Chinese Shallow Semantic Parsing Based on Multi-method of Machine Learning

Fucheng Wan, Xiangzhen He*, Dongjiao Zhang, Guo Qi, Ao Zhu, Zhang Lei, Ning Zenan and Wang Yicheng

Key Laboratory of China's Ethnic Languages and Information Technology of Ministry of Education Northwest Minzu University, Lanzhou, China
E-mail: wanfucheng@126.com
**Corresponding Author*

Received 09 July 2020; Accepted 05 August 2020;
Publication 31 October 2020

Abstract

With the rapid development of 5G+ information intelligence, higher requirements are put forward for accurate and efficient semantic annotation methods. Semantic role annotation for any single method at present has its obvious and complementary advantages and disadvantages. Therefore, this paper attempts to introduce the above three mainstream and stable annotation methods into each task of semantic role annotation, and designs a Chinese semantic role annotation that integrates multi-method. This method integrates the statistical-based linear sequence method, the rule-based hierarchical tree method and the most advanced deep learning in the four processing modules of semantic role annotation. Multi-level linguistic features are introduced into the feature arrangement of the model to realize the mutual combination of multiple modules. Experiments show that the modular fusion of steps and methods effectively improves the annotation performance of each step of annotation.

Keywords: Semantic role labelling, multi-method, linear sequence, hierarchical tree, deep learning, modularization.

Journal of Web Engineering, Vol. 19_5–6, 685–706.
doi: 10.13052/jwe1540-9589.19565

1 Introduction

Fusion annotation method is a system-level integration optimization program, which deals with a multi-step annotation problem step by step, and realizes complementary advantages through fusion and optimization of various methods according to a certain combination method, which improves the overall annotation performance of the system. Because the method is easy to understand and relatively simple to implement, it can achieve good results, which has been widely used in many application fields of intelligent information processing.

Generally, fusion annotation methods can be divided into two categories according to their implementation methods: One is to select the optimal solution output by connecting multiple models in parallel; The second is to select the optimal processing units of multiple models to combine relay outputs. Machine translation is one of the earliest and most widely used intelligent information processing fields of fusion methods. In the past two years, this method has also been widely applied to multi-type natural language understanding tasks. For example, Huang et al. [1] obtained a machine translation system with a three-layer fusion model through multi-level integration of word and sentence-level machine translation models. In addition, fusion methods are also widely used in other fields of natural language processing. Liu et al. [2] applied this method to word sense disambiguation tasks. Lluis et al. [3] applied the idea of model fusion to the related tasks of shallow semantic analysis.

The system fusion program has been adopted by more and more natural language processing tasks, which is a model mutual aid program with great research prospects. The practice of this program in the field of machine translation is more prominent. At present, the machine translation system based on fusion method can be roughly divided into word-level, phrase-level and sentence-level fusion methods [4]. The core idea of its technology is to complete translation by step-by-step mutual relay of knowledge at the level of mutual reference words, phrase chunks and syntax among multiple modules. Fusion method technology promotes the development of machine translation to be more perfect and provides a good implementation model for system optimization. Therefore, the model fusion method in this paper will also draw more lessons from the system fusion research in the field of machine translation.

Based on the characteristics and applications of sub-model optimization technology and module fusion technology, and on the basis of theoretical accumulation and practical analysis, this paper proposes a system design

program of Chinese semantic role annotation integrating multi-methods. The multi-method model integrates the advantageous annotation units of the three semantic analysis models of linear sequence method, hierarchical tree method and deep learning in the four annotation steps, so that the annotation results transferred between modules are the optimal solution, and the semantic role annotation system is effectively optimized. The fusion method model provides technical support for Chinese information retrieval, human-computer interaction and Chinese public opinion analysis, which is of great significance to the research of Chinese information processing and Chinese syntactic-semantic relationship.

2 Related Works

Semantic role annotation is a practical program in current semantic analysis and processing, and it is also a major topic in computational linguistics research. In recent years, shallow semantic analysis has made great breakthroughs in computational language methods. It abandons the complexity of deep components and relationships, and deduces that the sentence meaning is displayed in a structured form by analyzing the lexical and syntactic structures of the target sentence. It can realize a fast analysis algorithm in the real corpus environment and obtain better accuracy than deep semantic analysis. In 1998, Lee [5] put forward the concept of semantic web. After that, semantic web and semantic analysis technology have developed rapidly. Semantic role annotation technology has gradually become a common research hotspot in international academic circles. During this period, many classical annotation methods emerged. In foreign countries, Gildea et al. [6, 7] proposed seven common basic features of semantic role annotation, and realized the important role of syntactic analysis in semantic role recognition and classification by statistical machine learning method based on context-free grammar to automatically label semantic roles, and pointed out the direction for future generations. After that, Prandhan et al. [8, 9] applied the machine learning method based on support vector machine to semantic role annotation, which further adds a variety of new features at the lexical level, and makes the annotation model obtain relatively ideal results. Xue et al. [10] tried to combine and label features of different levels, which proves that the annotation effect can be effectively improved after some features are combined. With the rise of artificial intelligence in the past two years, deep learnings have been applied to this field. Collobert et al. [11] applied deep neural networks to frame semantic role annotation. This method slows down

