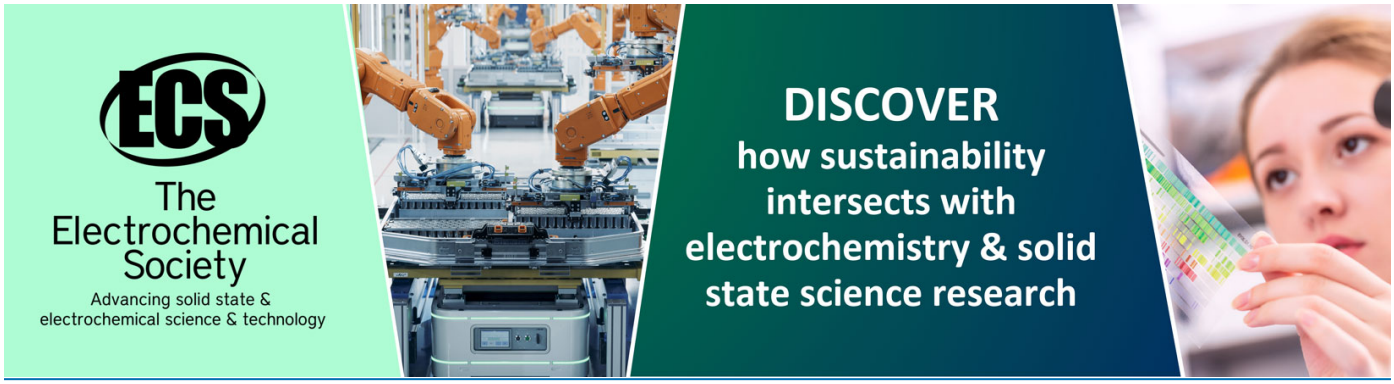
Semantic Dependency Graph Parsing of Financial Domain Questions Based on Deep Learning



Semantic Dependency Graph Parsing of Financial Domain Questions Based on Deep Learning

Peng Xin[1,2], Li Qiujun[1,2]

[1]School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

[2]Key Laboratory of Optical Communication and Networks, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

E-mail : 15730031613@163.com

Abstract. In order to effectively solve the problems of limited accuracy of semantic parsing and ambiguous semantic recognition in the current Question-Answer System based on Chinese Knowledge Map in intention recognition. This paper presents a method about Dependency Reduction in language dependency analysis(DR-BLSTM-CRF) based on Semantic Dependency Graph Parsing(SDGP), Bidirectional Long Short-term Memory(BLSTM) model and Conditional Random Field(CRF) of Xunfei Open Platform. The semantic dependency graph dependency reduction method is as follows:1) Semantic Dependency Graph Parsing of interrogative sentences based on Web API of Xunfei Open Platform to obtain a sentence representation containing semantic dependency information. 2) Named Entity Recognition (NER) algorithm combined with BLSTM and CRF is used to identify the named entity of the interrogative sentence to gain a sequence containing character label information, and then combine the analysis results of semantic dependency graph to obtain a more accurate semantic dependency graph through dependency reduction. The experimental results show that the precision, recall and F1 value of the model proposed in this paper are 33.4%, 33.9% and 34.2% higher than those of LTP in terms of the semantic dependency analysis on the 140,000 self-built data sets of financial domain questions. The model can effectively analyze the semantic dependency of financial domain questions.

1. Introduction

In recent years, financial technology has been in full swing, and various concepts have emerged in an endless stream, most of the common questions in financial client business consulting are repetitive and within a limited area. The traditional artificial customer service requires a large amount of labor costs, while the intelligent question answering system can explore the user's requirements through the user's questions, and it can provide accurate answers for the users, and then leads to service transformation. How to accurately identify the intention of the question is the primary problem of the intelligent question answering system. The traditional short text intent recognition methods mainly include the following two types: one is to reveal the syntactic structure by analyzing the dependencies between the components in the language unit; the other is to analyze the semantic association among the various language units of the sentence, and to present the semantic association in a dependent structure. Due to the serious lack of morphological changes in the former, there is no strict corresponding relationship between lexical classes and syntactic components, which leads to the low accuracy of Chinese syntactic analysis. There are two problems when the latter is used in specific domain questions: One is that the training and effects

of Semantic Dependency Parsing (SDP) depend on the corpus and cannot be widely applied to specific fields, and another one is that the dependence of SDP is too complex to extract semantic information well for some short questions.

