Sequential Tagging of Semantic Roles on Chinese FrameNet

Jihong LI
Computer Center
Shanxi University
lijh@sxu.edu.cn

Ruibo WANG, Yahui GAO
Computer Center
Shanxi University
{wangruibo,gaoyahui}@sxu.edu.cn

Abstract

In this paper, semantic role labeling(SRL) on Chinese FrameNet is divided into the subtasks of boundary identification(BI) and semantic role classification(SRC). These subtasks are regarded as the sequential tagging problem at the word level, respectively. We use the conditional random fields(CRFs) model to train and test on a two-fold cross-validation data set. The extracted features include 11 word-level and 15 shallow syntactic features derived from automatic base chunk parsing. We use the orthogonal array of statistics to arrange the experiment so that the best feature template is selected. The experimental results show that given the target word within a sentence, the best F-measures of SRL can achieve 60.42%. For the BI and SRC subtasks, the best F-measures are 70.55 and 81%, respectively. The statistical t-test shows that the improvement of our SRL model is not significant after appending the base chunk features.

1 Introduction

Semantic parsing is important in natural language processing, and it has attracted an increasing number of studies in recent years. Currently, its most important aspect is the formalization of the proposition meaning of one sentence through the semantic role labeling. Therefore, many large human-annotated corpora have been constructed to support related research, such as FrameNet (Baker et al., 1998), PropBank (Kingsbury and Palmer, 2002), NomBank (Meyers et al., 2004), and so on. On this basis, several international semantic evaluations have been organized, which include Senseval 3 (Litkowski, 2004), SemEval 2007 (Baker,et al., 2007), CoNLL 2008 (Surdeanu et al., 2008), CoNLL 2009 (Hajic et al., 2009), and so on.

The first SRL model on FrameNet was proposed by Gildea and Jurafsky(2002). The model consists of two subtasks of boundary identification(BI) and semantic role classification(SRC). Both subtasks were implemented on the pretreatment results of the full parsing tree. Many lexical and syntactic features were extracted to improve the accuracy of the model. On the test data of FrameNet, the system achieved 65% precision and 61% recall.

Most works on SRL followed Gildea's framework of processing the SRL task on English FrameNet and PropBank. They built their model on the full parse tree and selected features using various machine learning methods to improve the accuracy of SRL models. Many attempts have made significant progress, ssuch as the works of Pradhan et al. (2005), Surdeanu et al. (2007), and so on. Other researchers regarded the task of SRL as a sequential tagging problem and employed the shallow chunking technique to solve it, as described by Marquez at al. (2008).

Although the SRL model based on a full parse tree has good performance in English, this method of processing is not available in other languages, especially in Chinese. A systemic study of Chinese SRL was done by Xue et al. (2008). Like the English SRL procedure, he removed many

Proceedings of the 8th Workshop on Asian Language Resources, pages 22–29,
Beijing, China, 21-22 August 2010.

uncorrelated constituents of a parse tree and relied on the remainder to identify the semantic roles using the maximum entropy model. When human-corrected parse is used, the F-measures on the PropBank and NomBank achieve 92.0 and 69.6%, respectively. However, when automatic full parse is used, the F-measures only achieve 71.9 and 60.4%, respectively. This significant decrease prompts us to analyze its causes and to find a potential solution.

First, the Chinese human-annotated resources of semantic roles are relatively small. Sun and Gildea only studied the SRL of 10 Chinese verbs and extracted 1,138 sentences in the Chinese Tree Bank. The size of the Chinese PropBank and Chinese NomBank used in the paper of Xue is significantly smaller than the ones used in English language studies. Moreover, more verbs exist in Chinese than in English, which increases the sparsity of Chinese Semantic Role data resources. The same problem also exists in our experiment. The current corpus of Chinese FrameNet includes about 18,322 human-annotated sentences of 1,671 target words. There is only an average of less than 10 sentences for every target word. To reduce the influence of the data sparsity, we adopt a two-fold cross validation technique for train and test labeling.

