International Journal of Computer Trends and Technology
ISSN: 2231 – 2803 / https://doi.org/10.14445/22312803/IJCTT-V69I10P103

Volume 69 Issue 10, 23-27, Oct, 2021
© 2021 Seventh Sense Research Group®

Original Article

Semantic Role Labeling Based on Highway-BiLSTM-CRF Model

Xinxin Li[1], Xiangzhong Pu[2]

[1,2]School of Computer Science and Technology, Shandong University of Technology, Zibo, Shandong, China

Received Date: 05 September 2021
Revised Date: 07 October 2021
Accepted Date: 19 October 2021

Abstract *- Semantic Role Labeling is an important task in natural language processing. At present, the main approach of Semantic role labeling is based on BiLSTM. However, the BiLSTM network may have training difficulties and vanishing gradient problems with increased network depth. This paper proposes a Highway-BiLSTM-CRF model to solve this problem, which connects BiLSTM layers with highway networks. In the input layer, dependency relations, the distance between predicate and arguments are added to improve the experimental effect. Finally, the CRF layer is used to obtain the optimal tagging sequence. Experimental results of Chinese PropBank show that Chinese semantic role labeling achieves the best performance when BiLSTM depth is 8 layers, in which F1 value reaches 80.15%.*

Keywords — *Semantic role labeling, BiLSTM-CRF, Highway network, The dependency relation.*

I. INTRODUCTION

Semantic Role Labeling (SRL) is a fundamental task in natural language processing and has been widely used in information extraction, machine translation, and question answering systems. The study of semantic role labeling and improving its accuracy had played a significant role in the development of natural language processing. Semantic role labeling identifies all arguments of given predicates in a sentence and assigns them different semantic role types, such as agent, patient, time, place, etc. Normally, the core semantic role types are ARG0-ARG5, where ARG0 denotes the agent, ARG1 denotes the patient, and the remaining four types ARG2-ARG5 are assigned different semantic roles depending on the predicate. The other semantic roles are additional ones, denoted by ARG-X, e.g., location as ARG-LOC, time as ARG-TM, etc. An example is given in Figure 1.

sentence	Mike met John in the library yesterday.
roles	ARG0 Predicate ARG1 ARG-LOC ARG-TM

Fig. 1 An example of semantic role labeling

In this example, the word "meets" is the predicate, "Mike" is the agent, "John" is the patient, "yesterday is the time, and "library" is the place.

Most traditional SRL systems are based on syntactic analysis, which usually consists of five steps. Firstly, a syntactic tree needs to be constructed. Secondly, the candidate arguments for a given predicate are identified from the syntactic tree. Since there might be many candidate arguments in a sentence, those candidates that are least likely to be arguments need to be cut. Then, argument identification that determines the real arguments from former pruned arguments is performed as a binary classification problem. Finally, the semantic role labels of the arguments are obtained by a multi-class classifier. Traditional methods usually require hand-crafted extraction of many features and rely heavily on the syntactic analysis results.

With the development of deep learning, automatic feature learning makes it more efficient and accurate than traditional feature extraction methods and solves the problem of error accumulation. In recent years, deep learning has been widely applied for semantic role labeling. He et al. proposed a bidirectional long short-term memory network(BiLSTM), which achieved good results for English semantic role labeling [1]. Wang et al. applied LSTM networks to Chinese semantic role labeling and achieved good performance [2]. However, as the depth of the neural network increases, problems such as gradient disappearance and training difficulties arise.

A model based on Highway-BiLSTM-CRF is proposed to solve this problem. This model extended one-layer BiLSTM to a deep BiLSTM, and a highway neural network is added between different BiLSTM layers to solve gradient disappearance during the training process [3]. To improve the accuracy of our model, features including words, predicates, and dependent relations are added as feature vectors and fed as inputs to the neural network. Finally, the global optimal labeling sequence is output through the CRF layer to obtain the final results.

II. RELATED WORK

Traditionally, semantic role labeling is performed based on syntactic parsing of phrase structure. With the development of dependency parsing, a semantic role labeling system based on dependency trees has been proposed.