In order to solve the above problems and enable the question and answer system to better identify the intention of the question. This paper proposes a method of Dependency Reduction (DR) for the analysis results of the Semantic Dependency Graph of the Xunfei Open Platform. That is, the semantic association between each language unit of the sentence is analyzed by WebAPI of Semantic Dependency Graph in Xunfei Open Platform, and the semantic association is presented in a dependency structure, and then the dependency reduction model is used to reduce the semantic dependence to obtain the ideal semantic dependency analysis results.

For the time being, semantic dependence analysis can be divided into semantic dependence tree parsing (SDTP) and semantic dependence graph parsing (SDGP) [1]. Compared with the semantic dependence tree, the semantic dependence graph has a more comprehensive and in-depth analysis of the common phenomena in Chinese, such as linkage, concurrent speech and concept transposition. And this paper combines the Bidirectional Long Short-Term Memory (BLSTM) [2] and the Conditional Random Field (CRF) [3] to propose a semantic dependence analysis reduction model (DR-BLSTM-CRF), which makes SDP widely applicable to specific areas.

2. DR-BLSTM-CRF model

2.1. BLSTM network

Long Short-Term Memory (LSTM) is a special Recurrent Neural Networks (RNN) [4]. The essential difference lies in the introduction of clever controllable self-circulation by LSTM, which generates a path that allows gradients to flow continuously for a long time. LSTM is more competitive than RNN in dealing with tasks with very long time intervals and delays. The main reason is that LSTM adds a Cell State to replace the traditional hidden neuron nodes. It avoids the problem of gradient disappearance or gradient explosion with the increase of network layers in traditional RNN. The LSTM memory unit is shown in Figure 1.

The realization formula of LSTM memory unit is as follows:

$$\begin{aligned} f_t &= \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \\ i_t &= \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \\ \tilde{C}_t &= \tanh(W_i \cdot [h_{t-1}, x_t] + b_i) \\ C_t &= f_t * C_{t-1} + i_t * \tilde{C}_t \\ o_t &= \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \\ h_t &= o_t * \tanh(C_t) \end{aligned} \quad (1)$$

Where σ is the logical Sigmoid function, i, f, o, * and C represent input gate, forget gate, output gate, convolution-multiplication and cell vectors respectively. The dimensions of these vectors are consistent with the dimension of the hidden layer vector h. W_i, W_f and W_o represent the weight matrices of connecting input gates, forgetting gates and output gates respectively.