the manual intervention of traditional machine learning methods to deal with complex features and achieves ideal annotation results. Subsequently, multi-layer neural networks also began to be introduced into this field. Socher et al. [12] used the combination of neural network units and tree structure encoders to label, while Yin et al. [13] directly used multi-layer CNN models to label. In the application of shallow semantic analysis, Narayanan et al. [14] introduced semantic role annotation into the question answering system. Compared with the traditional pattern matching method, the new method greatly improves the accuracy of the answer. Wan [15] proposed the best answer extraction combining shallow semantic analysis to solve the problem in restricted domain, which improved the correlation between the answer sentence and the problem. Zhao et al. [16] took large-scale unlabeled data as a breakthrough to explore the automatic extraction of dependent syntactic features with high confidence from rough machined massive data.

In China, after more than ten years of development, semantic role annotation has also achieved rich research results. Liu Huaijun and others of Harbin Institute of Technology [17] proposed a maximum entropy semantic role annotation method for new features and their combination according to the characteristics of Chinese language. Chen Yaodong et al. [18] made syntactic and semantic classification of existing feature sets of statistical models in shallow semantic analysis. Li and others of Shanxi University [19, 20] treated the semantic role annotation task as a word sequence annotation problem for the first time. This program is different from most syntactic components as annotation units and opens up a brand-new idea. Wang Cheng et al. [21] proposed to make in-depth improvement on the basis of the semantic annotation model of conditional random fields, and incorporated multi-level linguistic features such as morphology and sentence patterns into the training process, which demonstrates that the annotation performance can be effectively enhanced. As deep learning continues to win the first place in the field of Chinese information processing, Wang Zhen et al. [22, 23] tried to apply the deep learning model of multi-layer network structure to the identification and classification of Chinese semantic roles. After that, the team tried to apply the bidirectional cyclic neural network algorithm to this field again. This method avoids a large number of complex feature extraction and can make better use of the information in the annotation sequence. Wang et al. [24] proposed to set up a "straight ladder unit" with information connection inside the multi-layer LSTM model unit. The annotation information can be quickly transmitted between different layers, while Li et al. [25] constructed a lightweight single-layer RNN model using external memory cells. The lightweight model has

the advantages of simple training, high annotation efficiency and the like, but its accuracy is close to that of the multi-level network model.

Generally speaking, although great progress has been made in the research on Chinese semantic role annotation, there is still a long way to go before it can be truly popularized and applied, and there are still few researches on its basic theory and application. As a syntactic and semantic cohesion tool, it is limited by corpus resources, Chinese lexical and syntactic constraints, as well as some differences brought by Chinese itself and English characteristics, which results in its relatively tortuous development and greater room for improvement.

3 Chinese Semantic Role Annotation Combining Multi-method

3.1 Method Introduction

System fusion technology has made gratifying achievements in machine translation, which greatly promotes the popularization and application of this method in the field of semantic analysis. For technologies with mature basic theoretical research such as semantic role annotation technology, fusion method provides a new expansion space for its performance improvement. However, there are few researches on semantic analysis of system fusion technology and it has only begun to appear in recent years. In foreign countries, Kontostathis [26] integrated vector space model into semantic analysis, which improves the performance of the system to a certain extent. Atreya et al. [27] incorporated BM25 technology into the latent semantic analysis model to better improve the analysis performance. Muhammad Hossain et al. [28] proposed to construct a semantic analysis model based on causality, which uses causality pairs to map into input/output pairs. In China, Zhang [29] proposed an analysis model combining morphology, syntax and semantics to construct a multi-level parallel joint model, which effectively avoided the spread of errors and local optimization problems, thus improving the analysis performance of each step. Xu et al. [30] integrated short syntax, dependent syntax model and related features into traditional semantic role annotation, and its performance was better than that of similar systems with a single method. Ren et al. [31] proposed to integrate clustering method word quantitative representation into the traditional model and combine ternary grammar model to obtain a potential semantic analysis model applied to speech recognition. It can be seen from this that the fusion method technology has shown a thriving and strong development momentum. At the same time,

semantic role annotation technology has also entered the stage of technology optimization after more than ten years of rapid growth, and the idea of fusion method undoubtedly provides a new breakthrough point for model optimization of semantic role annotation.

At present, most semantic role annotation systems rely on statistical machine learning methods, which are bound to be affected by the infinity of feature tags. Moreover, current research shows that when the existing semantic role annotation is applied to large-scale corpus sets, the hardware overhead required is extremely high, which greatly hinders its application and promotion. On the one hand, if large-scale complex feature problems are effectively split into multi-module step-by-step labeling methods, the complexity of modules will be greatly reduced, and the optional models of the step-by-step modules will be more extensive. After that, the multiple modules will be effectively fused by using the fusion method, which can not only solve the problem that the training characteristics are too complex, but also absorb the advantageous processing units of the multiple modules, so that the fused system performance can be brought into full play.