Second, because of the lack of morphological clues in Chinese, the accuracy of a state-of-the-art parsing system significantly decreases when used for a realistic scenario. In the preliminary stage of building an SRL model of CFN, we employed a Stanford full parser to parse all sentences in the corpus and adopted the traditional SRL technique on our data set. However, the experiment result was insignificant. Only 76.48% of the semantic roles in the data set have a constituent with the same text span in the parse tree, and the F-measure of BI can only achieves 54%. Therefore, we attempted to use another processing technique for SRL on CFN. We formalized SRL on CFN into a sequential tagging problem at the word level. We first extracted 11 word features into the baseline model. Then we added 15 additional base chunk features into the SRL model.

In this paper, the SRL task of CFN comprises two subtasks: BI and SRC. These are regarded as a sequential tagging problem at the word level. Conditional random fields(CRFs) model is employed to train the model and predict the result of the unlabeled sentence. To improve the accuracy of the model, base chunk features are introduced, and the feature selection method involving an orthogonal array is adopted. The experimental results illustrate that the F-measure of our SRL model achieves 60.42%. This is the best SRL result of CFN so far.

The paper is organized as follows. In Section 2, we describe the situation of CFN and introduce SRL on CFN. In Section 3, we propose our SRL model in detail. In Section 4, the candidate feature set is proposed, and the orthogonal-array-based feature selection method is introduced. In Section 5, we describe the experimental setup used throughout this paper. In Section 6, we list our experimental results and provide detailed analysis. The conclusions and several further directions are given at the end of this paper.

2 CFN and Its SRL task

Chinese FrameNet(CFN) (You et al., 2005) is a research project that has been developed by Shanxi University, creating an FN-styled lexicon for Chinese, based on the theory of Frame Semantics (Fillmore, 1982) and supported by corpus evidence. The results of the CFN project include a lexical resource, called the CFN database, and associated software tools. Many natural language processing(NLP) applications, such as Information Retrieval and Machine Translation, will benefit from this resource. In FN, the semantic roles of a predicate are called the frame elements of a frame. A frame has different frame elements. A group of lexical units (LUs) that evokes the same frame share the same names of frame elements.

The CFN project currently contains more than 1,671 LUs, more than 219 semantic frames, and has exemplified more than 18,322 annotated sentences. In addition to correct segmentation and part of speech, every sentence in the database is marked up to exemplify the semantic and syntactic information of the target word. Each annotated sentence contains only one target word.

(a). <medium-np-subj 第 1/m 章/q > <tgt=" 陈述" 介绍/v > <msg-np-obj 算法/n 与/c 数据/n

结构/n > ； /w

The CFN Corpus is currently at an early stage, and the available CFN resource is relatively limited, so the SRL task on CFN is described as follows. Given a Chinese sentence, a target word, and its frame, we identify the boundaries of the frame elements within the sentence and label them with the appropriate frame element name. This is the same as the task in Senseval-3.

3 Shallow SRL Models

This section proposes our SRL model architecture, and describes the stages of our model in detail.

3.1 SRL Model Architecture

A family of SRL models can be constructed using only shallow syntactic information as the input. The main differences of the models in this family mainly focus on the following two aspects.

i) model strategy: whether to combine the subtasks of BI and SRC?

ii) tagging unit: which is used as the tagging unit, word or chunk.

The one-stage and two-stage models are two popular strategies used in SRL tasks, as described by Sui et al. (2009). The word and the chunk are regarded as the two different tagging units of the SRL task.

In our SRL model, we consider BI and SRC as two stages, and the word is always used as the tagging unit. The detailed formalization is addressed in the following subsections.

3.2 BI

The aim of the BI stage is to identify all word spans of the semantic roles in one Chinese sentence. It can be regarded as a sequential tagging problem. Using the IOB2 strategy (Erik et al., 1999), we use the tag set $\{B,I,O\}$ to tag all words, where tag "B" represents the beginning word of a chunk, "I" denotes other tokens in the chunk, and "O" is the tag of all tokens outside any chunks. Therefore, the example sentence (a) can be represented as follows:

(b). 第 1$|_B$ 章 $|_I$ 介绍 $|_O$ 算法 $|_B$ 与 $|_I$ 数据 $|_I$ 结构 $|_I$ ； $|_O$

To avoid the problem of data sparsity, we use all sentences in our train data set to train the model of BI.