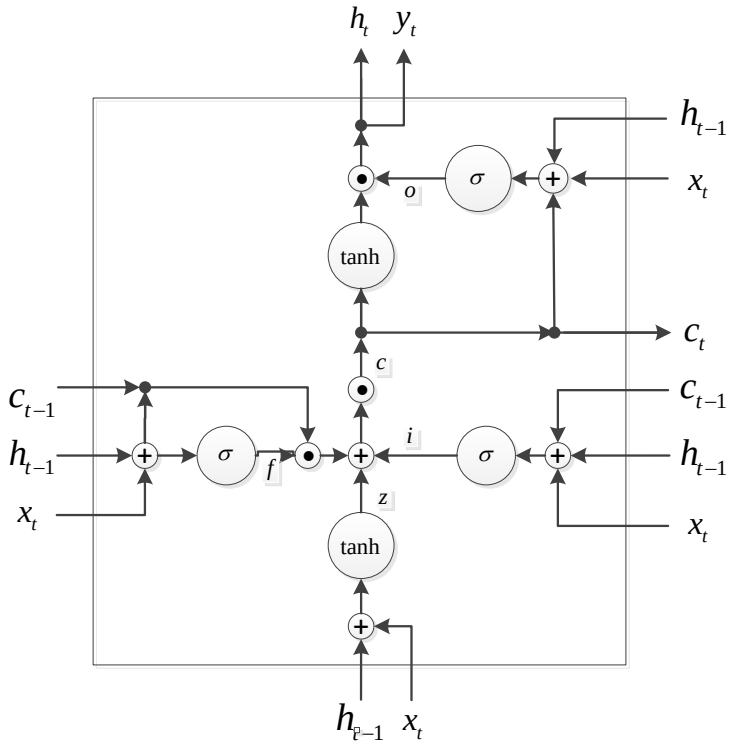


Figure 1. LSTM memory unit

BLSTM consists of two LSTM neural networks with opposite directions, namely forward LSTM and backward LSTM. Its operational principle is that the output of forward LSTM and backward LSTM cascade to form a new feature representation $h_t = [F_i : B_i]$,which indicates rich context information and so on.

2.2. BLSTM-CRF

This BLSTM-CRF model [5] combines BLSTM network with CRF layer, that is, a CRF linear layer is added behind the hidden layer of BLSTM network. The BLSTM-CRF structure is shown in Figure 2.

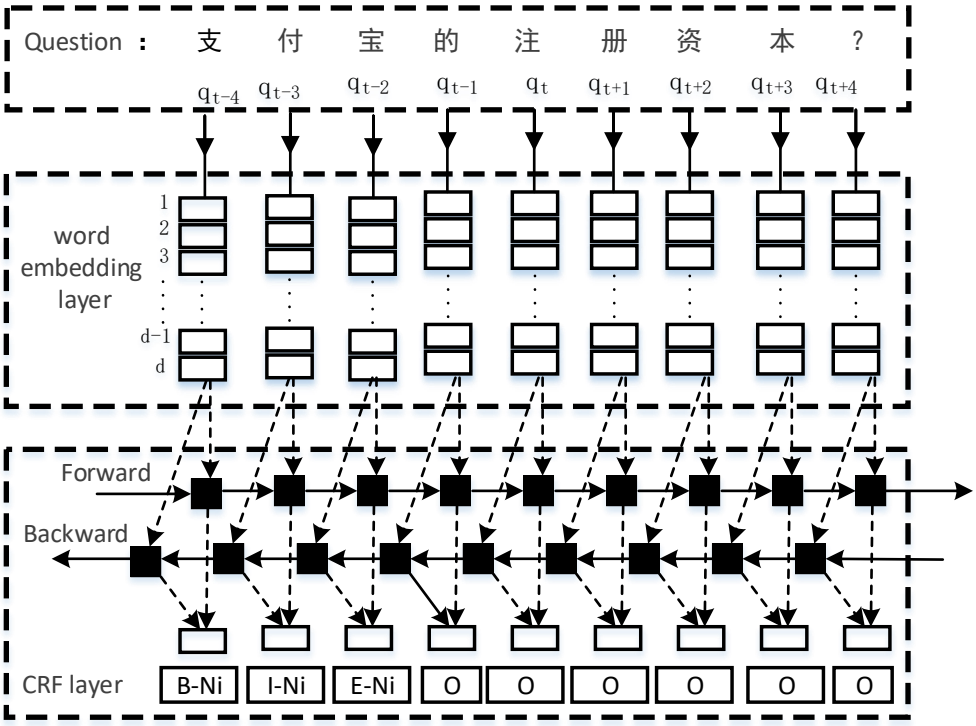


Figure 2. BLSTM-CRF neural network structure

This model not only can obtain the information of the sequence text context, but also can use the annotation information at the level of the whole sentence through CRF to ensure the accuracy of the annotation. This step is one of the key steps in the semantic dependency graph analysis of financial questions based on DR-BLSTM-CRF. The specific steps are as follows:

After modeling the financial domain questions through BLSTM, a sequence of Chinese characters containing context information can be obtained:

$$Char = (char_1, \ldots, char_n) \quad (2)$$

Where $Char$ is a word vector with d dimension.

