Research on Chinese Semantic Role Labeling Based on Binary Strategy

Yicheng Wang,1  Ning Ma,1  Fucheng Wan,2  Lei Hao2

(1Key Laboratory of National language intelligent processing, Gansu Province, Lanzhou, 730030)
(2Key Lab of China's National Linguistic Information Technology, Lanzhou, 730030)

Abstract. Chinese SRL(Semantic Role Labeling) is a cutting-edge technology in the field of Chinese information processing. Now it has become one of the hottest research areas in Chinese information processing research. In this paper, based on the analysis of the previous SRL theory, the Chinese SRL technology, the comparison of advantages and disadvantages, this paper classifies and discusses the basic algorithm ideas of Chinese SRL. At the same time, the thinking on the SRL system is formed and a Chinese SRL system based on the binary strategy is proposed. The system adopts a binary strategy SRL method, combined with a statistical method based on linear feature combination research, fusion of Chinese short-sentence analysis of the study, to achieve a binary strategy-based Chinese SRL system design and construction. Through the experimental comparison, it is proved that the SRL system based on the binary strategy has a good annotation effect.

1. SRL introduction
SRL(Semantic Role Labeling) is an important technology for natural language processing and has important significance for improving the performance of language information processing systems. SRL is to analyze the predicate in a sentence by analyzing the predicate in the sentence, and then make a corresponding annotation on the semantic component of the predicate in the sentence, such as agent, recipient, tool or additional language. The process can be expressed as: setting up a set of role classification system, marking the structural components of the sentence in part, so that the computer automatically obtains certain "understanding" capabilities.

With the continuous development of artificial intelligence and Chinese text information technology, as well as the improvement of Chinese syntactic and semantic processing related technologies, Chinese SRL has become a key technology that connects the previous and the next, and the demand for accuracy has also increased, adding to the urgency of mastering semantic role tagging techniques and related situations. After ten years of development, although Chinese semantic role annotation has made a series of progresses, there are few researches on SRL in Chinese semantic characters. In this paper, based on the existing research status of SRL, the improvement and integration of existing algorithms are proposed, which has a positive effect on the research of Chinese information technology. This article is based on this background to start the research work.

2. An Overview of Chinese SRL System
At present, the core categories of Chinese SRL can be roughly divided into Chinese semantic character labeling based on linear feature combination and Chinese semantic character labeling based on syntactic analysis. Using a high-performance semantic character labeling system can not only reduce system cost, but also create a high-accuracy labeling model.

2.1. Two SRL Systems

Chinese SRL can be divided into two categories according to the system implementation characteristics: SRL based on linear feature combination and SRL based on syntactic analysis. SRL algorithm based on linear feature combination mainly converts the SRL problem into a labeling problem, which is solved by using probabilistic statistical models such as HMM, CRF and SVM. SRL of syntactic analysis is divided into SRL based on phrase structure syntax tree and SRL based on dependency syntax tree, the former is the most mature technology in syntactic analysis, and its performance and results are all reliable. The latter is a syntactic structure based on parsing and analysis, which shows a dependency graph between words. This article only selects the semantic role annotation based on syntactic tree of phrase structure as the research direction.

1) SRL based on linear feature combinations. The basic idea is to use certain continuous words in the sentence as the basic unit of annotation, then extract the features of the reference related linguistics rules in the unit, and form learning examples with the semantic role types of the unit, and finally use machine learning. The algorithm learns these instances and labels them to predict new instances. The system uses syntactic elements as the basic unit of semantic annotation, and uses the statistical annotation model of linear features to identify and classify the semantic roles of predicates in sentences simultaneously.

2) SRL based on syntactic analysis. This algorithm builds semantic role annotation on the basis of syntactic analysis. Firstly, it constructs a shallow parser model, then uses SRL, extracts the characteristics of short sentences or dependency syntax, and prunes syntax trees to filter syntax. It is not possible to act as a chunk phrase unit and a relational node of a semantic role on the tree, and annotate the role class for a chunk or a node that acts as a semantic role. The phrase structure syntax tree chosen in this paper refers to the output of the sentence structure analysis of sentence structure as a tree structure, that is, it analyzes each input sentence by constructing the syntactic tree of the phrase structure. This "tree of compositional structure" not only reveals the grammatical relationships of sentences, but also reveals the hierarchy of sentences.

2.2. Comparison of Chinese SRL Study

Compared with the deep semantic analysis, the SRL has the characteristics of clear question, easy labeling and technical feasibility. It has broad application prospects in public opinion analysis, sentiment analysis, and intelligent question answering systems. Therefore, research on Chinese SRL for computer processing Natural language texts are of great significance.

In recent years, with the rapid development of Chinese information technology, especially in syntactic and semantic related technologies, the short statement tree has been staggering in the field of SRL in Chinese, and more and more scholars have begun to apply it to natural language processing. The short sentence tree is free from the dependence of the traditional statistical learning methods on the features of manual extraction and greatly reduces people's workload. At present, many natural language processing tasks have achieved better results with the statistical model.