Dependency structure is more flatter than phrase structure, making the distance between semantic roles and predicates in the syntactic tree relatively shorter. Moreover, the semantic role labeling methods based on dependency trees exploit the dependency relations between phrases and can focus on those phrases that have dependency relations with the predicates.

Hacioglu presented a semantic role labeling algorithm using dependency trees [4]. Xue proposed a method to label semantic roles for Chinese predicates and achieved good results [5]. Sun proposed a semantic role labeling approach using partial syntactic trees as input [6]. In addition, Yang et al. proposed a multi-predicate semantic role labeling method based on discriminative reordering, which was used to solve the phenomenon of multiple predicates in sentences and significantly improved the effectiveness of shared argument classification [7].

The above methods for semantic role labeling are mainly based on traditional statistical methods, such as support vector machines, conditional random fields, maximum entropy models, etc. These methods require manual extraction of many features, and the features might lead to model overfitting.

Many approaches based on deep neural networks are proposed to solve the problem. Collobert et al.'s work on English using CNN models showed that it reduced many features. Compared to traditional feature-based machine learning methods, it reached the best results of semantic role labeling for English [8,9].

Niu Yilin et al. proposed sememe attention over the target model (SAT), and they took the original information of words into account to improve the performance of semantic role labeling [10]. In 2016, Li et al. used an RNN model for Chinese semantic role labeling, and the experimental results improved substantially [11]. Xia et al. released a Chinese SemBank dataset for Chinese semantic role labeling in 2017 and proposed an evolutionary neural network model [12]. Wang et al. used BiLSTM networks for semantic role labeling and achieved the best results in models with no additional resources introduced [2]. Sha et al. investigated argument identification using dependent information and achieved an F1 value of 77.69% on CPB [13]; Guo et al. proposed unified neural network architecture for identifying and classifying multiple types of semantic relations between words in a sentence [14]. Wang et al. built a semantic role recognition model for Chinese frames by a neural network framework with feature fusion [15].

In recent years, most semantic role labeling systems have been based on LSTM. Inspired by the BiLSTM-CRF model, we proposed the Highway-BiLSTM-CRF model that extended a one-layer BiLSTM network to a multilayer LSTM and introduced a highway network to solve the gradient disappearance problem.

III. OUR MODEL

We propose a deep semantic role labeling model based on Highway-BiLSTM-CRF, and the network architecture is shown in Fig 2. The model combines the feature vectors of the word, part of speech, predicate, whether it is a predicate, the distance from the word to the predicate, and the dependency relationship as the network's input. The hidden layer in the model is composed of multiple layers of BiLSTM, and the highway networks connect adjacent BiLSTM layers. The CRF layer is used to output the best labeling sequence.

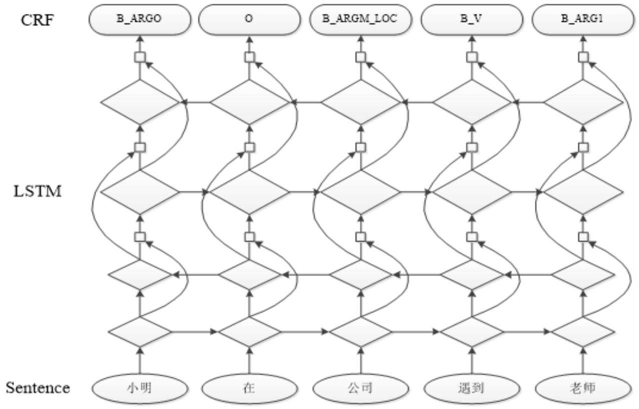

Fig. 2 Architecture of Highway-BiLSTM-CRF model

A. Word Embedding

The model uses five types of feature vectors as the input of the neural network, which includes the current word, the predicate, the distance from the current word to the predicate, the dependency relationship, and part of speech tags. The word embedding of the current word is trained using the ELMO model [16]. The dimension of the predicate vector is 1, which is obtained by determining whether the current word is a predicate or not. If the current word is a predicate, its value is 1; otherwise, it is 0. The distance between the current word and the predicate is the number of words from the current word to the predicate. The window for the current word and part of speech tags is [-2,2], which means that the tokens and part-of-speech tags of the current word and its neighbouring words in the window are used as features. The vector representation of each word is the concatenation of the five types of feature vectors.

