Structure Regularized Neural Network for Entity Relation Classification for Chinese Literature Text

Ji Wen[1], Xu Sun[1,2], Xuancheng Ren[1], Qi Su[3]

[1]MOE Key Lab of Computational Linguistics, School of EECS, Peking University
[2]Deep Learning Lab, Beijing Institute of Big Data Research, Peking University
[3]School of Foreign Languages, Peking University
`{wenjics, xusun, renxc, sukia}@pku.edu.cn`

Abstract

Relation classification is an important semantic processing task in the field of natural language processing. In this paper, we propose the task of relation classification for Chinese literature text. A new dataset of Chinese literature text is constructed to facilitate the study in this task. We present a novel model, named Structure Regularized Bidirectional Recurrent Convolutional Neural Network (SR-BRCNN), to identify the relation between entities. The proposed model learns relation representations along the shortest dependency path (SDP) extracted from the structure regularized dependency tree, which has the benefits of reducing the complexity of the whole model. Experimental results show that the proposed method significantly improves the F_1 score by 10.3, and outperforms the state-of-the-art approaches on Chinese literature text[1].

1 Introduction

Relation classification is the task of identifying the semantic relation holding between two nominal entities in text. Recently, neural networks are widely used in relation classification. Wang et al. (2016) proposes a convolutional neural network with two levels of attention. Zhang et al. (2015) uses bidirectional long short-term memory networks to model the sentence with sequential information. Bunescu and Mooney (2005) first uses SDP between two entities to capture the predicate-argument sequences. Wang et al. (2017) explores the idea of incorporating syntactic parse tree into neural networks. Liu et al. (2017) proposes a noise-tolerant method to deal with wrong labels in distant-supervised relation extraction with soft labels. In recent years, we

have seen a move towards deep learning architectures. Liu et al. (2015) develops dependency-based neural networks. Xu et al. (2015) applies long short term memory (LSTM) (Hochreiter and Schmidhuber, 1997) based recurrent neural networks (RNNs) along with the SDP.

In this paper, we focus on relation classification of Chinese literature text, which to our knowledge has not been studied before, due to the challenge. Chinese literature text tends to express intuitions and feelings. It has a wide range of topics. Many literature articles express feelings in a subtle and special way, making it more difficult to recognize entities. Chinese literature text is not organized very logically, whether among paragraphs or sentences. They tend to use various and flexible forms of sentences to create free feelings. The sentences are not associated with each other by evident conjunctions. Besides, Chinese is a topic-prominent language, the subject is usually covert and the usage of words is relatively flexible.

In short, sentences of Chinese literature text contain many non-essential words, and embody very complex and flexible structures. Existing methods make intensive use of the syntactical information, such as part-of-speech tags, and dependency relations. However, the automatically generated information is not reliable and of poor quality for Chinese literature text. It is of great challenge for the existing methods to achieve satisfying performance.

To mitigate the noisy syntactical information, we propose to apply structure regularization to the structures used in relation classification. Recently, many existing systems on structured prediction focus on increasing the level of structural dependencies within the model. However, the theoretical and experimental study of Sun (2014a) suggests that complex structures are tend to increase the overfitting risk, and can potentially be harm-
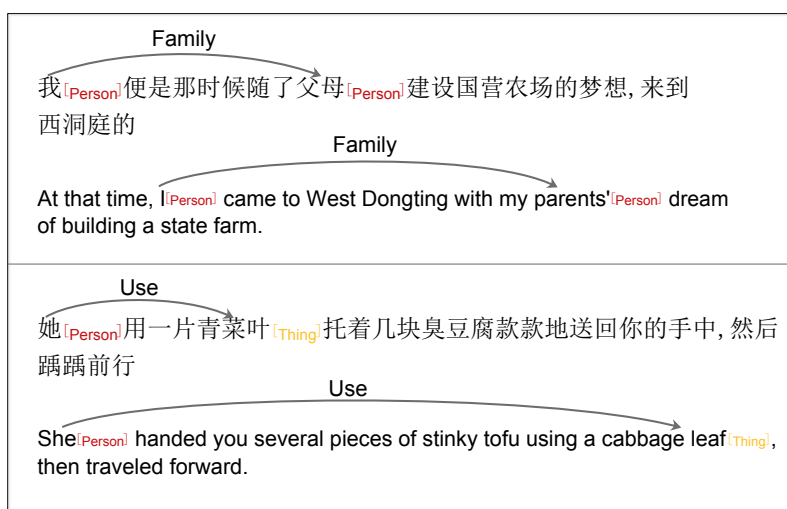

Figure 1: Examples from the Chinese literature text corpus.

ful to the model accuracy. As pointed out by Sun (2014a), complex structural dependencies have a drawback of increasing the generalization risk, because more complex structures are easier to suffer from overfitting.