The question word sequence in financial field is input into bidirectional LSTM neural network model, and the word sequence is constructed by BLSTM. The forward LSTM of bidirectional LSTM is used to model the word sequence and generate a vector representation $CharF_i$ containing the word sequence and the information above the word sequence. Similarly, the backward LSTM reads the word sequence reversely, and expresses the word sequence with its following information as $CharB_i$. Finally, the forward LSTM output of BLSTM and the backward LSTM output are cascaded to form a new feature representation $h_i = [CharF_i : CharB_i]$, h_i directly acts as a feature to make independent mark decision for each output y_i , this method effectively represents the word and its context information with vectors.

In the semantic dependency analysis task of the financial domain question of this paper, there is a strong dependency in the front and back character labels when identifying an entity of a text sequence. For example, for the text sequence " Registration capital of AICRobo Technology(Shenzhen) Co., Ltd.? ", BLSTM tends to obtain the information about the adjacent character labels[6][7][8]. When it receives the characteristic information containing "Robot Technology (Shenzhen) Co., Ltd", it will mistake the sequence as an organization name. The CRF is better at capturing the information of the whole sentence level. The sequence begins with the interference word. Compared with CRF, BLSM has a weaker dependence on the label of the head and tail characters. Therefore, CRF is used to model the dependence of each character on the label in the whole sentence. It is assumed that the output target sequence (i.e. the character label sequence containing the information at the beginning and the end of the question) of

the questions in the financial field is as follows:

$$y = (y_1, \ldots, y_n) \quad (3)$$

In order to obtain the target sequence of financial domain questions effectively, the model's score formula is as follows:

$$s(X, y) = \sum_{j=0}^{n} A_{y_j, y_{j+1}} + \sum_{i=1}^{n} P_{y_i, i} \quad (4)$$

Where P represents the output score matrix of the bidirectional LSTM, its size is $n \times k$, k represents the number of target labels, and n represents the length of the word sequence. A represents the transfer score matrix. When $j = 0$ denotes the beginning of a sequence and $j = n$ denotes the end of a sequence, the size of the square matrix is $K + 2$.

On the sequence of character tags for all question information, the probability that the CRF generates the target sequence y is:

$$p(y \mid X) = \frac{e^{s(X, y)}}{\sum_{y \in Y_X} e^{s(X, y)}} \quad (5)$$

Where Y_X represents all possible character tag sequences corresponding to the sequence of question information X.

In the training process, in order to obtain the correct character tag sequence of question information, the conditional likelihood logarithmic probability of maximizing the correct tag sequence will be adopted.

$$\begin{aligned} \log(p(y^* \mid X)) &= s(X, y) - \log\left(\sum_{y \in Y_X} e^{s(X, y)}\right) \\ &= s(X, y) - \operatorname*{logadd}_{y \in Y_X} s(X, y) \end{aligned} \quad (6)$$

It can be seen from the above expression that the training neural network is to output valid character tag sequence as much as possible in the financial domain question information. Therefore, the maximum score formula given by equation above is used to predict the most appropriate character tag sequence.

$$y^* = \operatorname*{argmax}_{y \in Y_X} s(X, y). \quad (7)$$

Since the interaction between the outputs is modeled, so dynamic programming is adopted to calculate the sum in equation (6) and the maximum posteriori sequence y* in equation (7).

2.3. DR-BLSTM-CRF

In order to solve the problem of semantic dependency analysis of financial entities including company names and investors in financial domain questions, this section proposes a semantic dependency analysis reduction method, which combines the Semantic Dependency Graph (SDG) analysis of Xunfei Open Platform, Bidirectional Long Short-Term Memory (BLSTM) network model and conditional random field(CRF), that is, DR-BLSTM-CRF model.

The process of semantic dependency parsing of financial domain questions based on DR-BLSTM-CRF is shown in Figure 3.

