Improving Chinese Semantic Role Classification with Hierarchical Feature Selection Strategy

Weiwei Ding
Institute of Computational Linguistics
Peking University
Beijing, 100871, China
weiwei.ding.pku@gmail.com

Baobao Chang
Institute of Computational Linguistics
Peking University
Beijing, 100871, China
chbb@pku.edu.cn

Abstract

In recent years, with the development of Chinese semantically annotated corpus, such as Chinese Proposition Bank and Normalization Bank, the Chinese semantic role labeling (SRL) task has been boosted. Similar to English, the Chinese SRL can be divided into two tasks: semantic role identification (SRI) and classification (SRC). Many features were introduced into these tasks and promising results were achieved. In this paper, we mainly focus on the second task: SRC. After exploiting the linguistic discrepancy between numbered arguments and ARGMs, we built a semantic role classifier based on a hierarchical feature selection strategy. Different from the previous SRC systems, we divided SRC into three sub tasks in sequence and trained models for each sub task. Under the hierarchical architecture, each argument should first be determined whether it is a numbered argument or an ARGM, and then be classified into fine-gained categories. Finally, we integrated the idea of exploiting argument interdependence into our system and further improved the performance. With the novel method, the classification precision of our system is 94.68%, which outperforms the strong baseline significantly. It is also the state-of-the-art on Chinese SRC.

1 Introduction

Semantic Role labeling (SRL) was first defined in Gildea and Jurafsky (2002). The purpose of SRL task is to identify and classify the semantic roles of each predicate in a sentence. The semantic roles are marked and each of them is assigned a tag which indicates the type of the semantic relation with the related predicate. Typical tags include Agent, Patient, Source, etc. and some adjuncts such as Temporal, Manner, Extent, etc. Since the arguments can provide useful semantic information, the SRL is crucial to many natural language processing tasks, such as Question and Answering (Narayanan and Harabagiu 2004), Information Extraction (Surdeanu et al. 2003), and Machine Translation(Boas 2002). With the efforts of many researchers (Carreras and Màrquez 2004, 2005, Moschitti 2004, Pradhan et al 2005, Zhang et al 2007), different machine learning methods and linguistics resources are applied in this task, which has made SRL task progress fast.

Compared to the research on English, the research on Chinese SRL is still in its infancy stage. Previous work on Chinese SRL mainly focused on how to transplant the machine learning methods which has been successful with English, such as Sun and Jurafsky (2004), Xue and Palmer (2005) and Xue (2008). Sun and Jurafsky (2004) did the preliminary work on Chinese SRL without any large semantically annotated corpus of Chinese. They just labeled the predicate-argument structures of ten specified verbs to a small collection of Chinese sentences, and used Support Vector Machines to identify and classify the arguments. This paper made the first attempt on Chinese SRL and produced promising results. After the PropBank (Xue and Palmer 2003) was built, Xue and Palmer (2005) and Xue (2008) have produced more complete and systematic research on Chinese SRL.

Moschitti et al. (2005) has made some preliminary attempt on the idea of hierarchical semantic

Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing, pages 324–333,
Honolulu, October 2008.

role labeling. However, without considerations on how to utilize the characteristics of linguistically similar semantic roles, the purpose of the hierarchical system is to simplify the classification process to make it less time consuming. So the hierarchical system in their paper performs a little worse than the traditional SRL systems, although it is more efficient.

Xue and Palmer (2004) did very encouraging work on the feature calibration of semantic role labeling. They found out that different features suited for different sub tasks of SRL, i.e. semantic role identification and classification. For semantic analysis, developing features that capture the right kind of information is crucial. Experiments on Chinese SRL (Xue and Palmer 2005, Xue 2008) reassured these findings.

In this paper, we mainly focus on the semantic role classification (SRC) process. With the findings about the linguistic discrepancy of different semantic role groups, we try to build a 2-step semantic role classifier with hierarchical feature selection strategy. That means, for different sub tasks, different models will be trained with different features. The purpose of this strategy is to capture the right kind of information of different semantic role groups. It is hard to do manual selection of features since there are too many feature templates which has been proven to be useful in SRC; so, we designed a simple feature selection algorithm to select useful features automatically from a large set of feature templates. With this hierarchical feature selection architecture, our system can outperform previous systems. The selected feature templates for each process of SRC can in turn reassure the existence of the linguistic discrepancy. At last, we also integrate the idea of exploiting argument interdependence to make our system perform better.