3.3 SRC

After predicting the boundaries of semantic role chunks in a sentence, the proper semantic role types should be assigned in the SRC step. Although it can be easily modeled as a classification problem, we regarded it as a sequential tagging problem at the word level. An additional constraint is employed in this step: the boundary tags of the predicting sequence of this stage should be consistent with the the output of the BI stage.

One intuitive reason for this model strategy is that the SRC step can use the same feature set as BI, and it can further prove the rationality of our feature optimization method.

3.4 Postprocessing

Not all predicted IOB2 sequences can be transformed to the original sentence correctly; therefore, they should satisfy the following compulsory constraints.

(1) The tagging sequence should be regular. "I...", "... OI...", "I-X...", "... O-I-X...", "... B-X-I-Y...", and "B-I-X-I-X-I-Y..." are not the regular IOB2 sequences.

(2) The tag for the target word must be "O".

We use the Algorithm 1 to justify whether the IOB2 sequences are regular.

Moreover, at the SRC stage, the boundary tags of the IOB2 sequence must be consistent with the given boundary tags.

For the BI stage, we firstly add an additional chunk type tag X to all "B" and "I" tags in the IOB2 sequences, and then use Algorithm 1 to justify the regularity of the sequences.

In the testing stage of the SRL model, we use the regular sequence with the max probability as the optimal output.

Algorithm 1. justify the regular IOB2 sequence

Input: (1) IOB2 sequence:$S = (s_1, .., s_n)$
where $s_i \in \{B-X, I-X, O\}$, and $1 \leq i \leq n$
(2) The position of target word in sentence pt

1, Initialization:
(1) Current chunk type: $ct = NULL$;
(2) Regularity of sequence: $state =' REG'$;
2, Check the tag of target word: s_{pt}**:**
(1) If $s_{pt} ==' O'$: go to Step 3;
(2) If $s_{pt} <>' O'$: $state =' IRR'$, and go to Step 4;
3,$For(i = 1; i \leq n; i++)$
(1) If $s_i ==' B-X'$: $ct =' X'$;
(2) If $s_i ==' I-X'$ and $ct <>' X'$: $state =' IRR'$, and go to Step 4;
(3) If $s_i ==' O'$: $ct = NULL$;
4, Stop

Output: Variable $state$;

3.5 Why Word-by-word?

We ever tried to use the methods of constituent-by-constituent and chunk-by-chunk to solve our SRL task on CFN, but the experiment results illustrate that they are not suitable to our task.

We use the Stanford Chinese full parser to parse all sentences in the CFN corpus and use the SRL model proposed by Xue et al.(2008) in our task. However, the results is insignificant. Only 66.72% of semantic roles are aligned with the constituents of the full parse tree, and the F-measure of BI only achieves 52.43%. The accuracy of the state-of-the-art Chinese full parser is not high enough, so it is not suitable to our SRL task.

Chunk-by-chunk is another choice for our task. When We use base chunk as the tagging unit of our model, only about 15% of semantic roles did not align very well with the boundary of automatically generated base chunks, and the F-measure is significantly lower than the method of word-by-word, as described by Wang et al.(2009).

Therefore, words are chosen as the tagging unit of our SRL model, which showed significant results from the experiment.

4 Feature Selection and Optimization

Word-level features and base-chunk features are used in our SRL research.

Base chunk is a Chinese shallow parsing scheme proposed by Professor Zhou. He constructed a high accuracy rule-based Chinese base chunk parse (Zhou, 2009), the F-measure of which can achieve 89%. We use this parse to generate all base chunks of the sentences in our corpus and to extract several types of features from them. The automatically generated base chunks of example sentences (a) are given as follows:

(c).[mp-ZX 第 1/m 章/q] [vp-SG 介绍/v] [np-SG 算法/n] 与/c [np-AM 数据/n 结构/n] ； /w

4.1 Candidate Feature Set

Three types of features are given as follows:

Features at the word level:

Word: The current token itself;

Part-of-Speech: The part of speech of the current token;

Position: The position of the current word relative to the target word(before, after, or on);