In this paper, we focus on the study of applying structure regularization to the relation classification task of Chinese literature text. To summarize, the contributions of this paper are as follows:

- To our knowledge, we are the first to develop a corpus of Chinese literature text for relation classification. The corpus contains 837 articles. It helps alleviate the dilemma of the lack of corpus in Chinese Relation Classification.

- We develop the tree-based structure regularization methods and make progress on the task of relation classification. The method of structure regularization is normally used on the structure of sequences, while we find a way to realize it on the structure of trees. Comparing to the original model, applying structure regularization substantially improves the F_1 score by 10.3.

2   Chinese Literature Text Corpus

In Figure 1, we show two examples from the annotated corpus. We label the entities and relations of the text on a sentence level. There are 6 kinds of entities and 9 kinds of relations. Details of the tags are shown in Table 1. The task aims at predicting the labels of these relations, given the sentences as well as the entities and their types. The corpus is part of the work of Xu et al. (2017).

We obtain over 1,000 Chinese prose articles from the Internet and then filter and extract 837 articles. Articles that are too short or too noisy are not included. Due to the difficulty of tagging Chinese prose text, we divide the annotation process into three steps.

First, we attempt to annotate the raw articles based on defined entity and relation tags. Second, we design several generic disambiguation rules to ensure the consistency of annotation guidelines. For example, remove all adjective words and only tag "entity header" when tagging entities (e.g., change "a girl in red cloth" to "girl"). In this stage, we re-annotate all articles and correct all inconsistency entities based on the heuristic rules. Even though the heuristic tagging process significantly improves dataset quality, it is too hard to handle all inconsistency cases based on limited heuristic rules. Finally, we introduce a machine auxiliary tagging method. The core idea is to train a model to learn annotation guidelines on the subset of the corpus and produce predicted tags on the rest data. The predicted tags are used to be compared with the gold tags to discovery inconsistent entities, which largely reduce annotators' efforts. After all annotation steps, we also manually check all entities and relations to ensure the correctness of corpus.

In prior work, Chinese literature text corpus is very rare. Many tasks cannot achieve a satisfying result on Chinese literature text compared to other corpus. However, understanding Chinese literature text is of great importance to Chinese literature research.

Tag	Description	Example	%
Located	locate in	幽兰(orchid)-山谷(valley)	37.43
Part-Whole	be a part of	花(flower)-仙人掌(cactus)	23.76
Family	be family members	母亲(mother)-奶奶(grandmother)	10.25
General-Special	be a general range and a special kind of it	鱼(fish)-鲫鱼(carp)	6.99
Social	be socially related	母亲(mother)-邻里(neighbour)	6.02
Ownership	be in possession of	村民(villager)-旧屋(house)	5.10
Use	do something with	爷爷(grandfather)-毛笔(brush)	4.76
Create	make happen or exist	男人(man)-陶器(pottery)	2.93
Near	a short distance away	山(hill)-县城(town)	2.76

Table 1: The set of relation tags. The last column indicates each tag's relative frequency in the full annotated data.

3 Structure Regularized BRCNN

3.1 Basic BRCNN

The Bidirectional Recurrent Convolutional Neural Network (BRCNN) model is used to learn representations with bidirectional information along the shortest dependency path (SDP).

Given a sentence and its dependency tree, we build our neural network on its SDP extracted from tree. Along the SDP, recurrent neural networks are applied to learn hidden representations of words and dependency relations, respectively. A convolution layer is applied to capture local features from hidden representations of every two neighbor words and the dependency relations between them. A max pooling layer thereafter gathers information from local features of the SDP and the inverse SDP. We have a softmax output layer after pooling layer for classification in the unidirectional model RCNN.