The rest of the paper is organized as follows. In section 2, the semantically annotated corpus - Chinese Propbank is discussed. The architecture of our method is described in section 3. The feature selection strategy is discussed in section 4. The settings of experiments can be found in section 5. The results of the experiments can be found in section 6, where we will try to make some linguistic explanations of the selected features. Section 7 is conclusions and future work.

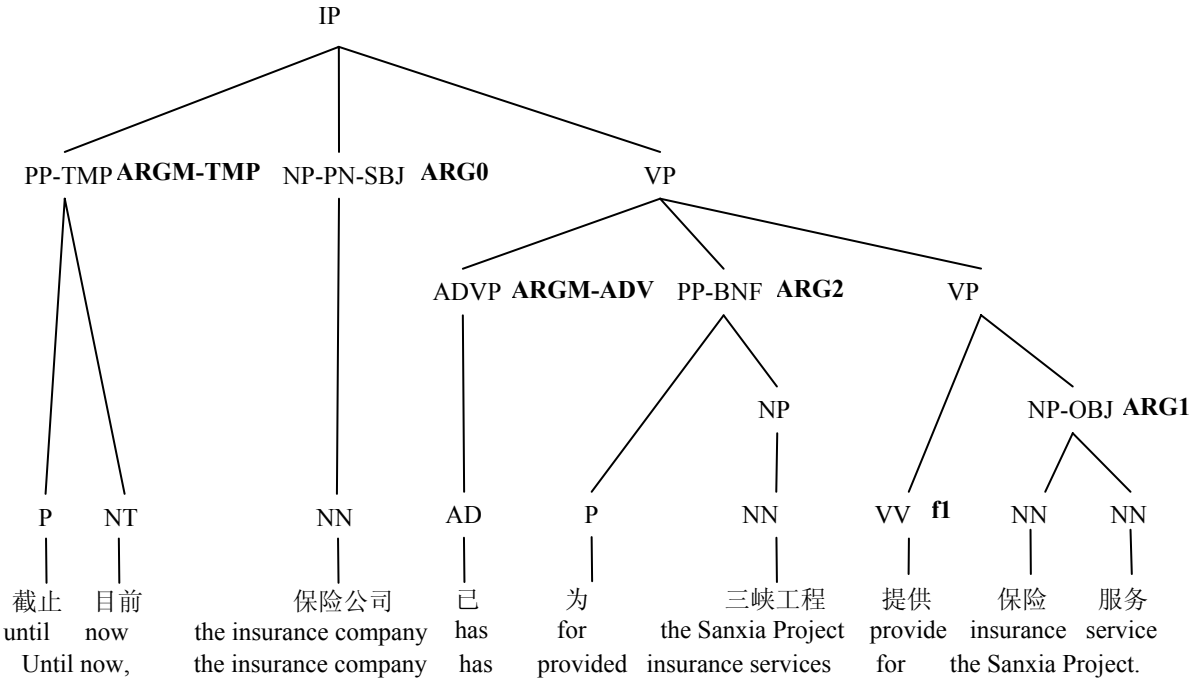

Figure 1. an example from PropBank

2 The Chinese PropBank

The Chinese PropBank has labeled the predicate-argument structures of sentences from the Chinese TreeBank (Xue et al. 2005). It is constituted of two parts. One is the labeled data, which indicates the positions of the predicates and its arguments in the Chinese Treebank. The other is a dictionary which

lists the frames of all the labeled predicates. Figure 1 is an example from the PropBank[1]. We put the word-by-word translation and the translation of the whole sentence below the example.

It is quite a complex sentence, as there are many semantic roles in it. In this sentence, all the semantic roles of the verb 提供 (provide) are presented in the syntactic tree. We can separate the semantic roles into two groups.

The first group of semantic roles can be called the core arguments, which capture the core relations. In this sentence, there are three arguments of verb 提供 (provide) in this sentence. 保险公司 (the insurance company) is labeled as ARG0, which is the proto-agent of the verb. Specifically to the verb 提供 (provide), it is the provider. 保险服务 (insurance services) is the direct object of the verb, and it is the proto-patient, which is labeled as ARG1. Specifically to the verb 提供 (provide), it represents things provided. 为三峡工程 (for the Sanxia Project) is another kind of argument, which is labeled as ARG1, and it represents the receiver.

