) are 0.72062 and 0.00145511, respectively. Hydrocephalus is not a disease but a pathological result caused by a variety of causes. According to the context, RED can effectively handle the relation classification, which also demonstrates that our method can perform well on short-range text.

Table 1: Case study of RED and BiGRU-Att

	Text B	Text C
Entity Pair	Atherosclerosis, Muscle	Acute bacterial meningitis, Hydrocephalus
Length of Sentence	46	16
Entity distance	8	2
Relation	Disease_Site	Disease_Complication
Result of BiGRU-Att	0.4563	0.4069
Result of RED	**0.7085**	**0.72062**

5 Related Work

Relation extraction is one of the crucial areas in KG. There are mainly two types of approaches: supervised methods and distant supervised methods.

In supervised methods, relation extraction can be taken as a multi-classification problem. Many efforts have been focused on this. GuoDong, Jian, Jie and Min [2005]proposed a feature-based relation extraction method with the incorporation of diverse lexical, syntactic and semantic knowledge based on SVM. Kernel methods such as dependency tree kernel and subsequence kernel are used in some other work [Bunescu and Mooney, 2005], where the data is processed with the help of NLP tools. Recently, with the increasing popularity of deep learning [Bengio *et al.*, 2003], deep neural networks have been introduced here. Zeng, Liu, Lai, Zhou and Zhao [2014] extracted lexical and sentence level features using a convolutional deep neural network (CNN), and based on which a Classification by Ranking CNN (CR-CNN) model is proposed [Santos *et al.*, 2015]. Tree LSTM-like structure are used in reasoning, inference and sentiment analysis [Tai *et al.*, 2015].

These methods are effective in precision and recall. An inevitable drawback of supervised methods is that the data they use need human annotation and labeling, which is time-consuming and makes them hard to be applied to a large corpus.

To alleviate this limitation, distant supervision methods have been widely explored by existing work [Surdeanu *et al.*, 2012]. Distant supervision is one of the most important techniques in practice for relation extraction due to its ability to generate large-scale labeled training data automatically by aligning the plain text to KGs. They align the plain text to given KGs heuristically and thus a relation extracting model can be learned. The large scale of structured data in KGs is regarded as the distant supervision information. In the work of Mintz et al. [2009], features from all sentences are extracted and fed to a classifier. Graphical models are used to select sentences and predict relations in the works of Riedel et al. [2010], and they use multi-instance single-label learning or multi-instance multi-label learning to alleviate the wrong labeling problem. These are feature-based methods with the help of NLP tools. So in consequence, the quality of the generated features is important and there are error propagation problems.

6 Conclusion and Future Works

In this paper, we combine the data-driven and knowledge-driven methods together by adding dependency information to BiGRU with attention mechanism. With BIGRU and attention mechanism, useful information in the text can be automatically extracted without human labor, while the noise is de-emphasized. It is important to incorporate dependency parses, as we showed that the word distances between entities in medical text are typically longer than what observed in general domain.

Experiment results show that, our model achieves better results than the stat-of-the-art methods in the medical domain. Also, experiments show that in our model, less data is needed to achieve comparable or even better performance compared to model without dependency information.

Dependency is useful linguistic knowledge. With the coming era of deep learning, performance in RE is highly improved. However, the traditional linguistic knowledge should not be abandoned, for the pure data-driven methods cannot make full use of priori knowledge. We believe that the most promising avenues for future research include exploring how to combine data-driven model and linguistic model the key part of relation extraction, by considering the application of linguistic knowledge such as inversion transduction grammar, tree substitution grammar and tree adjoining grammar, and so on.

7 Acknowledgement

This work was financially supported by the National Natural Science Foundation of China (No.61602013), and the Shenzhen Fundamental Research Project (Grant No. JCYJ20170818091546869).

References

[Bengio *et al.*, 2003] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *J. Mach. Learn. Res.*, 3:1137–1155, March 2003.