B. Dependency Relation

Dependency parsing aims to represent the syntactic structure by describing the relationships among its components, such as subject-predicate relationship and verb-object relationship. The core component in a sentence is the predicate, which governs all other words in the sentence. The dependency tree shows the relationship between words and the importance of verb predicate in the sentence, which is related to semantic role labeling. Therefore, this paper introduces dependency features in the semantic role annotation process.

Stanford Dependency Parser obtains the dependency trees of the training sentences. The syntactic tree is used to determine whether there are dependencies between different words. Then, we build a dependency co-occurrence matrix U where each element stores the number of dependency relations between word i and word j. The matrix is reduced in dimension by singular value decomposition and used to generate the dependency embedding for each word. The formula of singular value decomposition is shown as follows.

$$M = U\Sigma V^{*} \qquad (1)$$

U is an m*m order unitary matrix, ∑ is a positive semi-definite m*n order diagonal matrix, and *V**, the conjugate transpose of *V*, is a n*n order unitary matrix.

C. BiLSTM Network

A recurrent Neural Network (RNN) can memorize historical information and apply it to the current state. But it is difficult to use RNN to train a long-distance sentence because of the problems of gradient disappearance and gradient explosion.

LSTM, a variant of RNN, can effectively capture long-distance dependencies. The LSTM model is composed of input word X_t at time *t*, cell state C_t, hidden layer state h_t, forget gate f_t, memory gate i_t, and output gate O_t. LSTM is calculated by remembering new information in the cell state and discarding useless information. The formulas are shown as follows.

$$\overset{:}{C}_t = \tanh(W_c x_t + U_i h_{t-1} + b_c) \qquad (1)$$

$$i_t = \sigma(W_t x_t + U_i h_{t-1} + b_i) \qquad (2)$$

$$f_t = \sigma(W_f x_t + U_f h_{t-1} + b_f) \qquad (3)$$

$$C_t = f_t C_{t-1} + i_t \overset{:}{C}_t \qquad (4)$$

$$O_t = \sigma(W_o x_t + U_o h_{t-1} + b_o) \qquad (5)$$

$$h_t = O_t \tanh(C_t) \qquad (6)$$

In sequence labeling, both past and new information can contribute to sequence labeling. BiLSTM combines the forward hidden layer h_z and backward hidden layer h_f to r_t.

$$r_t = [h_z, h_f] \qquad (7)$$

D. Highway Network

The depth of the neural network is important for its performance. However, as the depth increases, training the network becomes more difficult, and occur gradient disappearance. We introduce a highway network to solve this problem [3].

We use transition gate r_t to control the weights between different layers. The output h_t is:

$$r_t = \sigma(W_t h_{t-1} + W_r x_t + b_r) \qquad (8)$$

$$h_t^{'} = O_t \tanh(C_t) \qquad (9)$$

$$h_t = r_t h_t^{'} + (1 - r_t) W_h x_t \qquad (10)$$

E. CRF Layer

After BiLSTM networks, the probabilities of all tags for each word can be directly calculated through a softmax layer, and the tag with the highest probability can be selected as the output. However, this method only considers the context of the current word but ignores the other words in the sentence. Therefore, this paper introduces the CRF layer, takes the sentence, and outputs the optimal global results. In the linear-chain conditional random field, given observation sequence X, the probability of an output sequence Y can be defined as:

$$P(Y \mid X) = \frac{1}{Z(X)} \exp\Big(\sum_{i=1}^{n}\big(\sum_{j}\lambda_j t_j(y_{i-1}, y_i, X, i) + \sum_{k} u_k s_k(y_i, X, i)\big)\Big) \qquad (11)$$

The cost function is

$$L(\lambda, D) = -\log\Big(\prod_{m=1}^{N} p(Y_m \mid X_m, W)\Big) + c\frac{1}{2}\|W\|^2 \qquad (12)$$

Where *Z(x)* is the normalization factor, t_j and s_k are feature functions.