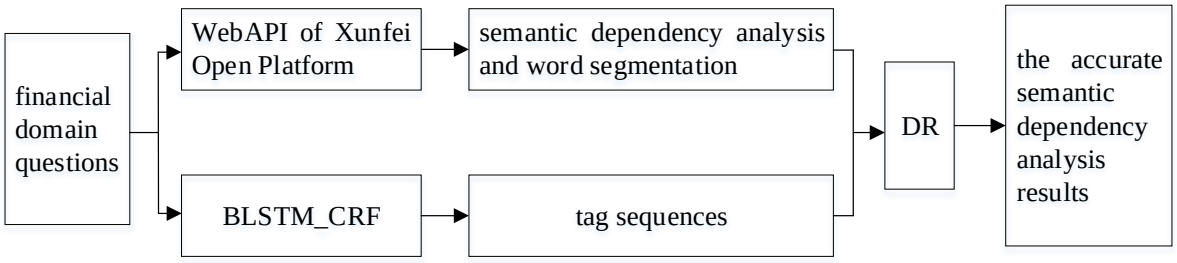


Figure 3. The process of semantic dependency parsing

Firstly, the collected financial domain questions are input into WebAPI of Xunfei Open Platform for semantic dependency analysis and word segmentation, and BLSTM-CRF for financial domain entity recognition respectively. Based on semantic dependence analysis and word segmentation results and tag sequences, the accurate semantic dependency analysis results would be obtained by DR.

The procedure of DR-BLSTM-CRF algorithm and the DR algorithm are shown as follows:

Algorithm 1 DR-BLSTM-CRF Algorithm

Input: Questions in the financial field

1. Input the question into the WebAPI of Xunfei Open platform for word segmentation and semantic dependence analysis, and obtain the word segmentation result cws_result and semantic dependency graph result sdgp_result;
2. Input the question into trained BLSTM-CRF model, predicting the entity sequence result blstm_crf_result containing tag information;
3. Get the blstm_crf_result and set the corresponding entity tag principle, and output the blstm_crf_result as the form of the blstm_crf_reschange, as follows:
   If blstm_crf_result = {"H":"B-Ni", "I": "I-Ni","J":"E-Ni","K":"O"}, then blstm_crf_reschange = ["HIJ", "K"];
4. For the results cws_result, sdgp_result, and blstm_crf_reschange obtained from above steps, the DR algorithm is used to reduce the dependency of the sdgp_result;

Output: The results of reduced and optimized semantic dependency graph parsing: sdgp_result_dr, blstm_crf_reschange.

Algorithm 2 DR Algorithm

Input:cws_result, sdgp_result, blstm_crf_reschange

1. Create an index mapping list master_list from cws_result and blstm_crf_reschange as follows: If cws_result = [‘A’,’B’,’C’,’D’,’E’,’F’,’? ’], Blstm_crf_reschane = [‘ABC’,’D’,’EF’,’? ’], the following results master_list = [[0,1,2],3,[4,5],6];
2. The sdgp_result is grouped according to the index mapping relationship in the master_list obtained in step 1, to obtain sub_sdgp_result;
3. Traverse the master_list, if it contains the sublist sub_list, continue to traverse the sublist, and traverse sub_sdgp_result according to the sublist element, when the "parent" value in sub_sdgp_result is included in the sub_list, delete the dictionary at this time;
4. Update the values of "parent" and "id" in sub_sdgp_result according to master_list to obtain sdgp_result_dr.

   An example is as follows:

if sdgp_result = [{"id":0,"parent":2,"relate": "Nmod"}, {"id":1,"parent":2,"relate":"Desc"}, {"id":2,"parent":3,"relate":"Desc"},{"id":3,"parent":1,"relate":"Root"},{"id":4,"parent":5,"relate":"Exp"},{"id":5,"parent":3,"relate":"mAux"},{"id":6,"parent":4,"relate":"mPunc"}],

we can get the following result: sdgp_result_dr=[{"id":0,"parent":3,"relate":"Desc"}, {"id":1,"parent":2,"relate":"mAux"},{"id":2,"parent":1,"relate":"Root"},{"id":3,"parent":2,"relate":"mPunc"}].

Output: Optimized semantic dependency graph parsing results.

