Chinese Semantic Role Labeling with Bidirectional Recurrent Neural Networks

Zhen Wang, Tingsong Jiang, Baobao Chang, Zhifang Sui
Key Laboratory of Computational Linguistics, Ministry of Education
School of Electronics Engineering and Computer Science, Peking University
Collaborative Innovation Center for Language Ability, Xuzhou 221009 China
wzpkuer@gmail.com, {tingsong, chbb, szf}@pku.edu.cn

Abstract

Traditional approaches to Chinese Semantic Role Labeling (SRL) almost heavily rely on feature engineering. Even worse, the long-range dependencies in a sentence can hardly be modeled by these methods. In this paper, we introduce bidirectional recurrent neural network (RNN) with long-short-term memory (LSTM) to capture bidirectional and long-range dependencies in a sentence with minimal feature engineering. Experimental results on Chinese Proposition Bank (CPB) show a significant improvement over the state-of-the-art methods. Moreover, our model makes it convenient to introduce heterogeneous resource, which makes a further improvement on our experimental performance.

1 Introduction

Semantic Role Labeling (SRL) is defined as the task to recognize arguments for a given predicate and assign semantic role labels to them. Because of its ability to encode semantic information, there has been an increasing interest in SRL on many languages (Gildea and Jurafsky, 2002; Sun and Jurafsky, 2004). Figure 1 shows an example in Chinese Proposition Bank (CPB) (Xue and Palmer, 2003), which is a Chinese corpus annotated with semantic role labels.

Traditional approaches to Chinese SRL often extract a large number of handcrafted features from the sentence, even its parse tree, and feed these features to statistical classifiers such as CRF, MaxEnt and SVM (Sun and Jurafsky, 2004; Xue, 2008; Ding and Chang, 2008; Ding and Chang, 2009; Sun, 2010). However, these methods suffer from three major problems. Firstly, their performances are heavily dependent on feature engineering, which needs domain knowledge and laborious work of feature extraction and selection. Secondly, although sophisticated features are designed, the long-range dependencies in a sentence can hardly be modeled. Thirdly, a specific annotated dataset is often limited in its scalability, but the existence of heterogenous resource, which has very different semantic role labels and annotation schema but related latent semantic meaning, can alleviate this problem. However, traditional methods cannot relate distinct annotation schemas and introduce heterogeneous resource with ease.

WORD:	警察	正在	调查	事故	原因
	Police	now	investigate	accident	cause
ROLE:	[A0]	[AM-TMP]	REL	[A1]	
IOBES:	S-A0	S-AM-TMP	REL	B-A1	E-A1

Figure 1: A sentence with semantic roles labeled from CPB.

Concerning these problems, in this paper, we propose bidirectional recurrent neural network (RNN) with long-short-term memory (LSTM) to solve the problem of Chinese SRL. Our approach makes the following contributions:

- We formulate Chinese SRL with bidirectional LSTM RNN model. With bidirectional RNN, the dependencies in a sentence from both directions can be captured, and with LSTM architecture, long-range dependencies can be well modeled. The test results on the bechmark dataset CPB show a significant improvement over the state-of-the-art methods.
- Compared with previous work that relied on a huge number of handcrafted features, our model can achieve much better performance only with minimal feature engineering.
- The framework of our model makes the introduction of heterogeneous resource efficient

Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 1626–1631,
Lisbon, Portugal, 17-21 September 2015.

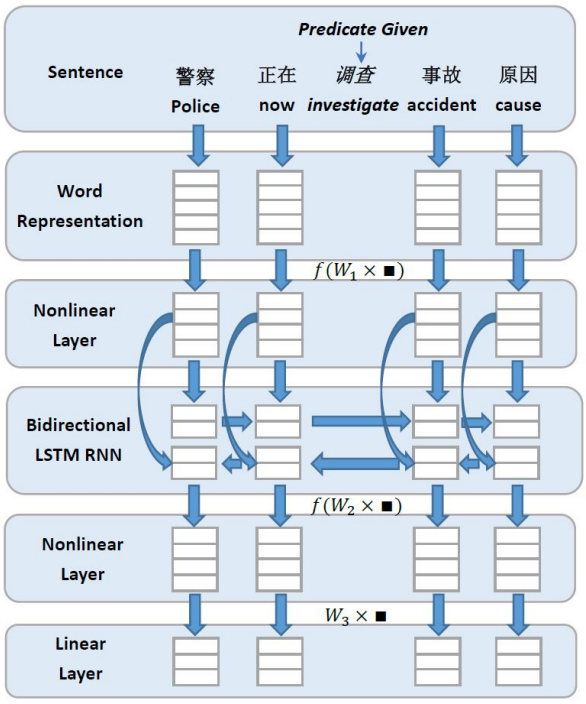

Figure 2: The model architecture.