On the other hand, the in-depth research on linear sequence method, hierarchical tree method and deep learning in the previous chapters found that these three annotation methods mainly have the following characteristics in the annotation process:

(1) In the pruning, preprocessing and post-processing stages, the hierarchical tree method has greater advantages than other methods. Using the pruning algorithm based on hierarchical tree can eliminate non-semantic role lexical elements to the greatest extent.
(2) The argument recognition stage is essentially a binary classification problem, so the linear sequence method and the deep learning have natural advantages. Considering the performance ratio of argument recognition, the binary classification model constructed by the linear sequence method is lighter and concise, with less system overhead, and is more suitable for the argument recognition stage.
(3) As the stage of role classification is equivalent to multiple classification problems, this paper flexibly adds multiple groups of hierarchical linguistic features, which increases the difficulty of classification. Through experimental verification, deep learning is more competent for classification and annotation problems with strong hierarchy.

In addition, a large number of studies show that increasing the scale of training corpus can improve the annotation performance based on statistical methods. However, when the training corpus reaches a certain scale, this

method will no longer work. Continuing to increase the training corpus will only lead to a large number of interference noises and redundant labels in the model, which will lead to performance degradation due to too large a model.

3.2 System Construction

The research of system construction focuses on the idea of fusion method and different semantic role annotation methods mentioned in the previous three chapters. A large number of preliminary experiments and reading literatures are carried out according to the research status of this problem, and a joint mutual aid model of Chinese semantic role annotation that integrates multi-method is proposed. Firstly, the Chinese semantic role annotation corpus with multi-level linguistic features is selected as the basic corpus. Through Bootstrapping corpus self-expansion mechanism, short syntactic features and dependent syntactic features are added to the basic corpus and manually proofread as the fusion method corpus. Then, the semantic role pruning module based on hierarchical tree method is constructed and trained with the constructed fusion method corpus. After the pruning operation is completed, a "Pruning" tag column is added, the corpus with pruning tag is sent into a conditional random field classification model based on two classifications, and an argument recognition module based on linear sequence method is trained and constructed. After the argument recognition stage is completed, the corpus is added with an "Argument" tag column on the basis of the previous step, and the recognized corpus is sent to a semantic role classification model based on multi-classification, and a role classification module based on deep learning is trained and constructed. In the post-processing stage, according to simple post-processing rules and semantic role post-processing methods based on short syntax tree, the morphology of syntax tree and subtree is analyzed, argument boundary correction and other processing are carried out, and a post-processing module based on short syntax tree is constructed. Finally, the multi-method combination and system module coupling technology are used to interface the four modules in a unified way and connect them in series in sequence to obtain a semantic role annotation model integrating multi-method. The general technical route is shown in Figure 1.

3.2.1 Pruning module of hierarchical tree method

Since most syntactic components in the short syntax tree are non-predicate-argument parts, heuristic algorithm is used to prune them. The pruning of sentences mainly includes parenthesis and parallel structure pruning. Parentheses in sentences are generally independent components. Removing them

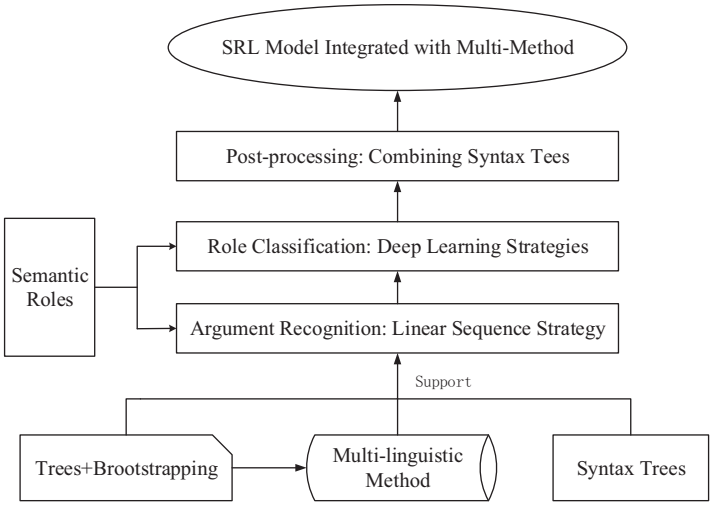

Figure 1 Technology roadmap of fusion method.

from the main meaning of the sentence will not change and the syntactic tree structure can be simplified. For the juxtaposed components in the juxtaposed structure, because they are of the same importance in the sentence, the first of the juxtaposed components is usually reserved and the rest is cut off.