IV. EXPERIMENTS

A. The Dataset and Model Setting

In the experiment, we use the Chinese PropBank (CPB) as our semantic role labeling dataset. To evaluate our model, we use Precision, Recall, and F1 as criteria.

The parameters in our neural network model are important. We set the following parameters:

(1) The vector dimension: set the dimension of the word embedding to 150, set the dimension to 1 for whether it is a predicate, set the distance between the current word to the predicate to 1, and set the dimensions of part of speech and dependency relation to 50.

(2) Other parameter settings: the dimension of hidden layer: 240; learning rate: 0.01; Droupout rate in BiLSTM layers: 0.5; regularization factor: 0.0002.

B. Experimental Results and Analysis

To investigate the best depth of BiLSTM layers in our model, we perform experiments with different depths. The experimental results are shown in Figure 3. Figure 3 analyzes the effect of the depth of BiLSTM layers and highway network for the model. When it uses two BiLSTM layers, the model without adding highway networks performs better than the model with highway networks. But as the depth of the BiLSTM layers increases, the model's performance without highway networks decreases sharply, but the performance of the model with highway networks improves. Meanwhile, when the depth of BiLSTM layers is 9, the experimental results of both models decrease. Therefore, the performance of the model achieves its best performance with 8 BiLSTM layers.

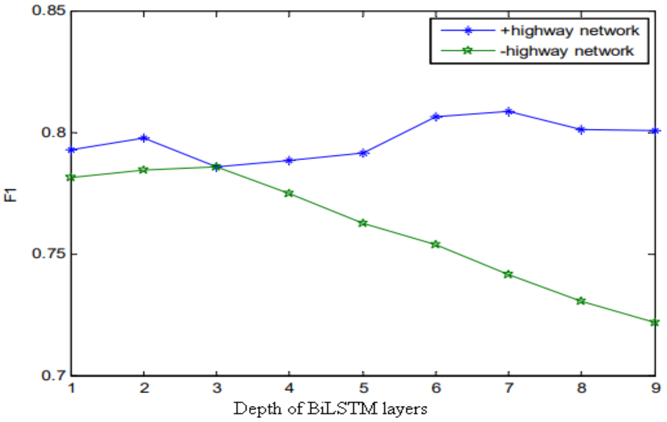

Fig. 3 Results on the different depths of BiLSTM layers

Table 1 shows the experimental results of the model using 8 BiLSTM layers with/without highway network. When the highway network is not added, the F1 value is 74.90%. When the highway network is added, the F1 value of the model reaches 79.95%, which is an increase of about 5.05%. It can be seen that when the number of BiLSTM layers increases, highway networks are beneficial to improve the performance of the model.

In order to verify the benefit of dependency relations on the model, this paper conducts experiments using the golden dependency relations and the automatic dependency relations. Penn2Malt was used to transform the syntactic tree of the phrase structure of CTB to obtain the golden dependencies. Automatic dependency relations are extracted using Stanford Dependency Parser [17]. The experimental results are shown in Table 1. The F1 value of the model was 79.40% when the dependency relations were not added, and it reached 80.65% with golden correct dependency relations, which was a 1.25% increase. The F1 value on our model with automatic dependency relations reached 80.15%, an increase of 0.75% than the model with dependency relations.

Table 1. Results with the highway network and dependency relations

Model	Precision(%)	Recall (%)	F1(%)
- highway network	77.74	73.57	74.90
+ highway network	81.95	76.86	79.95
+ dep relation	82.65	76.76	80.15
- windows	82.63	78.72	80.12
+ golden dep relation	81.72	78.85	80.65

The experiments are also performed on the model with context windows. The results in Table 1 show that the F1 value of the model is only 80.12% when the context window is not added. When the current words and part of speech features of the context window of [-2,2] are added, it reached 80.15%, an increase of 0.03%. Therefore, adding the context window of the current word and part of speech features to the model is necessary.