and convenient, and this can further improve our experimental performance.

2 Chinese SRL with RNN

Following previous work, we regard Chinese SRL as a task of sequence labeling, which assigns a label for each word in the sequence. To identify the boundary information of semantic roles, we adopt the IOBES tagging schema for the labels as shown in Figure 1. For sequence labeling, it is important to capture dependencies in the sequence, especially for the problem of SRL, where the semantic role label for a word not only relies on its local information, but also is determined by long-range dependencies from other words. The advantage of RNN is the ability to better capture the contextual information, which is beneficial to capture dependencies in SRL. Moreover, we enrich the basic RNN model with bidirectional LSTM RNN, which can model bidirectional and long-range dependencies simultaneously.

2.1 Model Architecture

The architecture of our approach is illustrated in Figure 2. Given a sentence, we first get representation for each word to be labeled. Then after a nonlinear transformation, bidirectional LSTM RNN layer is designed to combine the local information of a word and its contextual information from both directions. With a nonlinear layer to form more complex features, a linear output layer follows. For each word to be labeled, there is an output vector, whose each dimension is a score corresponding to a kind of semantic role label.

2.2 Word Representation

Word representation captures the features locally embedded around the word. The features used in our work are: the current word, the current POS tag, the predicate, left and right words, left and right POS tags, distance to the predicate. Note that these features are all basic information about the word, hence we alleviate the heavy job of feature engineering. All these features are introduced by embeddings. After concatenation, we get the word representation feature vector.

To get more complex features, we adopt a nonlinear transformation:

$$z_t = f(W_1 x_t) \quad 1 \leq t \leq N$$

where x_t is the word representation of the t-th word, $W_1 \in \mathbb{R}^{n_1 \times n_0}$, n_0 is the length of word representation, f is an activation function and we use *tanh* in our experiments, N is the number of words to be labeled in the sequence.

2.3 Bidirectional LSTM RNN

Representation z_t only captures the local information. Here we adopt RNN to capture contextual information. Traditional RNN has the problem of vanishing or exploding gradients, which means the long-term dependencies can hardly be modeled. LSTM is designed to mitigate this problem.

At each word position t, the LSTM RNN computes six internal vectors: $\widetilde{C}$, g_i, g_f, g_o, C_t and h_t for the memory cell, which is a structure used in LSTM to store information. $\widetilde{C}$ computes the candidate value for the state of the memory cell:

$$\widetilde{C} = f(W_c z_t + U_c h_{t-1} + b_c)$$

The activations of the memory cell's input gate, forget gate and output gate are defined as:

$$g_j = \sigma(W_j z_i + U_j h_{t-1} + b_j)$$

where j stands for i, f or o. σ is taken $sigmoid$ in experiments. Then we can compute C_t, the memory cell's new state at position t:

$$C_t = g_i \odot \widetilde{C} + g_f \odot C_{t-1}$$

where $\odot$ indicates elementwise vector multiplication. With the new state of the memory cell, we can compute the value of output state h_t:

$$h_t = g_o \odot f(C_t)$$

h_t contains the information not only from local representation z_t, but also from previous output state h_{t-1}, hence can capture dependencies in a sentence. Because the dependencies forward and backward are both important to label semantic roles, we extend LSTM with bidirectional approach, resulting in:

$$a_t = [\overrightarrow{h_t}^T, \overleftarrow{h_t}^T]^T \quad 1 \leq t \leq N$$

Further, a nonlinear transformation follows:

$$v_t = f(W_2 a_t) \quad 1 \leq t \leq N$$

where $W_2 \in \mathbb{R}^{n_3 \times n_2}$, n_2 is the dimension of a_t.