The pruning process starts from the predicate node of the syntax tree to merge upward, first merging the brother nodes of the same layer of the predicate node, and then merging the parent nodes of the current node upward in turn until the root node of the syntax tree. The model judges whether there are parentheses or parallel structures in this layer. If it contains parentheses, the parentheses are cut off. When the parentheses contain predicates, the predicates and related arguments in the parentheses are preserved. If a parallel structure is included and the parallel structure does not contain predicates, the parallel component is removed. As shown in Figure 2, the pruning process of the short syntax tree in the example sentence "the supervision team has been stationed in the epidemic area and is rapidly carrying out epidemic prevention work" is as follows: locate the core predicate of the sentence as "VV-carry out", merge from the node, and first merge the sibling node "NP-((NN-epidemic prevention) (NN-work))" of the node into the candidate queue; Then move up to node VP in turn, and the sibling nodes of node VP are modifiers, so merge "ADVP-is" and "ADVP-fast" into the candidate queue, and continue to move the pointer to another VP node on the previous layer, whose sibling nodes are connected by "PU-," and belong to parallel

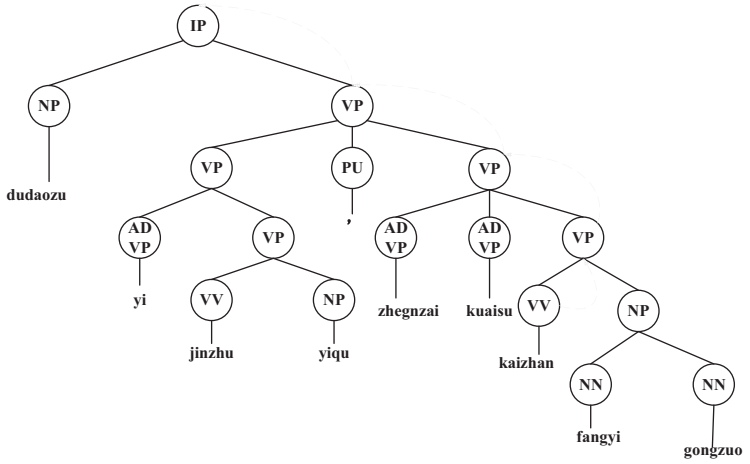

Figure 2 Schematic diagram of pruning process.

components, so cut them off; Finally, it moves to the upper VP node, merges the brother node "NP-Supervision Group" into the candidate queue, and the algorithm terminates.

3.2.2 Argument identification module of linear sequence method

The argument recognition module of the linear sequence method takes the CRF annotation model set up in the third chapter as the template, CRFs model still adopts CRF-L2 algorithm, and the best adjustment parameter $\mathrm{C} = 4.0$ and domain offset factor $\varphi(\Theta) = 1.0$ of the first-order transfer feature of the output sequence. After pruning, the results are output to the newly added "IS_Pruning" tag column, and the corpus with pruning tag is sent to the conditional random field classification model based on two classifications, and the argument recognition module based on linear sequence method is trained and constructed. The recognition module has the advantages of light weight, high efficiency and low complexity. Each labeled word element is taken as a sample, that is, each word element and the "IS_Argument" label column to which it belongs are taken as a sample, so that the value of the sample is determined quickly and efficiently.

3.2.3 Role classification module of deep learning

According to the idea of constructing the method module in this section, the role classification module of the deep learning takes the neurons of Bi-LSTM neural network model designed in the fifth chapter as the template. The markup corpus in the previous section removes the "IS_Pruning"

markup columns and rearranges the order of the markup columns. The "IS_Argument" tag column is moved forward to the lexical feature and rearranged as the training corpus for the module. The Bi-LSTM network layer is supervised to be trained by using the multi-level linguistic feature tags corresponding to each lexical element vector as the basic input unit, and the multi-level linguistic feature tags corresponding to each lexical element are used. Semi-supervised learning of multi-level features and type tags of semantic roles, training and constructing a role classification module based on deep learning, so that the vector representation of semantic role information of lexical elements can be automatically obtained for the sentence corpus of the new input module. The training of each layer of the network adopts a bidirectional propagation algorithm composed of Forward pass and Backward pass, the whole technology route is like Figure 3 below.

3.2.4 Post-processing module with syntax tree

The post-processing module can further restrict the semantic role annotation model. By analyzing the obvious errors of model annotation, the corresponding constraint restrictions can be formulated according to the actual situation