Target word: The target word in the sentence;

Features at the base chunk level:

Syntactic label: The syntactic label of the current token, such as, *B-np*,*I-vp*, etc;

Structural label: The structural label of the current token, such as, *B-SG*, *I-ZX*, etc;

Head word and its Part of Speech: The head word and its part of speech of the base chunk;

Shallow syntactic path: The combination of the syntactic tags from the source base chunk, which contains the current word, to the target base chunk, which contains the target word of the sentence;

Subcategory: The combination of the syntactic tags of the base chunk around the target word;

Other Features:

Named entity: The three types of named entities are considered: person, location, and time. They can be directly mapped from the part of speech of the current word.

Simplified sentence: A boolean feature. We use the punctuation count of the sentence to estimate whether the sentence is the simplified sentence.

Aside from the basic features described above, we also use combinations of these features, such as *word/POS* combination, etc.

4.2 Feature Optimization Method

In the baseline model, we only introduce the features at the word level. Table 1 shows the candidate features of our baseline model and proposes their optional sizes of sliding windows.

For Table 1, we use the orthogonal array $L_{32}(4^9 \times 2^4)$ to conduct 32 different templates.

The best template is chosen from the highest F-measure for testing the 32 templates. The detailed orthogonal-array-based feature selection method was proposed by Li et al.(2010).

Table 1. Candidate features of baseline models

Feature type	Window size			
word	[0,0]	[-1,1]	[-2,2]	[-3,3]
bigram of word	-	[-1,1]	[-2,2]	[-3,3]
POS	[0,0]	[-1,1]	[-2,2]	[-3,3]
bigram of POS	-	[-1,1]	[-2,2]	[-3,3]
position	[0,0]	[-1,1]	[-2,2]	[-3,3]
bigram of position	-	[-1,1]	[-2,2]	[-3,3]
word/POS	-	[0,0]	[-1,1]	[-2,2]
word/position	-	[0,0]	[-1,1]	[-2,2]
POS/position	-	[0,0]	[-1,1]	[-2,2]
trigram of position	-	[-2,0]	[-1,1]	[0,2]
word/target word	-	[0,0]		
target word	[0,0]			

Compared with the baseline model, the features at the word and base chunk levels are all considered in Table 2.

Table 2. Candidate features of the base chunk-based model

Feature type	Window size		
word	[0,0]	[-1,1]	[-2,2]
bigram of word	-	[-1,1]	[-2,2]
POS	[0,0]	[-1,1]	[-2,2]
bigram of POS	-	[-1,1]	[-2,2]
position	[0,0]	[-1,1]	[-2,2]
bigram of position	-	[-1,1]	[-2,2]
word/POS	-	[0,0]	[-1,1]
word/position	-	[0,0]	[-1,1]
POS/position	-	[0,0]	[-1,1]
trigram of position	-	[-2,0]	[-1,1]
syntactic label	[0,0]	[-1,1]	[-2,2]
syn-bigram	-	[-1,1]	[-2,2]
Syn-trigram	-	[-1,1]	[-2,2]
head word	[0,0]	[-1,1]	[-2,2]
head word-bigram	-	[-1,1]	[-2,2]
POS of Head	[0,0]	[-1,1]	[-2,2]
POS-bigram of head	-	[-1,1]	[-2,2]
syn/head word	[0,0]	[-1,1]	[-2,2]
stru/head word	[0,0]	[-1,1]	[-2,2]
shallow path	-	[0,0]	[-1,1]
subcategory	-	[0,0]	[0,0]
named Entity	-	[0,0]	[0,0]
simplified Sentence	-	[0,0]	[0,0]
target word(compulsory)	[0,0]		

The orthogonal array $L_{54}(2^1 \times 3^{25})$ is employed to select the best feature template from all candidate feature templates in Table 2. To distinguish it from the baseline model, we call the model based on the table 2 as the "base chunk-based SRL model".

For both feature sets described above, the target word is the compulsory feature in every template, and the boundary tags are introduced as features during the SRC stage.

The feature templates in Table 2 cannot contain the best feature template selected from Table 1. This is a disadvantage of our feature selection method.