2.4 Output Representation

For each word to be labeled, we adopt linear transformation to get the output vector o_t:

$$o_t = W_3 v_t \quad 1 \leq t \leq N$$

$W_3 \in \mathbb{R}^{n_4 \times n_3}$, where n_4 is the number of semantic role labels in IOBES tagging schema. Therefore, the resulting vector o_t for the t-th word is of length n_4, and each dimension corresponds to the score of a certain semantic role label.

2.5 Training Criteria

Because there are dependencies among word labels in a sentence, isolated training approach which independently considers each word will be inappropriate. Therefore, we adopt sentence tag approach, in which we encourage the correct path of tags, while discouraging all other valid paths.

Given all our training examples:

$$T = (x^{(i)}, y^{(i)})$$

where $x^{(i)}$ denotes the i-th training sentence, $y^{(i)}$ is the corresponding N_i (the number of words to be labeled) dimension vector, which indicates the correct path of tags, and $y_t^{(i)} = k$ means the t-th word has the k-th semantic role label. The score of $x^{(i)}$ along the path $y^{(i)}$ is defined as follows:

$$s(x^{(i)}, y^{(i)}, \theta) = \sum_{t=1}^{N_i} o_{t y_t^{(i)}}$$

where θ is an ensemble of all the parameters in the network.

The log likelihood with a single sample is then:

$$\begin{aligned} logp(y^{(i)}|x^{(i)}, \theta) &= log \frac{\exp\left(s(x^{(i)}, y^{(i)}, \theta)\right)}{\sum_{y\prime} \exp\left(s(x^{(i)}, y\prime, \theta)\right)} \\ &= s(x^{(i)}, y^{(i)}, \theta) - log \sum_{y\prime} \exp\Big(s(x^{(i)}, y\prime, \theta)\Big) \end{aligned}$$

where $y\prime$ ranges from all the valid paths of tags.

The full log likelihood of the whole training corpus is as follows:

$$J(\theta) = \sum_{i=1}^{T} logp(y^{(i)}|x^{(i)}, \theta)$$

To compute the network parameter θ, we maximize the log likelihood $J(\theta)$ using stochastic gradient ascent in the experiments.

2.6 Introducing Heterogeneous Resource

A single annotated corpus with semantic role labels is often limited in its scalability. Heterogeneous resource in our work is defined as another dataset annotated with semantic roles, which also provides predicate-argument structure annotation, but uses very different semantic role labels and annotation schema. However, in spite of these differences, the latent semantic meaning may be highly correlated. Therefore, the introduction of heterogeneous data can alleviate the problem of scalability with a single annotated corpus.

Traditional approaches hardly concern the existence of heterogeneous resource and are difficult to relate different annotation schemas, but in the framework of our model, heterogeneous data can be introduced in a relatively convenient way. Specifically, we learn a bidirectional LSTM RNN model based on heterogeneous data, then with the fine-tuned word embeddings we initialize the model on our experimental dataset. The principle behind is that the words almost convey the same semantic meaning albeit in distinct annotation schemas. The introduction of heterogenous resource in this way is efficient and can lead to performance improvement on our experiment.

3 Experiments

We conduct experiments to compare our model with previous landmark methods on the benchmark dataset CPB for Chinese SRL. The result

Remark	Choice
Word embedding dimension	$n_{word} = 50$
POS tag dimension	$n_{pos} = 20$
Distance dimension	$n_{dis} = 20$
Nonlinear layer	$n_1 = 200$
RNN layer	$n_h = 100$
Nonlinear layer	$n_3 = 100$
Learning rate	$\alpha = 10^{-3}$

Table 1: Hyper parameters of our model.

Method	F1(%)
Xue (2008)	71.90
Collobert and Weston (2008)	74.05
Sun et al. (2009)	74.12
Yang and Zong (2014)	75.31
Ours (Random Initialization)	**77.09**
+ Standard Pre-training	77.21
+ Heterogenous Resource	**77.59**

Table 2: Results comparison on CPB dataset.

reveals that even with our basic model, which does not resort to other resources, our approach can significantly outperform all of the competitors. Moreover, we enrich our work with introducing heterogenous resource to make a further improvement on performance. And the result also shows the influence of heterogeneous resource is more evident than the standard method of pre-training for word embeddings.