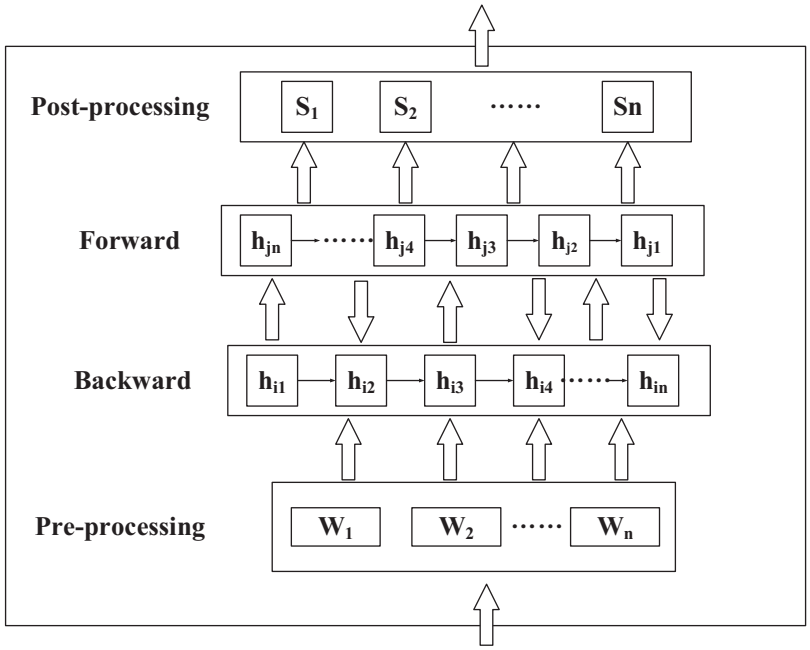

Figure 3 Schematic diagram of role classification module based on Bi-LSTM.

and needs. According to the characteristics of predicates in the short syntax tree, the following post-processing rules are formulated:

(1) Each predicate in the syntactic tree cannot have two or more identical semantic roles.
(2) The scope of the qualifying predicate. The scope of predicates is divided into two categories. Class I: the scope of predicates is the whole short syntax tree, such as "say", "tell", "answer", "you", etc. Class II: The scope of predicate is the subtree of the syntactic tree where it is located.
(3) Among the annotation results with nested relation, the semantic role with the highest probability is retained.
(4) In order to prevent the semantic tags from being sparse, a small number of semantic role types are merged.

Through the above rules, the argument roles that have been classified are filtered to solve the conflict roles that do not meet the semantic structure restrictions of sentences and obtain the final annotation results, thus improving the accuracy of classification.

For example, AA represent for Parent Node Part of Speech, BB represent for Sub-node Part of Speech, CC represent for Dependency Type, DD represent for Phrase Structure Type, EE represent for Parent Node Part of Speech, FF represent for Sub-node Part of Speech, GG represent for Dependency Type, HH represent for Phrase Structure Type, II represent for Parent Node Part of Speech, JJ represent for Sub-node Part of Speech, KK represent for Dependency Type, LL represent for Phrase Structure Type.

Table 1 Double syntactic rules used in post-processing (part)

AA	BB	CC	DD	EE	FF	GG	HH	II	JJ	KK	LL
V	D	ADV	VP	Q	C	LAD	QP	N	N	COO	NP
V	P	ADV	VP	Q	M	RAD	QP	N	M	IS	NP
V	V	ADV	VP	P	P	COO	PP	N	N	IS	NP
V	V	COO	VP	P	C	LAD	PP	D	P	ADV	ADVP
V	N	VOB	VP	P	N	POB	PP	R	N	SBV	IP
V	V	VOB	VP	N	N	ATT	NP	R	V	SBV	IP
Q	Q	COO	QP	N	U	ATT	NP	V	NR	VOB	VP
Q	V	IS	QP	N	NS	ATT	NP	T	N	SBV	IP

4 Fusion Method Experiment

First of all, Bootstrapping self-expansion mechanism is adopted to expand the corpus selected in the experiment with double syntactic feature tags, and the linguistic professionals are handed over for simple manual proofreading. A total of 2853 sentences are obtained, including 2710 training corpus and 143 test corpus. Three groups of experiments were carried out to test the contribution of different modules to the system annotation performance. The performance of semantic role annotation system is evaluated, and Precision, Recall and F-Score are used to evaluate the performance of the system.

4.1 Multi-feature Corpus Construction

The original corpus used in the experiment is the dependent corpus of Tsinghua University and Harbin Institute of Technology, which is oriented to the news field. Referring to the short syntactic information annotation standard of Chinese Penn Treebank Tag Set, feature column processing is carried out to realize the construction of multi-feature Chinese semantic role annotation corpus. In this process, it integrates various linguistic features such as morphology and syntax, and inherits the criteria for the construction of traditional semantic role annotation corpus such as predicate division and semantic role recognition. After processing and screening, a total of 22,000 sentences of corpus were obtained, including 20,000 sentences of training corpus and 2,000 sentences of testing corpus. Statistics on the main semantic roles in the corpus are shown in Table 2.

As shown in Figure 4 below, a sentence column string is extracted from the original corpus, and it can be seen that it contains multiple column features, which need to be screened. Where ID is the word element number column; Seg is a lexical ontology column; PoS is the original part-of-speech granularity feature mark column; Gran is the feature label column of part of

Table 2 Occurrence frequency statistics of main semantic roles

Semantic role	Frequency	Semantic role	Frequency
Core predicate	21981	Time	2997
Agent	8188	Connection	11288
Patient	10692	Preposition	8621
Degree	4074	Azimuth	3539
Premises	3651	Reason	372

ID	Seg	PoS	Gran	Syn	Rel	Dep
1	We	r	Rr	_	6	Agent
2	Today	t	T	_	6	Time
3	Also	d	D	_	6	Comments
4	No	v	V	_	6	Predicate
5	Thoroughly	a	Ad	_	6	Degree
6	Resolve	v	V	_	0	Core
7	Corruption	a	An	_	8	Limit
8	Problem	n	N	_	6	Patient

ID	Seg	C_G	SRL
1	We	r	Agent
2	Today	t	Time
3	Also	d	*
4	No	v	Predicate
5	Thoroughly	a	Degree
6	Resolve	v	Core component
7	Corruption	a	*
8	Problem	n	Patient

Figure 4 An example diagram of coarse-grained corpus construction of part of speech.

speech with fine granularity. Syn is a short syntactic signature column and has not been added yet. Rel is the syntactic pointing label feature tag column; Dep is a dependent feature tag column; C_G is the coarse graininess feature column of part of speech; SRL annotates columns for semantic information.