On the basis of RCNN model, we build a bidirectional architecture BRCNN taking the SDP and the inverse SDP of a sentence as input. During the training stage of a (K+1)-relation task, two fine-grained softmax classifiers of RCNNs do a (2K+1)-class classification respectively. The pooling layers of two RCNNs are concatenated and a coarse-grained softmax output layer is followed to do a (K+1)-class classification. The final (2K+1)-class distribution is the combination of two (2K+1)-class distributions provided by fine grained classifiers during the testing stage.

We use two bidirectional LSTMs to capture the features of words and relations separately. After we obtain representations of words and relations, we concatenate them to get a representation of a complete dependency unit. The hidden state of a relation is denoted as r_{ab}. Words on its sides have the hidden states denoted as h_a and h_b. [h_a

h_{ab} h_b] denotes the representation of a dependency unit L_{ab}. Then we utilize a convolution layer upon the concatenation. We have

$$L_{ab} = f(W_{con} \cdot [h_a \oplus h'_{ab} \oplus h_b] + b_{con}) \quad (1)$$

where W_{con} is the weight matrix and b_{con} is a bias term. We choose $tanh$ as our activation function and apply max pooling following the activation.

Two RCNNs pick up information along the SDP and its reverse. A coarse-grained softmax classifier is applied on the global representations $\overrightarrow{G}$ and $\overleftarrow{G}$. Two fine-grained softmax classifier are applied to to give a more detailed prediction of 2K+1 class.

$$\overrightarrow{y} = softmax(W_f \cdot \overrightarrow{G} + b_f) \quad (2)$$

$$\overleftarrow{y} = softmax(W_f \cdot \overleftarrow{G} + b_f) \quad (3)$$

During training, our objective is the penalized cross-entropy of three classifiers. Formally,

$$J = \sum_{i=1}^{2K+1} \overrightarrow{t_i} \log \overrightarrow{y_i} + \sum_{i=1}^{2K+1} \overleftarrow{t_i} \log \overleftarrow{y_i}$$
$$+ \sum_{i=1}^{K} t_i \log y_i + \lambda \cdot \|\theta\|^2 \quad (4)$$

When decoding, the final prediction is a combination of $\overrightarrow{y}$ and $\overleftarrow{y}$

$$y_{test} = \alpha \cdot \overrightarrow{y} + (1 - \alpha) \cdot z(\overleftarrow{y}) \quad (5)$$

where α is the fraction of the composition of distributions. We apply a function z to transform $\overleftarrow{y}$ to a corresponding forward distribution like $\overrightarrow{y}$.

3.2 Structure Regularized BRCNN

The basic BRCNN model can handle the task to some extent, but there still remains some weakness, especially dealing with long sentences with

367

	Models	Information	F_1 Score
Baselines	SVM (Hendrickx et al., 2010)	Word embeddings, NER, WordNet, HowNet, POS, dependency parse, Google n-gram	48.9
	RNN (Socher et al., 2011)	Word embeddings	48.3
		+ POS, NER, WordNet	49.1
	CNN (Zeng et al., 2014)	Word embeddings	47.6
		+ word position embeddings, NER, WordNet	52.4
	CR-CNN (dos Santos et al., 2015)	Word embeddings	52.7
		+ word position embeddings	54.1
	SDP-LSTM (Xu et al., 2015)	Word embeddings	54.9
		+ POS + NER + WordNet	55.3
	DepNN (Liu et al., 2015)	Word embeddings, WordNet	55.2
	BRCNN (Cai et al., 2016)	Word embeddings	55.0
		+ POS, NER, WordNet	55.6
Our Model	**SR-BRCNN**	Word embeddings	**65.2 (+9.6)**
		+ POS, NER, WordNet	**65.9 (+10.3)**

Table 2: Comparison of relation classification systems on Chinese literature text.

complicated structures. The SDP generated from a more complicated dependency tree consists more irrelevant words. Sun (2014b) shows both theoretically and empirically that structure regularization can effectively control overfitting risk and lead to better performance. Sun et al. (2017a) and Sun et al. (2017b) also show that complex structure models are prone to the structure-based overfitting. Therefore, we propose the structure regularized BRCNN.