Note: According to the practical application requirements, it is found that when dependency reduction is carried out, only one entity ’C’ in the corresponding reduced entity (e.g.[‘A’,’B’,’C’]) is made up of other elements of the sentence as the parent node, and the semantic relation type of the reduced entity ’ABC’ is consistent with that of ’C’.

3. Experimental description

In order to verify the semantic dependency analysis effect of the method proposed in this paper on questions in the financial field, the web crawler crawls some information in the financial field on the Internet, based on the crawled information and the practical applications of question and answer in financial field, about 140,000 questions were constructed for experiment.

3.1. Dataset

The dataset in this paper has crawled about 50,000 financial data from the platforms of Eastern Fortune Network, Finance, and causal trees, and based on the information collected, combined with the practical application of question and answer in financial field, finally constructed about 140,000 questions. From the observation of the data, it is found that the proportion of financial entities is about 65%. For the above data, training set and test set are randomly selected to according to the ratio of 8:2.

3.2. Experimental setting

For the above model, this paper uses TensorFlow to build. TensorFlow is a deep learning framework developed by Google's artificial intelligence team, and is widely used in the programming implementation of various machine learning algorithms. In the experiment, the batch_size is set to 32 of the training set, the batch_size of the test set is set to 64, and the dimension of the hidden state in the front and back direction of the BLSTM are 100, and in order to prevent over-fitting during model training, the Dropout training method is adopted proposed by Hinton et al.[9], the value is set to Dropout=0.5; Learning Rate (LR) also plays an important role in model training, and learning rate determines the speed at which parameters move to the optimal value. Therefore, after a large number of experimental comparisons, it is found that the learning rate LR = 0.01 is the best.

Early stopping strategy is also introduced in this experiment. That is to say, when introducing Early Stopping strategy, the accuracy of verification set, recall rate and F1-score are calculated after each Epoch end. When F1-score is no longer improved, the training of the model will stop, but only when

F1-score is not significantly improved many times. Otherwise, the training effect of the model will easily be less than optimal. In this experiment, the F1-score did not improve significantly after 20 consecutive Epochs, so the training stopped updating iterations.

4. Simulation Results and Analysis

On the dataset, this paper adopts the following methods for comparison: BLSTM, CNN+CRF, BLSTM+CRF, DR+BLSTM, DR+CNN+CRF, DR+BLSTM+CRF. The evaluation indicators used in this paper are Precision, Recall, and F1-score. The experimental results are shown in the following tables.

Table 1. Various financial entities identification result

scheme	Precision	Recall	F1-score
BSLTM	0.891	0.859	0.875
CNN+CRF	0.913	0.897	0.905
BLSTM+CRF	**0.934**	**0.949**	**0.941**

Table 2. Financial domain question SDTP results

scheme	Precision	Recall	F1-score
Xunfei Platform	0.578	0.571	0.574
DR+BLSTM	0.809	0.773	0.791
DR+CNN+CRF	0.883	0.865	0.874
DR+BLSTM+CRF	**0.923**	**0.911**	**0.917**

Table 3. Financial domain question SDGP results

scheme	Precision	Recall	F1-score
Xunfei Platform	0.583	0.574	0.578
DR+BLSTM	0.814	0.786	0.800
DR+CNN+CRF	0.886	0.871	0.878
DR+BLSTM+CRF	**0.927**	**0.913**	**0.920**

By comparing the experimental results, it can be found that RNN-based schemes such as BLSTM are better than CNN-based methods, because RNN-based models are more prominent for dynamic sequence problems, while CNN [10] is characterized by efficient extraction of static features.

It can be seen from Table 1 that the introduction of the language model CRF can bring about a certain degree of improvement to the model effect, because the naming of entities in the financial field, such as the name of the institution, may include the someone's name, location, etc., that may easily interfere with entity recognition[11][12], and the introduction of the language model can make the model pay more attention to the overall relationship of the question, and to some extent reduce the influence of the local information on the entity recognition of the model.