Aiming at the non-linear mapping relationship between syntax and semantics, based on the flexible addition of multi-level features by inheriting linear sequence method, short syntactic features and dependent syntactic feature tags are added to the basic corpus by using Bootstrapping corpus self-expansion mechanism, and simple manual proofreading is carried out

as the corpus of fusion method. After the annotation is completed, a short syntactic tag and a dependent syntactic tag column are added to the sentence column corpus. By counting the semantic roles of different syntactic tags, the semantic role types of different syntactic roles and the ability of this syntactic role to act as semantic roles can be obtained, which paves the way for the following pruning and post-processing.

4.2 Modular Experiment

Experiment 1: The processed training corpus is sent to the pruning module of hierarchical tree method and the Chinese syntax parser of Stanford University to perform pruning operation respectively. The results are compared as shown in Table 3 below.

The experiment shows that the pruning effect of the pruning method based on hierarchical tree is greatly improved compared with the pruning effect of Stanford University's Chinese syntax parser. The latter has a slight advantage in recall rate, but there is a big gap in the accurate ratio of pruning. Results Error analysis found that the recall rate of the pruning method based on hierarchical tree was insufficient because the pruning operation of syntactic tree removed semantic role nodes that were difficult to identify in subtrees.

Experiment 2: After the pruning operation is completed, the test results of whether to prune "Y/N" are imported into the "IS_Pruning" feature column, and then sent into the conditional random field classification module based on two classifications. Each word element in the sentence column is taken as a unit to train and test the recognition accuracy of the module, and the final test results are imported into the corresponding "IS_Argument" feature column of the word element. When the markup is complete, The "IS_Pruning" feature column in the corpus is removed, Move the "IS_Argument" feature column forward after the lexical feature column, The rearranged training corpus is used as a role classification module, the sentence column corpus is converted into lexical element vectors as basic input units, the role classification module based on deep learning is trained and constructed, the vector expression of semantic role information of lexical elements is obtained after the training is

Table 3 Comparison of experimental results of two types of pruning modules

Pruning device	Argument Role Recall	Node pruning
Hierarchical Tree Pruning Module	83.70	89.34
Standford Parser	84.59	86.12

completed, and the semantic role components are imported into the columns to be labeled.

As can be seen from Table 4, when the model in this chapter is tested with the same corpus as Bi-LSTM, the annotation results are improved compared with the single improved Bi-LSTM model built in the previous chapter, which shows that the multi-method and modular annotation model in this chapter is feasible. However, due to the high complexity of the model and the transmission of annotation results by multiple sub-modules, there are some model efficiency disadvantages compared with a single class of neural network model.

Comprehensive analysis shows that: The performance of lexical vectors trained by using the multi-feature corpus constructed in this paper is better than that provided by the conference CoNLL-2009. The main reason is that the corpus integrates multi-level linguistic features, and after pruning and linear argument recognition module processing, the training data sent into Bi-LSTM model has wide dimensional information and rich semantic information. Therefore, experiments show that adding multi-level linguistic features to role classification tasks can significantly improve classification performance.

Experiment 3: After the role classification is completed, the basic information of lexical elements, short syntactic features and the final semantic role annotation results are selected from the annotation results and sent to the post-processing module combining the syntactic tree, and the abnormal node deletion operation in the syntactic tree is executed according to the processing rules. Therefore, the syntactic component nodes corresponding to the final semantic role are determined, and the role classification is merged.

As can be seen from Table 5, experiment 3 adds post-processing rules to experiment 2, and the overall F1 value reaches 81.72%. The application

Table 4 Comparison of experimental results of semantic role annotation process

Annotation Task	Accuracy	Recall	F1
Argument recognition	86.29	85.51	85.96
Role Classification	83.34	79.73	81.50

Table 5 Comparison of experimental results of post-processing modules

	Accuracy	Recall	F1
Before post-treatment	83.34	79.73	81.50
After post-treatment	83.56	79.96	81.72

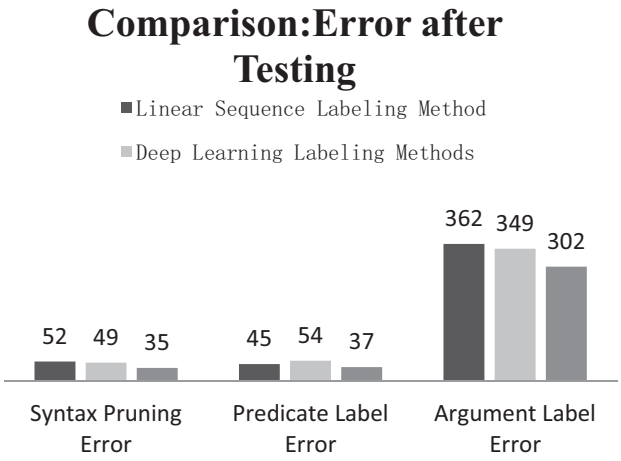

Figure 5 error mark statistic of experimental results of each method.

of post-processing module can also improve the overall performance of the system.