[Bunescu and Mooney, 2005] Razvan C. Bunescu and Raymond J. Mooney. Subsequence kernels for relation extraction. *Advances in Neural Information Processing Systems*, pages 171–178, 2005.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart Van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder-decoder for statistical machine translation. *Computer Science*, 2014.

[Chung *et al.*, 2014] Junyoung Chung, Caglar Gulcehre, Kyung Hyun Cho, and Yoshua Bengio. Empirical evaluation of gated recurrent neural networks on sequence modeling. *Eprint Arxiv*, 2014.

[Culotta, 2004] Aron Culotta. Dependency tree kernels for relation extraction. In *In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics (ACL-04*, pages 423–429, 2004.

[Fundel *et al.*, 2007] Katrin Fundel, Robert Küffner, Ralf Zimmer, and Satoru Miyano. Relex–relation extraction using dependency parse trees. *Bioinformatics*, 23, 2007.

[Guodong *et al.*, 2005] Zhou Guodong, Su Jian, Zhang Jie, and Zhang Min. Exploring various knowledge in relation extraction. In *ACL 2005, Meeting of the Association for Computational Linguistics, Proceedings of the Conference, 25-30 June 2005, University of Michigan, Usa*, pages 419–444, 2005.

[Lin *et al.*, 2016] Yankai Lin, Shiqi Shen, Zhiyuan Liu, Huanbo Luan, and Maosong Sun. Neural relation extraction with selective attention over instances. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016.

[Mikolov *et al.*, 2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *Computer Science*, 2013.

[Mintz *et al.*, 2009] Mike Mintz, Steven Bills, Rion Snow, and Daniel Jurafsky. Distant supervision for relation extraction without labeled data. In *ACL 2009, Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 2-7 August 2009, Singapore*, pages 1003–1011, 2009.

[Miyao *et al.*, 2008] Yusuke Miyao, Rune Sætre, Kenji Sagae, Takuya Matsuzaki, and Jun'ichi Tsujii. Task-oriented evaluation of syntactic parsers and their representations. In *ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA*, pages 46–54, 2008.

[Riedel *et al.*, 2010] Sebastian Riedel, Limin Yao, and Andrew Mccallum. Modeling relations and their mentions without labeled text. In *European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 148–163, 2010.

[Santos *et al.*, 2015] Cicero Nogueira Dos Santos, Bing Xiang, and Bowen Zhou. Classifying relations by ranking with convolutional neural networks. *Computer Science*, pages 132–137, 2015.

[Shen *et al.*, 2015] Ying Shen, Joël Colloc, Armelle Jacquet-Andrieu, and Kai Lei. Emerging medical informatics with case-based reasoning for aiding clinical decision in multi-agent system. *Journal of Biomedical Informatics*, 56:307–317, 2015.

[Shen *et al.*, 2018] Ying Shen, Lizhu Zhang, Jin Zhang, Min Yang, Buzhou Tang, Yaliang Li, and Kai Lei. CBN: constructing a clinical bayesian network based on data from the electronic medical record. *Journal of Biomedical Informatics*, 88:1–10, 2018.

[Surdeanu *et al.*, 2012] Mihai Surdeanu, Julie Tibshirani, Ramesh Nallapati, and Christopher D Manning. Multi-instance multi-label learning for relation extraction. In *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 455–465, 2012.

[Tai *et al.*, 2015] Kai Sheng Tai, Richard Socher, and Christopher D. Manning. Improved semantic representations from tree-structured long short-term memory networks. *Computer Science*, 5(1):: 36., 2015.

[Xu *et al.*, 2015] Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. Classifying relations via long short term memory networks along shortest dependency paths. In *EMNLP*, pages 1785–1794, 2015.

[Zeng *et al.*, 2014] D. Zeng, K. Liu, S. Lai, G. Zhou, and J. Zhao. Relation classification via convolutional deep neural network. 2014.

[Zeng *et al.*, 2015] Daojian Zeng, Kang Liu, Yubo Chen, and Jun Zhao. Distant supervision for relation extraction via piecewise convolutional neural networks. In *Conference on Empirical Methods in Natural Language Processing*, pages 1753–1762, 2015.