3.1 Experimental Setting

To facilitate comparison with previous work, we conduct experiments on the standard benchmark dataset CPB 1.0.[1] We follow the same data setting as previous work (Xue, 2008; Sun et al., 2009), which divided the dataset into three parts: 648 files (from chtb_081.fid to chtb_899.fid) are used as the training set. The development set includes 40 files, from chtb_041.fid to chtb_080.fid. The test set includes 72 files, which are chtb_001.fid to chtb_040.fid, and chtb_900.fid to chtb_931.fid. We use another annotated corpus[2] with distinct semantic role labels and annotation schema, which is designed by ourselves for other projects, as heterogeneous resource. This labeled dataset has 17,308 annotated sentences, and the semantic roles concerned are like "agent" and "patient", resulting in 21 kinds of types, which are all distinct from the semantic roles defined in CPB. We use the development set of CPB for model selection, and the hyper parameter setting of our model is reported in Table 1.

3.2 Chinese SRL Performance

Table 2 summarizes our SRL performance compared to previous landmark results. The work of Collobert and Weston (2008) was conducted on English SRL, we implement their approach on CPB for comparison. As indicated by this table, our approach significantly outperforms previous state-of-the-art methods even with all parameters randomly initialized, that is without introducing other resources. This result can prove the ability of our model to capture useful dependencies for Chinese SRL with minimal feature engineering.

Further, we conduct experiments with the introduction of heterogenous resource. Previous work found that the performance can be improved by pre-training the word embeddings on large unlabeled data and using the obtained embeddings to make initialization. With the result in Table 2, it is true that these pre-trained word embeddings have a good effect on our performance (we use *word2vec*[3] on Chinese Gigaword Corpus for word pre-training). However, as shown in Table 2, compared to standard pre-training, the influence of heterogenous data is more evident. We can explain this difference via the distinction between these two kinds of methods for performance improvement. The information provided by standard pre-training with unlabeled data is more general, while that of heterogenous resource is more relevant to our task, hence is more informative and evident.

4 Related Work

Semantic Role Labeling (SRL) was first defined by Gildea and Jurafsky (2002), who presented a a system based on statistical classifiers trained on hand-annotated corpus FrameNet. Sun and Jurafsky (2004) did the preliminary work on Chinese SRL without any large semantically annotated corpus and produced promising results. After CPB (Xue and Palmer, 2003) was built, Xue and Palmer (2005) and Xue (2008) produced more complete and systematic research on Chinese SRL. Ding and Chang (2009) established a

[1] https://catalog.ldc.upenn.edu/LDC2005T23

[2] This Chinese dataset is available on request.

[3] https://code.google.com/p/word2vec/

word based Chinese SRL system, which is quite different from the previous parsing based ones. Sun et al. (2009) extended the work of Chen et al. (2006), performed Chinese SRL with shallow parsing, which took partial parses as inputs. Yang and Zong (2014) proposed multi-predicate SRL, which showed improvements both on English and Chinese Proposition Bank. Different from most work relying on a large number of handcrafted features, Collobert and Weston (2008) proposed a convolutional neural network for SRL. Their approach achieved competitive performance on English SRL without requiring task specific feature engineering. However, by max-pooling operation, the convolution approach only preserved the most evident features in a sentence, thus can only weakly model the dependencies. With our bidirectional LSTM RNN model, this problem can be well alleviated.

Our model is based on recurrent neural network (RNN), which uses iterative function loops to store contextual information. To remedy the problem of vanishing and exploding gradients when training the standard RNN, Hochreiter and Schmidhuber (1997) proposed long-short-term memory (LSTM), which has been shown capable of storing and accessing information over very long time spans. Bidirectional RNN (Schuster and Paliwal, 1997) and bidirectional LSTM RNN (Graves et al., 2005) are the extensions of RNN and LSTM RNN with the capability of capturing contextual information from both directions in the sequence. In recent years, RNN has shown the state-of-the-art results in many NLP problems such as language modeling (Mikolov et al., 2010) and machine translation (Sutskever et al., 2014; Bahdanau et al., 2014). Sundermeyer et al. (2014) also used bidirectional LSTM RNN model to improve strong baselines when modeling translation. More recently, Zhou and Xu (2015) proposed LSTM RNN approach for English Semantic Role Labeling, which shared similar idea with our model. However, the features used and the network architecture were different from ours. Moreover, it is delightful that our work can achieve a rather good result with a relatively simpler model architecture.