We conduct structure regularization on the dependency tree of the sentences. Based on the heuristic rules, several nodes in the dependency tree are selected. The subtrees of these selected nodes are cut from the whole dependency tree. With these selected nodes as the roots, these subtrees form a forest. The forest will be connected by lining the roots of the trees of the forest. Traditional SDP is extracted directly from the dependency tree, while in our model, the SDP is extracted from the final forest. We call these kinds of SDPs as SR-SDPs. We build our BRCNN model on the SR-SDP.

3.3 Various Structure Regularization Methods

We experiment with three kinds of regularization rules. First, the punctuation is a natural break point of the sentence. The resulting subtrees usually keep similar syntax to traditional dependency trees. Another popular method to regularize the structure is to decompose the structure randomly. In our model, we randomly select several nodes in the dependency tree and then cut the subtrees under these nodes. Finally we decide to cut the dependency tree by prepositions. Especially in Chi-

(a) The dependency tree and the SDP before flattening.

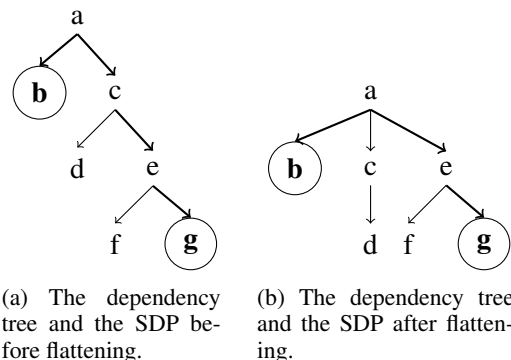(b) The dependency tree and the SDP after flattening.

Figure 2: An example of the proposed method. The two words in circles are the entities, and the thick edges form the SDP. By flattening the dependency tree, the path becomes shorter.

nese literature text, there usually are many decorations to describe the entities, and the using of prepositional phrases is very common for that purpose. So we also try to decompose the dependency trees using prepositions.

4 Experiments

We evaluate our model on the Chinese literature text corpus. It contains 9 distinguished types of relations among 837 articles. The dataset contains 695 articles for training, 58 for validation, and 84 for testing.

4.1 Experiment settings

We use pre-trained word embeddings, which are trained on Gigaword with word2vec (Mikolov et al., 2013). Word embeddings are 200-dimensional. The embeddings of relation are initialized randomly and are 50-dimensional. The

hidden layers of LSTMs to extract information from entities and relations are the same as the embedding dimension of entities and relations. We applied L2 regularization to weights in neural networks and dropout to embeddings with a keep probability 0.5. AdaDelta (Zeiler, 2012) is used for optimization.

4.2 Experimental Results

Table 2 compares our SR-BRCNN model with other state-of-the-art methods on the corpus of Chinese literature text, including the basic BR-CNN method (Cai et al., 2016). Structure regularization helps improve the result substantially. The method of structure regularization could prevent the overfitting of poor quality SDPs.

4.3 Analysis: Effect of SR

Figure 2a and Figure 2b show an example of structure regularized SDP. The relation is between the two circled elements. The main idea of the method is to avoid the incorrect structure from the dependency trees generated by the parser. The SDP in Figure 2a is longer than the SR-SDP in Figure 2b. However, the dependency tree of the example is not completely correct. The longer the SDP is, the more incorrect information the model learns.

The structure regularized BRCNN has shown obvious improvements. We attribute the improvements to the simplified structures that generated by structure regularization. The internal relations of components of a sentence are more obscure due to the feature of Chinese literature text. By conducting structure regularization on the dependency tree, we get several subtrees with simpler structure, and then we extract SDP from the lined forests. In most cases, the distance between two entities will be shortened along the new SR-SDP. Without the redundant information along the original SDP. The model that benefits from the intensive dependencies will capture more effective information for classification.

4.4 Analysis: Effect of Different Regularization Methods

The punctuation is a natural break point of the sentence, which makes subtrees more like the traditional dependency trees in the aspect of integrity. However, the original dependency trees cannot be sufficiently regularized. Despite its drawbacks, this method still shows obvious improvements on the model and leads to further experiments.

Classifier	F_1 score
BRCNN	55.6
SR by punctuation	59.7
SR by random	62.4
SR by preposition	65.9

Table 3: Different structure regularization results on Chinese literature text.

Regularizing the structure by decomposing the structure randomly will solve the insufficient decomposition problems. The method of structure regularization has shown that the degree of loss of information is not a serious problem. It gives a slightly better result compared to cutting dependency trees by punctuations.