5 Experimental Setup and Evaluation Metrics

5.1 Data Set

The experimental data set consists of all sentences of 25 frames selected in the CFN corpus. These sentences have the correct POS tags and CFN semantic information; they are all auto parsed by the rule-based Chinese base chunk parser. Table 3 shows some statistics on these 25 frames.

Table 3. Summary of the experimental data set

Frame	FEs	Sents	Frame	FEs	Sents
感受	6	569	因果	7	140
知觉特征	5	345	陈述	10	1,603
思想	3	141	拥有	4	170
联想	5	185	适宜性	4	70
自主感知	14	499	发明	12	198
查看	9	320	计划	6	90
思考	8	283	代表	7	80
非自主感知	13	379	范畴化	11	125
获知	9	258	证明	9	101
相信	8	218	鲜明性	9	260
记忆	12	298	外观	10	106
包含	6	126	属于某类	8	74
宗教信仰	5	54	**Totals**	**200**	**6,692**

5.2 Cross-validation technique

In all our experiments, three groups of two-fold cross-validation sets are used to estimate the performance of our SRL model. All sentences in a frame are cut four-fold on average, where every two folder are merged as train data, and the other two folds are used as test data. Therefore, we can obtain three groups of two-fold cross-validation data sets.

Estimating the parameter of fold number is one of the most difficult problems in the cross-validation technique. We believe that in the task of SRL, the two-fold cross validation set is a reasonable choice, especially when the data set is relative small. With a small data set, dividing it in half is split of data set is the best approximation of the real-world data distribution of semantic roles and the sparse word tokens.

5.3 Classifiers

CRFs model is used as the learning algorithm in our experiments. Previous SRL research has demonstrated that CRFs model is one of the best statistical algorithms for SRL, such as the works of Cohn et al. (2005) and Yu et al. (2007).

The crfpp toolkit[1] is a good implementation of the CRF classifier, which contains three different training algorithms: CRFL1, CRFL2, and MIRA. We only use CRFL2 with Gaussian priori regularization and the variance parameter C=1.0.

5.4 Evaluation Metrics

As described in SRL reseach, precision, recall, and F-measure are also used as our evaluation metrics. In addition, the standard deviation of the F-measure is also adopted as an important metric of our SRL model. The computation method of these metrics is given as follows:

Let P_j^i, R_j^i and F_j^i be the precision, recall, and F-measure of the jth group of the ith cross validation set, where $j = 1, 2, 3$ and $i = 1, 2$. The final precision(P), recall(R), and F-measure(F) of our SRL model are the expectation values of the P_j^i, R_j^i, and F_j^i, respectively.

The estimation of the variance of cross-validation is another difficult problem in the cross-validation technique. Although it has been proven that the uniform and unbiased estimation of the variance of cross-validation does not exist (Yoshua et al., 2007), we adopted the method proposed by Nadeau et al. (2007), to estimate the variance of the F-measure of cross-validation sets. This method is proposed hereinafter.

Let F_j be the average F-measure of the j group experiment, that is, $F_j = \frac{1}{2}(F_j^1 + F_j^2)$, where $j = 1, 2, 3$. The proposed estimator of the variance of F_j in the work of Nadeau et al. (2007) is as follows:

$$\begin{aligned}\widehat{Var}(F_j) &= (\frac{1}{K} + \frac{n_2}{n_1})\sum_{i=1}^{2}(F_j^i - F_j) \\ &= (\frac{1}{2} + 1)\sum_{i=1}^{2}(F_j^i - F_j)\end{aligned}$$

where, K is the fold number of cross-validation and n_1 and n_2 are the counts of training examples and testing examples. In our experimental setting, $K = 2$ and $\frac{n_2}{n_1} \approx 1$. Moreover, the estimation of the variance of the total F-measure is as follows:

$$\begin{aligned}Var(F) &= Var(\frac{1}{3}(F_1 + F_2 + F_3)) \\ &= \frac{1}{9}\sum_{j=1}^{3} Var(F_j)\end{aligned}$$

Using $\widehat{Var}(F_j)$ to estimate $Var(F_j)$, we can obtain:

$$\begin{aligned}\widehat{Var}(F) &= \frac{1}{9}\sum_{j=1}^{3}\widehat{Var}(F_j) \\ &= \frac{1}{6}\sum_{j=1}^{3}\sum_{i=1}^{2}(F_j^i - F_j)\end{aligned}$$

Finally, we can derive the standard deviation of the F-measure, that is, $std(F) = \sqrt{\widehat{Var}(F)}$.