3. Chinese SRL based on binary strategy

With the rapid development of artificial intelligence, higher requirements have been put forward for precise and efficient word segmentation strategies. Based on the research of a variety of mainstream segmentation strategies, this paper integrates and improves a variety of existing methods, and completes an accurate and efficient word segmentation algorithm.

In this paper, based on the semantic feature annotation based on linear feature combination, the mature and stable short sentence tree algorithm based on syntactic analysis is incorporated, which
makes the combination of linear feature combination and syntactic analysis based SRL organically. The short sentence method tree gets rid of the dependence of the traditional statistical learning methods on the characteristics of manual extraction and greatly reduces people's workload. At present, many natural language processing tasks have achieved better results than the statistical model. Therefore, this paper will introduce the combination of the two into the SRL task and propose a Chinese SRL method based on the binary strategy, in order to provide a certain reference for the study of SRL. The overall framework diagram of the system is shown in Figure 1.

```
3.1. Research content and difficulties
The main research of this paper focuses on the development of Chinese SRL, and proposes a Chinese semantic role-based labeling mutual help model based on a binary strategy. Firstly, based on the open source Chinese semantic role annotation corpus, we screened and built the Chinese SRL corpus as experiment corpus, and determined the mark set used in this paper. Then we built the Chinese SRL corpus as experiment corpus and constructed and trained linear features based on them. The combined semantic role annotation model and the semantic role annotation model based on the short sentence tree; after that, it adds the trained word-of-speech feature vector, and combines the model of the linear feature combination with the model based on the short sentence tree with a certain weight ratio. After constructing and training again, a SRL model based on a binary strategy is obtained. This method can not only reduce the degree of reliance on the results of syntactic analysis, but also can more effectively perform semantic role annotation in the case of greatly reducing the workload of manually extracting features. The study is as follows:

1) Chinese SRL Based on Linear Feature Combination
2) Research on Chinese SRL Based on Short Statement Tree
3) Research on Chinese SRL Based on Linear Feature Combination and Short Sentence Method Tree

The main points of his research are the following:

1) Construction and training of semantic role annotation model based on linear feature combination;
2) Construction and training of semantic role annotation model based on short sentence tree;
3) Construction and training of fusion models.
```

Figure 1. Roadmap of Chinese SRL technology based on binary strategy
3.2. **SRL Model Based on Linear Feature Combination**

The conditional random field model is a probabilistic framework that uses conditional distribution probabilities. It is also a typical discriminant model. Compared to other sequence annotation models, the conditional random field model has many advantages. Therefore, selecting the conditional random field model in practical tasks can always obtain better performance. Considering the SRL as a sequence labeling problem, this paper selects a SRL method based on the linear chain conditional random field model.

1. **Conditional Random Field Model**: CRFs (Conditional Random Fields) is a discriminant probability model, which is used to label or analyze sequence materials. This paper improves the algorithm in the selection of the basic model of CRF. CRFs usually use the first-order linear chain structure shown in Figure 1.

![Figure 2 First-order linear chain structure of CRFs](image)

2. **Extraction of block-level features**: The features selected for this experiment include two parts, word-level features and block-level features. Word-level features refer to features based on words, parts of speech, positions of words relative to target words, and combinations of these features.

3. **Feature selection**: In the conditional random field algorithm, many important information of the data is given by the characteristic function. The CRF tool provides the definition of these feature functions in the form of feature templates, making it convenient to select features and define feature functions.

This experiment uses a method of selecting optimal features based on orthogonal table experiments. Multiple features based on the word level were selected. The algorithm implemented by the orthogonal table method is very simple and easy to implement parallel processing. Therefore, the selection of feature templates is easier. The general steps of Chinese semantic character labeling based on linear CRF are shown in Figure 3.

![Figure 3. Flow chart of Chinese SRL model based on linear CRF](image)

3.3. **SRL model based on short sentence tree**
In English, the related attempts to establish semantic role annotation based on the short sentence tree have achieved good results. In Chinese, shallow syntactic analysis has also made some progress. Therefore, this paper selects short sentence tree as another research idea for Chinese semantic roles.

Chinese SRL is based on shallow parsing. First, a chunk-based shallow parser is implemented, followed by SRL. At the syntactic analysis stage, according to the characteristics of Chinese language, the "false center morpheme" of words was used to simulate the central morphemes of words, which effectively alleviated the data sparsity problem and improved the performance of the syntactic analyzer. In the semantic role annotation stage, the morpheme characteristics of the target verb are obtained by means of word formation, and the characteristics of the verb are described from the perspective of the internal structure of the verb. The basic flow is shown in Figure 4.

![Figure 4. Flowchart based on a short sentence tree](image)

### 3.4. Implementation Based on Binary Strategy

(1) The implementation of corpus

This study will be based on a certain scale of SRL corpora. Through corpus acquisition, computerized word segmentation and part-of-speech labeling, manual proofreading of word segmentation and part of speech tagging, computer automatic syntactic labeling, and manual proofreading of syntactic labeling, a certain number of short-sentence tree libraries were constructed, based on a short sentence tree bank. Semantic roles were manually annotated. In addition, the tagging system also uses the HowNet grid frame dictionary as a reference for the SRL to ensure the accuracy of the tagging.