After classifying and counting the annotation errors, it is found that after adding multiple types of linguistic features, model training and decoding test experiments are carried out, and 500 test corpus are tested. The results are compared with those shown in Figure 5 below.

The results show that the hierarchical tree method combining linear sequence and deep learning has obvious effect on improving the performance of semantic role annotation, especially when combining double syntactic tree analysis. The F1 value of the hierarchical tree model is increased by nearly 1.5% from the F value labeled by the best single method, which proves the effectiveness of the method.

5 Conclusion

Traditional semantic role labelling task need this process that the syntactic structure of the sentence is given, Roth and Lapata (2016) construct an deep-learning model to obtain the syntactic dependency paths information; while Marcheggiani and Titov (2017) construct Graph Convolutional Networks to encode the dependency structure of the sentence. Although He et al. (2017)'s approach is a pure end-to-end learning, they have included an analysis of adding syntactic dependency information into English SRL in the discussion section. Cai et al. (2018) have compared syntax-agnostic and syntax-aware

approaches and Xia et al. (2019) have compared different ways to represent and encode the syntactic knowledge.

This paper briefly introduces the research status of fusion method and its related application research in the field of natural language processing. On the basis of previous studies, the prominent points of each annotation method are comprehensively analyzed, and a multi-method fusion annotation idea is proposed. A multi-module fusion annotation method based on the pruning module of the hierarchical tree method, the argument recognition module of the linear sequence method, the role classification module of the deep learning and the post-processing module combined with syntax tree is realized by integrating the linear sequence method, the hierarchical tree method and the deep learning. The experimental results show that the fusion method is an effective multi-level semantic role annotation optimization program, and reasonable fusion of the advantageous annotation units of multi-method can better improve the annotation performance.

Acknowledgements

This research is supported by Key talent projects in Gansu Province (2020RCXM106) and Dual first-class and characteristic development guidance Special Fund project.

References

[1] F. Huang, K. Papineni. Hierarchical System Combination for Machine Translation// In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Prague: Association for Computational Linguistics, 2007. 277–286.

[2] Y.P. Liu, S. Li, T.J. Zhao, Systematic Fusion Based on Word Net Word Sense Disambiguation. Acta Automata Sinica, 2010, 36 (11): 1575–1580.

[3] L. Mdrquez, M. Surdeanu, P. Cmas, et al. A robust combination strategy for semantic role annotation// In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing. Stroudsburg: Association for Computing Machinery, 2005. 644–651.

[4] M.X. Li, C.Q. Zong, Review of Machine Translation System Fusion Technology. Chinese Journal of Information Technology, 2010, 24 (4): 74–84.

[5] R. Lawrence, Rabiner Digital Speech Processing Theory and Application Beijing. Electronic Industry Press.
[6] D. Gildea, D. Jurafsky, Automatic annotation of semantic roles. Computational Linguistics, 2002, 28(3): 245–288.
[7] D. Gildea, M. Palmer, The necessity of syntactic parsing for predicate argument recognition//In Proceedings of ACL-2002. Philadelphia, USA, 2002: 239–246.
[8] S. Pradhan, K. Hacioglu, V. Krugler, et al. Support vector learning for semantic argument classification. Machine Learning Journal, 2005, 60(1): 11–39.
[9] S. Pradhan, W. Ward, K. Hacioglu, et al. Semantic role annotation using different syntactic views//In Proceedings of ACL-2005. Ann Arbor, USA, 2005: 581–588.
[10] N. Xue, M. Palmer, Calibrating features for Semantic Role Labelling//In Proceedings of EMNLP-2004. Barcelona, Spain, 2004: 88–94.
[11] C. Ronan, W. Jason. A unified architecture for natural language processing: Deep neural networks with multitask learning// Proceedings of the 25th international conference on machine learning. 2008: 160–167.
[12] R. Socher, E. Huang, Jeffrey Pennington, et al. Dynamic Pooling and Unfolding Recursive Autoencoders for Paraphrase Detection// In Proceedings of the NIPS, 2011: 801–809.
[13] W.P. Yin, H. Schutze. Convolutional Neural Network for Paraphrase Identification// In Proceedings of the HLT-NAACL, 2015: 901–911.
[14] S. Narayanan, S. Harabagiu, Question answering based on semantic structures// Proceedings of the 20th International Conference on Computational Linguistics. Geneva, Switzerland, 2004: 693–701.
[15] F.C. Wan, Extracting Algorithm for the Optimum Solution Answer Oriented Towards the Restricted Domain. IPPTA: Quarterly Journal of Indian Pulp and Paper Technical Association. 2018, 30(5): 590–597.
[16] H. Zhao, W.L. Chen, C. Kit, Semantic dependency parsing of NomBank and PropBank: an efficient integrated approach via a large-scale feature selection//Proceedings of the CoNLL-2009. Boulder: ACL Press, 2009: 30–39.
[17] H.J. Liu, W.X. Che, T. Liu, Feature engineering of Chinese semantic role annotation. Chinese Journal of Information Technology. 2007, 22 (1): 79–84.
[18] Y.D. Chen, T. Wang, H.W. Chen, Shallow semantic analysis of semi-supervised learning and active learning. Chinese Journal of Information Technology, 2008 (02): 70–75.