5.5 Significance Test of Two SRL Models

To test the significance of SRL models A and B, we use the following statistics S.

$$S = \frac{F(A) - F(B)}{\sqrt{Var(F(A)) + Var(F(B))}} \sim t(n)$$

where $F(A)$ and $F(B)$ are the F-measures of models A and B, and n is the freedom degree of t-distribution, an integer nearest to the n'.

$$n' = \frac{3(Var(F(A)) + Var(F(B)))^2}{(Var(F(A))^2 + Var(F(B))^2)}$$

We use the $p - value(\cdot)$ to test the significance of SRL models A and B, which are given as follows:

$$p - value(F(A), F(B)) = P(S \geq t_{1-\alpha/2}(n))$$

If $p - value(F(A), F(B)) \leq 0.05$, the difference of the F-measures between models A and B is significant at 95% level.

[1] crfpp toolkit: http://crfpp.sourceforge.net/

6 Experimental Results and Discussion

We summarized the experiment results of every stage of our SRL model, that is, BI, SRC and a combination of these two steps (BI+SRC).

6.1 Baseline SRL Model

The results of the baseline model are given in Table 4, which only uses the features in Table 1.

Table 4. Results of the baseline model

	P(%)	R(%)	F(%)	std(F)
BI	74.42	66.80	70.40	0.0031
SRC	-	-	80.32	0.0032
BI+SRC	62.87	56.44	59.48	0.0050

In Table 1, because the results of the SRC stage are based on human-corrected boundary information, the precision, recall, and F-measure of this stage are the same. Therefore, we only give the F-measure and its deviation at the SRC stage.

In the baseline model, the BI stage is the bottleneck of our SRL model. Its F-measure only achieves 70.4%, and the recall is lower than the precision. Moreover, the F-measure of the final model only achieves 59.48%, and its standard deviation is larger than both stages.

6.2 Base chunk-based SRL Model

When base chunk features, proposed in Table 2, are employed in the SRL model, we can obtain the results summarized in Table 5.

Table 5. Results of the base chunk-based model

	P(%)	R(%)	F(%)	std(F)
BI	74.69	66.85	70.55	0.0038
SRC	-	-	81.00	0.0029
BI+SRC	63.97	57.25	60.42	0.0049

A comparison of Table 4 and Table 5 provides the following two conclusions.

(1) When base chunk features are used, all P, R, F at every stage slightly increase ($< 1\%$).

(2) The significance test values between the baseline model and the base chunk-based model are given in Table 6. For every stage, the performance boost after introducing the base chunk features is not significant at 95% level. However, the impact of base chunk features at the SRC stage is larger than that at the BI stage.

Table 6. Test values between two SRL models

	BI	SRC	BI+SRC
$p - value$	0.77	0.166	0.228

7 Conclusions and Further Directions

The SRL of Chinese predicates is a challenging task. In this paper, we studied the task of SRL on the CFN. We proposed a two-stage model and exploited the CRFs classifier to implement the automatic SRL systems. Moreover, we introduced the base chunk features and the OA-based method to improve the performance of our model. Experimental results shows that the F-measure of our best model achieves 60.42%, and the base chunk features cannot improve the SRL model significantly.

In the future, we plan to introduce unlabeled data into the training phase and use the EM-schemed semi-supervised learning algorithms to boost the accuracy of our SRL model.

Acknowledgement

The authors would like to thank Prof. Kaiying LIU for his comments and Prof. Qiang ZHOU for the base-chunk parser.

References

Baker, C., Fillmore, C., and John B. 1998. The Berkeley Framenet project. *In Proceedings of COLING-ACL*, 86-90, Montreal, Canada.