A more elaborate method is to cut the dependency tree by prepositions. In Chinese literature text, prepositional phrases are used frequently. Cutting by prepositions will regularize the tree more sufficiently. Meanwhile, the subtrees under the prepositional nodes are usually internally linked.

5 Conclusions

In this paper, we present a novel model, Structure Regularized BRCNN, to classify the relation of two entities in a sentence. We demonstrate that tree-based structure regularization can help improve the results, while the method is normally used in sequence-based models before. The proposed structure regularization method makes the SDP shorter and contain less noise from the unreliable parse trees. This leads to substantial improvements on the relation classification results. The results also show how different ways of regularization act in the model of BRCNN.

We also develop a corpus on Chinese literature text focusing on the task of Relation Classification. The new corpus is large enough for us to train models and verify the models.

References

Razvan Bunescu and Raymond Mooney. 2005. A shortest path dependency kernel for relation extraction. In *Proceedings of Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 724–731.

Rui Cai, Xiaodong Zhang, and Houfeng Wang. 2016. Bidirectional recurrent convolutional neural network for relation classification. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 756–765.

Iris Hendrickx, Su Nam Kim, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó. Séaghdha, Sebastian Padó, Marco Pennacchiotti, Lorenza Romano, and Stan Szpakowicz. 2010. Semeval-2010 task 8: Multi-way classification of semantic relations between pairs of nominals. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 33–38.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Tianyu Liu, Kexiang Wang, Baobao Chang, and Zhifang Sui. 2017. A soft-label method for noise-tolerant distantly supervised relation extraction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1790–1795.

Yang Liu, Furu Wei, Sujian Li, Heng Ji, Ming Zhou, and Houfeng WANG. 2015. A dependency-based neural network for relation classification. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 285–290.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Cicero dos Santos, Bing Xiang, and Bowen Zhou. 2015. Classifying relations by ranking with convolutional neural networks. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 626–634.

Richard Socher, Jeffrey Pennington, Eric H. Huang, Andrew Y. Ng, and Christopher D. Manning. 2011. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 151–161.

Xu Sun. 2014a. Structure regularization for structured prediction. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014*, pages 2402–2410.

Xu Sun. 2014b. Structure regularization for structured prediction: Theories and experiments. *CoRR*, abs/1411.6243.

Xu Sun, Xuancheng Ren, Shuming Ma, Bingzhen Wei, Wei Li, and Houfeng Wang. 2017a. Training simplification and model simplification for deep learning: A minimal effort back propagation method. *CoRR*, abs/1711.06528.

Xu Sun, Weiwei Sun, Shuming Ma, Xuancheng Ren, Yi Zhang, Wenjie Li, and Houfeng Wang. 2017b. Complex structure leads to overfitting: A structure regularization decoding method for natural language processing. *CoRR*, abs/1711.10331.

Linlin Wang, Zhu Cao, Gerard de Melo, and Zhiyuan Liu. 2016. Relation classification via multi-level attention CNNs. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1298–1307.

Yizhong Wang, Sujian Li, Jingfeng Yang, Xu Sun, and Houfeng Wang. 2017. Tag-enhanced tree-structured neural networks for implicit discourse relation classification. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 496–505.

Ji Wen. 2017. Structure regularized bidirectional recurrent convolutional neural network for relation classification. *CoRR*, abs/1711.02509.

Jingjing Xu, Ji Wen, Xu Sun, and Qi Su. 2017. A discourse-level named entity recognition and relation extraction dataset for chinese literature text. *CoRR*, abs/1711.07010.

Yan Xu, Lili Mou, Ge Li, Yunchuan Chen, Hao Peng, and Zhi Jin. 2015. Classifying relations via long short term memory networks along shortest dependency paths. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1785–1794.

Matthew D. Zeiler. 2012. ADADELTA: an adaptive learning rate method. *CoRR*, abs/1212.5701.

Daojian Zeng, Kang Liu, Siwei Lai, Guangyou Zhou, and Jun Zhao. 2014. Relation classification via convolutional deep neural network. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 2335–2344.

Shu Zhang, Dequan Zheng, Xinchen Hu, and Ming Yang. 2015. Bidirectional long short-term memory networks for relation classification. In *Proceedings of the 29th Pacific Asia Conference on Language, Information and Computation*, pages 73–78.