[19] J.H. Li, R.B. Wang, W.L. Wang, et al. Automatic annotation of semantic roles in Chinese frames. Acta Software Sinica, 2010, 30 (4): 597–611.
[20] J.H. Li, Research on Automatic Annotation Technology of Chinese Frame Semantic Roles. Taiyuan: Shanxi University. 2010.
[21] Y.C. Wang, F.C. Wan, N. Ma, Multi-clue Chinese Semantic Role Annotation Based on Conditional Random Fields. Journal of Yunnan University (Natural Science Edition), 2020, 42 (3): 474–480.
[22] Z. Wang, T.S. Jiang, B.B. Chang, et al. Chinese Semantic Role Annotation with Bidirectional Recurrent Neural Networks// Lisbon, Portugal: Proceedings of 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 1626–1631.
[23] W. Zhen, B.B. Chang, Z.F. Sui, Chinese semantic role annotation based on hierarchical output neural network. Chinese Journal of Information Technology, 2014, 28 (6): 56–61.
[24] M.X. Wang, Liu Q. Semantic role annotation based on deep neural network. Chinese Journal of Information Technology, 2018, 32 (02): 50–57.
[25] T.S. Li, Q. Li, W.H. Wang, B.B. Chang, Text Retelling Discriminant Model Based on External Memory Unit and Semantic Role Knowledge. Chinese Journal of Information Technology, 2017, 31 (06): 33–40.
[26] A. Kontostathis. Essential Dimensions of Latent Semantic Indexing (EDLSI) // In: Proceedings of the 40th Annual Hawaii International Conference on System Sciences. Kona Hawaii: IEEE CS Press, 2007. 73–80.
[27] A. Atreya, C. Elkan. Latent Semantic Indexing (LSI) Fails for TREC collections. SIGKDD Explorations, 2011, 12(2): 5–10.
[28] M.M. Hossain, V. Prybutok, N. Evangelopoulos. Causal Latent Semantic Analysis (cLSA): An Illustration. International Business Research, 2011, 4(2): 38–50.
[29] M.S. Zhang, Research on Joint Analysis Model of Chinese Lexical, Syntactic and Semantic. Harbin Institute of Technology, 2014.
[30] J. Xu, J.H. Li, Q.M. Zhu, et al. Chinese semantic role annotation based on phrases and dependent syntactic structures. Computer Engineering, 2011, 37 (24): 169–172.
[31] J.S. Ren, Z.Y. Wang, A New Language Model for Latent Semantic Analysis. High Technology Communications, 2005, 15 (8): 1–5.

Biographies

Fucheng Wan, (1985-), male, China, Liaoning Province, Northwest Minzu University, associate professor, master's tutor, research direction contain natural language processing, Tibetan-Chinese machine translation, information extraction, automatic question and answer research. Published more than 20 core papers, writing 4 books, access to patents and software copyright more than 10 items.

Xiangzhen He, (1977-) China, Ningxia Province, Northwest Minzu University, associate professor, master's tutor, research direction contain natural language processing and motion capture. Published more than 40 core papers.

Dongjiao Zhang, born in 1996 in Qitaihe, Heilongjiang Province, is now a graduate student in Northwest University for nationalities. Her research direction is data visualization and has published a paper.

Guo Qi, graduate student of China National Information Technology Research Institute, Northwest University for Nationalities, whose main research interests are natural language processing and information extraction.

Ao Zhu, graduate student of China National Institute of Information Technology, Northwest University for Nationalities. He is from shanxi Province. His research direction is shallow semantic analysis, and have published a paper and applied for a soft copy.

Zhang Lei is a graduate student at Northwest University for Nationalities since 2019. He researches automatic question answering technology. He has published a paper and a software book.

Ning Zenan, born in 1996 in Yuncheng City, Shanxi Province. I was a graduate student in Northwest University, and published a research paper in the direction of national science and technology.

Wang Yicheng, from Luliang, Shanxi Province. He obtained a master's degree from Northwest University for nationalities. His research direction is semantic role analysis. He has published two CSCD papers.