Baker, C., Ellsworth, M., Erk, K. 2007. SemEval'07 Task 19: Frame semantic structure extraction. *Proceedings of the 4th International Workshop on Semantic Evaluations*, 99-104, Prague, Czech Republic.

Cohn, T., Blunsom P. 2005. Semantic role labeling with tree conditional random fields. *Proceedings of CoNLL 2005*, ACL, 169-172.

Erik F., and John V. 1999. Representing text chunks. *In Proceedings of EACL'99*, 173-179.

Fillmore, C. 1982. Frame Semantics. *In The Linguistic Society of Korea*, Seoul: Hanshin.

Gildea, D., and Jurafsky, D. 2002. Automatic labeling for semantic roles. *Computational Linguistics*, 28(3):245-288.

Hajic, J., Ciaramita, M., Johansson, R., Kawahara, D., Marti, M., Màrquez, L., Meyers, A., Nivre, J., Padó, S., Stěpánek, J., Stranak, P., Surdeanu, M., Nianwen X., Zhang, Y. 2009. The CoNLL-2009 shared task: syntactic and semantic dependencies in multiple languages. *In Proceedings of CoNLL 2009*, 1-18, Boulder, CO, USA..

Jihong, L., Ruibo, W., Weilin, W., and Guochen, L. 2010. Automatic Labeling of Semantic Roles on Chinese FrameNet. *Journal of Software*, 2010, 21(4):597-611.

Jiangde, Y., Xiaozhong, F., Wenbo, P., and Zhengtao, Y. 2007. Semantic role labeling based on conditional random fields *Journal of southeast university (English edition)*, 23(2):5361-364.

Liping, Y., and Kaiying, L. 2005. Building Chinese FrameNet database. *In Proceedings of IEEE NLP-KE'05* , 301-306.

Litkowski, K. 2004. Senseval-3 task automatic labeling of semantic roles. *Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, 9-12, Barcelona, Spain.

Màrquez, L., Carreras, X., Litkowski, K., Stevenson, S. 2008. Semantic Role Labeling: An Introduction to the Special Issue. *Computational Linguistics*, 34(2):145-159.

Meyers, A., Reeves, R., Macleod, C., Szekely, R., Zielinska, V., Young, B., and Grishman, R. 2004. The NomBank Project: An interim report. *In Proceedings of the NAACL/HLT Workshop on Frontiers in Corpus Annotation*, 24-31, Boston, MA, USA.

Nadeau, C., and Bengio, Y. 2003. Inference for the generalization error. *Machine Learning*, 52: 239-281.

Nianwen, X. 2008. Labeling Chinese Predicates with Semantic Roles. *Computational Linguistics*, 2008, 34(2): 225-255.

Paul, K., and Martha, P. 2002. From TreeBank to PropBank. *In Proceedings of LREC-2002*, Canary Islands, Spain.

Pradhan, S., Hacioglu, K., Krugler, V., Ward, W., Martin, J., Jurafsky, D. 2005. Support vector learning for semantic argument classification. *Machine Learning*, 2005, 60(1):11-39.

Qiang, Z. 2007. A rule-based Chinese base chunk parser. *In Proc. of 7th International Conference of Chinese Computation (ICCC-2007)*, 137-142, Wuhan, China.

Ruibo, W. 2004. Automatic Semantic Role Labeling of Chinese FrameNet Based On Conditional Random Fields Model. *Thesis for the 2009 Master's Degree of Shanxi University*, Taiyuan, Shanxi, China.

Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L., Nivre, J. 2008. The CoNLL 2008 shared task on joint parsing of syntactic and semantic dependencies. *In Proceedings of CoNLL 2008*, 159-177, Manchester, England, UK.

Surdeanu, M., Màrquez, L., Carreras, X., Comas, P. 2007. Combination strategies for semantic role labeling. *Journal of Artificial Intelligence Research*, 29:105-151.

Yoshua, B., and Yves, G. 2004. No unbiased estimator of the variance of K-fold cross-validation *Journal of Machine Learning Research*, 5:1089-1105.

Weiwei, S., Zhifang, S., Meng, W., and Xing, W. 2009. Chinese semantic role labeling with shallow parsing. *In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing(EMNLP 2009)*, ACL, 1475-1483.