There are many ways to obtain word embedding. We perform experiments on three different word embedding models: the CBOW model, Glove model, and ELMO model [18,19] to get the best word embedding. The experimental results are shown in Fig 4. From Fig 4, we can see that the model with randomly initialized word embedding is the worst. The CBOW, Glove, and ELMO models perform better, indicating that the three models can capture linear relationships between words. Among them, the ELMO model achieves the best results. The ELMO model is a dynamic word vector training method and changes with the different context information.

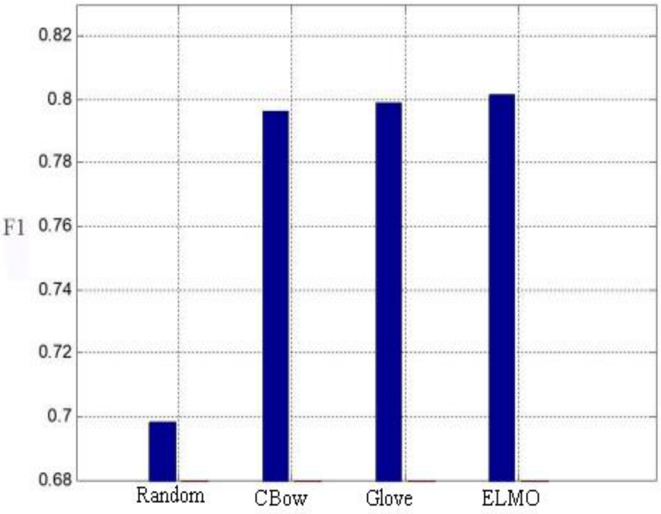

Fig. 4 Results with different word embeddings

We also analyze the effect of different iteration numbers on the performance of the model. The experimental results are shown in Fig 4. When the iteration number increases from 0 to 180, the F1 value of the model increases. When the iteration number reaches 180, the results tend to be stable and achieve the best performance when the model takes 320 iterations.

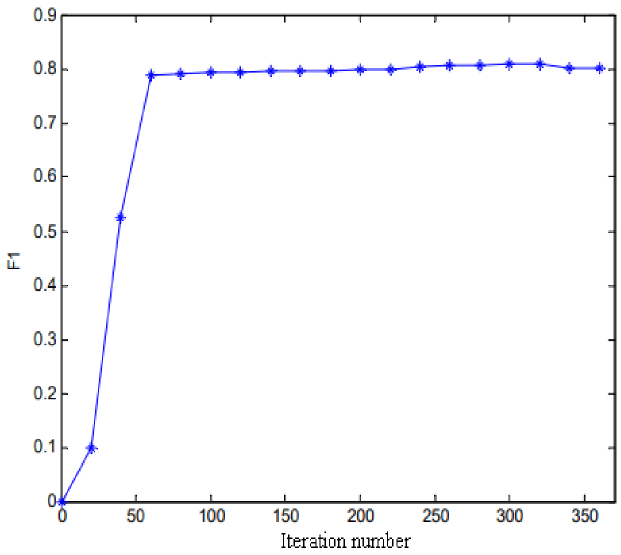

Fig. 5 Results with different iteration numbers

Finally, we compare our model with previous works, and the results are shown in Table 2. Wang et al. used a single-layer LSTM for this task [2]. Sha et al. used a maximum entropy classifier for Chinese semantic role labeling[13]. Zhang introduced a gated mechanism and BiLSTM-CRF [20]. As shown in the table, our model with an 8-layer BiLSTM-CRF network and adding dependencies improve 0.75% F1 value than previous work.

Table 2. Comparison with previous work

Models	**Precison（%）**	**Recall（%）**	**F1(%)**
Wang et al.[2]	-	-	77.09
Sha et al.[13]	-	-	77.69
Zhang et al.[20]	82.44	76.57	79.40
Our model	82.65	76.76	80.15

V. CONCLUSION

To address the training difficulties and gradient disappearance problem of deep neural networks, we propose a highway-BiLSTM-CRF model for semantic role labeling. Dependency relations are also introduced to improve model performance. The best labeling sequence is obtained through the CRF layer. The results show the effectiveness of the model proposed in this paper. Both highway network and dependency features improve the performance.