The results in Table 2 show that due to the diversification of entity naming in the financial field, this brings a lot of interference factors to entity recognition. Such as the name of the organization may include the person name, location, etc., therefore, the effect of semantic dependency analysis in the financial field directly by the Xunfei Open Platform is not good. In this paper, the DR-BLSTM-CRF model is proposed to solve the problem that semantic dependency parsing cannot be widely applied in specific fields. It can be seen from the experimental results that the proposed model can optimize the domain on the semantic dependency parsing of Xunfei Open Platform, and the performance is improved by about 34%.

By comparing the experimental results in Table 2, we can find that the DR+BLSTM+CRF model proposed in this paper has better performance than other schemes, because the proportion of questions in the dataset includes about 65% of the financial domain entities, and this part of the question is due to the diversification of the entity name in the financial field when the entity is identified, the internal

interference factor of the entity increases, so the semantic dependency analysis of the financial field question should be from the global consideration to reduce the influence of local relationships.

Finally, comparing Tables 2 and 3, it can be seen that the semantic dependency graph parsing(SDGP) is slightly better than the semantic dependency tree parsing(SDTP). Through statistical analysis of the results, we found that this is because about 0.5% of the questions in the dataset are easily confused with each other in semantic dependency parsing, resulting in incomplete semantic description of dependency tree structure.

The main difference between the SDTP and the SDGP is that in the tree structure, any one component cannot depend on two or more components, while in the graph structure, the components in the sentence are allowed to depend on two or two. Moreover, intersections between dependency arcs are allowed in the dependency graph, but not in the dependency tree. Therefore, this paper chooses the semantic dependency graph to analyze the questions in the financial field.

5. Conclusion

In this paper, we propose DR-BLSTM-CRF model to semantic dependency graph parsing for questions in financial field. A preliminary semantic dependency parsing of questions in the financial field is carried out by using WebAPI of Semantic dependency graph analysis in Xunfei Open Platform. Then, the BLSTM-CRF model is used to identify the entities in the financial domain. Finally, the semantic dependency graph parsing of questions in the financial field is completed by optimizing the results of semantic dependency parsing with DR-BLSTM-CRF model. By comparing and analyzing the results of different schemes, it is concluded that the proposed scheme based on DR-BLSTM-CRF is superior to other schemes.

References

[1] Wanxiang Che, Zhenghua Li, Ting Liu. LTP: A Chinese Language Technology Platform. In Proceedings of the Coling 2010: Demonstrations. 2010.08, pp13-16, Beijing, China.
[2] Zhou P, Shi W, Tian J, et al. Attention-Based Bidirectional Long Short-Term Memory Networks for Relation Classification[C]// Meeting of the Association for Computational Linguistics. Berlin,2016:207-212.
[3] Sammut C. Webb G.I. Conditional Random Field [J]. Encyclopedia of Machine Learning and Data Mining.2017,10(3):9-16.
[4] Karjala T W, Himmelblau D M, Miikkulainen R. Data rectification using recurrent (Elman) neural networks[C]// International Joint Conference on Neural Networks. 1992.
[5] Lample G , Ballesteros M , Subramanian S , et al. Neural Architectures for Named Entity Recognition[J]. 2016.
[6] Yang Z , Salakhutdinov R , Cohen W W . Transfer Learning for Sequence Tagging with Hierarchical Recurrent Networks[J]. 2017.
[7] Yang Z, Salakhutdinov R, Cohen W. Multi-Task Cross-Lingual Sequence Tagging from Scratch[J]. 2016.
[8] Ma X, Hovy E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF[J]. 2016.
[9] Hinton G E, Srivastava N, Krizhevsky A, et al. Improving neural networks by preventing co-adaptation of feature detectors[J]. Computer Science, 2013, 3(4): 212-223.
[10] Strubell E , Verga P , Belanger D , et al. Fast and Accurate Entity Recognition with Iterated Dilated Convolutions[J]. 2017.
[11] Peng N , Dredze M . Improving Named Entity Recognition for Chinese Social Media with Word Segmentation Representation Learning[J]. 2016.
[12] Gao J, Li M, Wu A, et al. Chinese Word Segmentation and Named Entity Recognition: A Pragmatic Approach.[J]. Computational Linguistics, 2005, 31(4):págs. 531-574.