5 Conclusion

In this paper, we formulate Chinese SRL problem with the framework of bidirectional LSTM RNN model. In our approach, the bidirectional and long-range dependencies in a sentence, which are important for Chinese SRL, can be well modeled. And with the framework of deep neural network, the heavy job of feature engineering is much alleviated. Moreover, our model makes the introduction of heterogenous data, which can alleviate the problem of scalability with a single annotated corpus, more convenient. Experiments show that our approach achieves much better results than previous work, and the introduction of heterogenous resource can make further improvement on performance.

Acknowledgments

This research is supported by National Key Basic Research Program of China (No.2014CB340504) and National Natural Science Foundation of China (No.61375074,61273318). The contact authors of this paper are Baobao Chang and Zhifang Sui.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Wenliang Chen, Yujie Zhang, and Hitoshi Isahara. 2006. An empirical study of chinese chunking. In *Proceedings of the COLING/ACL on Main conference poster sessions*, pages 97–104. Association for Computational Linguistics.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM.

Weiwei Ding and Baobao Chang. 2008. Improving chinese semantic role classification with hierarchical feature selection strategy. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 324–333. Association for Computational Linguistics.

Weiwei Ding and Baobao Chang. 2009. Word based chinese semantic role labeling with semantic chunking. *International Journal of Computer Processing Of Languages*, 22(02n03):133–154.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic labeling of semantic roles. *Computational linguistics*, 28(3):245–288.

Alex Graves, Santiago Fernández, and Jürgen Schmidhuber. 2005. Bidirectional lstm networks for improved phoneme classification and recognition. In *Artificial Neural Networks: Formal Models and*

Their Applications–ICANN 2005, pages 799–804. Springer.

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

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *INTERSPEECH 2010, 11th Annual Conference of the International Speech Communication Association, Makuhari, Chiba, Japan, September 26-30, 2010*, pages 1045–1048.

Mike Schuster and Kuldip K Paliwal. 1997. Bidirectional recurrent neural networks. *Signal Processing, IEEE Transactions on*, 45(11):2673–2681.

Honglin Sun and Daniel Jurafsky. 2004. Shallow semantic parsing of chinese. In *Proceedings of NAACL 2004*, pages 249–256.

Weiwei Sun, Zhifang Sui, Meng Wang, and Xin Wang. 2009. Chinese semantic role labeling with shallow parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3-Volume 3*, pages 1475–1483. Association for Computational Linguistics.

Weiwei Sun. 2010. Improving chinese semantic role labeling with rich syntactic features. In *Proceedings of the ACL 2010 Conference Short Papers*, pages 168–172. Association for Computational Linguistics.

Martin Sundermeyer, Tamer Alkhouli, Joern Wuebker, and Hermann Ney. 2014. Translation modeling with bidirectional recurrent neural networks. In *Proceedings of the Conference on Empirical Methods on Natural Language Processing, October*.

Ilya Sutskever, Oriol Vinyals, and Quoc VV Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Nianwen Xue and Martha Palmer. 2003. Annotating the propositions in the penn chinese treebank. In *Proceedings of the second SIGHAN workshop on Chinese language processing-Volume 17*, pages 47–54. Association for Computational Linguistics.

Nianwen Xue and Martha Palmer. 2005. Automatic semantic role labeling for chinese verbs. In *IJCAI*, volume 5, pages 1160–1165. Citeseer.

Nianwen Xue. 2008. Labeling chinese predicates with semantic roles. *Computational linguistics*, 34(2):225–255.

Haitong Yang and Chengqing Zong. 2014. Multi-predicate semantic role labeling. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 363–373.

Jie Zhou and Wei Xu. 2015. End-to-end learning of semantic role labeling using recurrent 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 1127–1137, Beijing, China, July. Association for Computational Linguistics.