(2) Combination of quantitative analysis and qualitative analysis

Traditional language research is mainly introspective qualitative research and lacks scientific and objective data support. The simple quantitative analysis of the lack of guidance in linguistic theory will also be difficult. More and more language researchers have realized that if we want to analyze and study language in a comprehensive and objective manner, we must combine the two. This paper uses the theoretical knowledge of linguistics to analyze the statistical results of corpus in order to better verify some problems in linguistics.

(3) Combination based on rules and statistics

The rule-based approach refers to the establishment of a set of linguistic rules from the perspective of linguistics and cognitive science to make computers understand human languages. The statistics-based approach combines computer science and statistics to achieve a computer's understanding of a given natural language by processing large-scale corpus. In language information processing, setting a set of linguistic rules for the computer can help the machine better understand the complex natural language. Computers can use these rules to understand more new corpus and generate new language processing rules through statistics. These new rules can be applied to the next round of machine understanding. The two methods complement each other, making the process of natural language understanding faster.

(4) Man-machine mutual assistance method

In the semantic role annotation, manual labeling is time-consuming and laborious. It can no longer meet the needs of today's big data era. Machine labeling is not yet mature. Therefore, it is the most scientific and effective method to use the method of man-machine mutual assistance to mark up corpus. Through multiple trainings and refinements, an efficient, high-performance automated semantic tagging system will be trained.

### 4. Experimental setup and analysis

The experiment constructed Chinese trigger word dictionary, Chinese syntax tagging and entity semantic tagging corpus. At the same time, this experiment establishes a Chinese SRL model based on
statistical training, and adopts a short sentence tree to model and classify Chinese agents, affair, and verbs. Through the expert's role of manually annotating the role relations of the arguments and establishing the arguments of the arguments of the arguments, the construction of the Chinese SRL model based on the short sentence tree is realized. Two kinds of classical SRL models are compared based on linear feature combination and short sentence tree. Combining the advantages and disadvantages of these two models and selecting the best weight ratio, the above methods are integrated, and a Chinese SRL model of performance-optimized fusion binary strategy is proposed.

This experiment uses the data provided by the CoNLL2005 review as a training and test set, mainly from the Prop Bank with the Wall Street Journal (WSJ) as its source. Section 02-21 is the training set, Section 23-24 is the test set, and the statistical training method is used to repeat the experimental test and refer to the experimental data published by peer researchers. SRL methods and other single strategy SRL methods have a significant increase in performance and accuracy.

5. Conclusions and Next Steps
In this paper, based on the semantic feature annotation based on linear feature combination, the semantic role annotation of short sentence tree is combined, and a Chinese semantic role annotation method based on binary strategy is proposed. Compared to the traditional single strategy semantic role annotation method, this method is a new attempt to label Chinese semantic roles. The experimental results show that the two types of SRL methods are organically combined to preliminarily prove that the labeling effect is better and the accuracy of Chinese semantic character labeling is improved. In the follow-up research work, we will select more abundant and more effective related features and combine them with the features of phrases to further improve the system performance.

ACKNOWLEDGEMENT
This work is supported by the Fundamental Research Funds for the Central Universities(31920170155)" 
This work is supported by “the National Natural Science Foundation of China (Grant: 61602387)” and "the Fundamental Research Funds for the Central Universities(31920170155)"

References
[1] Sun, Honglin, Daniel Jurafsky. Shallow Semantic Parsing of Chinese[C]//Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics. Boston. USA: 2004.
[2] Sameer S. Pradhan, Wayne Ward, and James H. Martin. Towards robust semantic role labeling [J]. Comput. Linguist. 2008, 34(2): 289-310.
[3] Nianwen Xue. Labeling Chinese Predicates with Semantic roles [J]. Computational Linguistics, 2008, 34 (2): 225-255.
[4] Miezianko R, Pokrajac D. Detecting and Recognizing Abandoned Objects in Crowded Environments[C]//Proc. of AVSS’ 07. London, UK: [s. n.], 2007.
[5] Wang Xin, Sun Weiwei, Sui Zhifang. Research on Chinese semantic role annotation based on shallow syntactic analysis[J]. Chinese Journal of Information, 2011: 116-122.
[6] Li Jihong. Research on automatic annotation of Chinese framework semantic roles [D]. Taiyuan: Shanxi University. 2010.06.
[7] Wan Fucheng, Yu Hongzhi, Wu Yihong. Research on Tibetan Short Sentences for Machine Translation[J]. CEA, 2015(13): 211-215.
[8] Yu Jiande, Fan Xiaozhong, Pang Wenbo, Yu Zhengtao. Semantic Role Labeling Based on Conditional Random Fields[J]. Journal of Southeast University, 2007,23(3):361-364.
[9] Ding Weiwei, Chang Bing. Chinese Semantic Role Labeling and Classification Based on Maximum Entropy Principle[J]. Chinese Journal of Information, 2008, 22(6):20-27.