REFERENCES

[1] Luheng He, Kenton Lee, Mike Lewis, Luke Zettlemoyer, Deep Semantic Role Labeling: What Works and What's Next. In Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (2017) 473-483.

[2] Wang Zhen, Jiang Tingsong, Chang Baobao, et al. Chinese semantic role labeling with the bidirectional recurrent neural network. Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. (2015) 1626-1631.

[3] Yu Zhang, Guoguo Chen, Dong Yu, Kaisheng Yaco, Sanjeev Khudanpur, and James Glass. High-way long short-term memory rnns for distant speech recognition. In 2016 IEEE International Conference on Acoustics, Speech and Signal Processing(ICASSP),(2016)5755–5759.

[4] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 9(8)(1997) 1735-1780.

[5] Xue Nianwen. Labeling Chinese predicates with semantic roles. Computational Linguistics, 34(2)(2008) 225-255.

[6] Sun Weiwei, Sui Zhifang, Wang Meng, et al. Chinese semantic role labeling with shallow parsing. Proceedings of the EMNLP. (2009) 1475-1483.

[7] Yang Haitong, Zong Chengqing. Multi-predicate semantic role labeling, Proceedings of the 2014 Conference on empirical Methods in Natural Language Processing. (2014) 363-373.

[8] Collobert R, Weston J. A unified architecture for natural language processing: Deep neural networks with multitask learning, Proceedings of the 25th ICML. (2008)160-167.

[9] Collobert R, Weston J, Bottou L, et al. Natural language-processing (almost) from scratch. The Journal of Machine Learning Research. 12(1) (2011) 2493-2537.

[10] Niu Yilin, Xie Ruobong, Liu Zhiyuan, et al. Improved Word Representation Learnong with Sememes. Proceedings of ACL. (2017) 2049-2058.

[11] Li Tianshi, Li Qi, Chang Bbaobao. Improving Chinese Semantic Role Labeling with English Proposition Bank. China National Conference on Chinese Computational Linguistics. (2016) 3-11.

[12] Xia Qiaolin, Chang Baobao, Sui Zhifang. A Progressive Learning Approach to Chinese SRL Using Heterogeneous Data. Proceedings off the 55th ACL. (2017) 2069-2077.

[13] Sha Lei, Jiang Tingsong, Li Sujian, et al. Capturing argument relationships for Chinese semantic role labeling. Proceedings of the Conference on Empirical Methods in Natural Language Processing. (2016) 2011-2016.

[14] Guo Jiang，Che Wanxiang，Wang Haifeng， et al. . A unified architecture for semantic role labeling and relation classification[C]// Proceedings of the 26th International Conference on Computational Linguistics. (2016) 1264-1274.

[15] Wang Y, Johnson M, Wan S, et al. How to best use syntax in semantic role labeling.Proceedings of the 57th Conference of the Association for Computational Linguistics. (2019)5338-5343.

[16] Peters, Matthew E., Neumann, Mark, Iyyer, Mohit and Gardner, Matt and Clark, Christopher and Lee, Kenton and Zettlemoyer, Luke. Deep contextualized word representations. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics. (2018) 2227-2237.

[17] Danqi Chen and Christopher D Manning. A Fast and Accurate Dependency Parser using Neural Networks. Proceedings of EMNLP (2014) 740-750.

[18] Mikolov, Tomas and Sutskever, Ilya and Chen, Kai and Corrado, Greg and Dean, Jeffrey. Distributed Representations of Words and Phrases and Their Compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems. (2013) 3111–3119.

[19] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. GloVe: Global Vectors for Word Representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. (2014) 1532–1543.

[20] Zhang Miaomiao, Zhang Yujie, Liu Mingtong, Xu Jin-an, Chen Yufeng. Chinese Semantic Role Labeling Based on Gated Mechanism and Bi-LSTM-CRF. Computer and Modernization. 4(1) (2018) 254-263.