The other group of semantic roles is called adjuncts. They are always used to reveal the peripheral information. There are two adjuncts of the target verb in this sentence: 截止目前 (until recently) and 已 (has), both of which are labeled as ARGM. However, the two ARGMs reveal information of different aspects. Besides the ARGM tags, the secondary tags "TMP" and "ADV" are assigned to the two semantic roles respectively. "TMP" indicates that 截止目前 (until recently) is a modifier representing the temporal information, and "ADV" indicates that 已 (has) is an adverbial modifier.

In the Chinese PropBank, the difference of the two groups is obvious. The core arguments are all labeled with numbers, and they are also called the numbered arguments. The numbers range from 0 to 4 in Chinese PropBank. The adjuncts are labeled with "ARGM".

3 Building a Hierarchical Semantic Role Classifier

In this section, we will discuss the linguistic fundaments of the construction of a hierarchical semantic role classifier. We use "hierarchical" to distinguish our classifier from the previous "flat" ones.

3.1 Linguistic Discrepancy of Different Semantic Role Groups

The purpose of the SRC task is to assign a tag to all the semantic roles which have been identified. The tags include ARG0-4, and 17 kinds of ARGMs (with functional tags). Previous SRC systems treat all the tags equally, and view the SRC as a multi-category classification task. However, we have different opinions of the traditional architecture.

Due to the discussions in section 2, we noticed that the semantic roles can be divided into two groups naturally according to the different kinds of semantic information represented by them. Here we will make some linguistic analysis of the two semantic role groups. Conversely to the process of the syntactic realization of semantic roles, we want to find out what linguistic features make a constituent ARG0 instead of ARG1, or another constituent ARGM-TMP instead of ARGM-ADV, i.e. what features capture the most crucial information of the two groups.

As what we have assumed, the linguistic features which made a syntactic constituent labeled as either one of the core arguments or one of the adjuncts varies greatly. Take the sentence in section 2 as an example, even if the only information we have about the phrase 截止目前 (until now) is that it is an adjunct of the verb, we can almost confirm, no matter where this node takes place in the parsing tree, this constituent will be labeled as ARGM-TMP. 已 (has) is also the same. According to its meaning, the only category can be assigned to it is ARGM-ADV. But, things are quite different to the core argument. In the same sentence, 保险公司 (the insurance company) is a good example. If we limit our observation to the phrase itself, we can hardly assert that it is the ARG0 of the target verb. Only when we extend our observation to the syntactic structure level, find out it is the subject of this sentence, and the voice of the sentence is active, the semantic type of 保险公司 (the insurance company) is finally confirmed. If we have another sentence in which 保险公司 (the insurance company) is not the subject, but rather the object, and the target verb is 开办 (set up), then it will probably be labeled as ARG1.

[1] This sentence is extracted from chtb_082.fid of Chinese PropBank 1.0, and we made some simplifications on it.

Due to the analysis above, we can conclude the linguistic discrepancy of the two semantic role groups as follows. Core arguments and adjuncts share different kinds of inner linguistic consistency respectively. For the core arguments, the specific type cannot be determined with the information of the arguments only. At this level, the core arguments are dependent on other information except the information about themselves. For example, the information of syntactic structures is crucial to the determination of the types of core arguments, and trivial differences of the syntactic structures will lead to the different output. Because of this, we can say that the core arguments are sensitive to the syntactic structures. Compared to the core arguments, adjuncts are the opposite. They are relatively independent on other information, since most of the adjuncts can be easily classified just based on the information about themselves[2]. And although the positions of the adjuncts in the syntactic structure can vary, the types of the adjuncts are fixed. In this sense, the adjuncts are insensitive to the syntactic structures.

After we made the linguistic discrepancy of the two semantic role groups, we can make a bold assumption that the differences of the two groups can be reflected in the capability of different kinds of features to capture the crucial information for the two groups. For example, the "voice" features seems to be crucial to the core arguments but useless to the adjuncts. This assumption provided us with the idea of a hierarchical feature selection system.

In this system, we first classify the constituents into two classes: core arguments and adjuncts. And then, the system classifies core arguments and adjuncts separately. For different subtasks we only select the most useful features and discard the less pertinent ones. We hope to take utilization of the most crucial features to improve semantic role classification.

3.2 System Architecture

Previous semantic role classifiers always did the classification problem in one-step. However, in this paper, we did SRC in two steps. The architectures of hierarchical semantic role classifiers can be found in figure 2, which is similar with that in Moschitti et al. (2005).

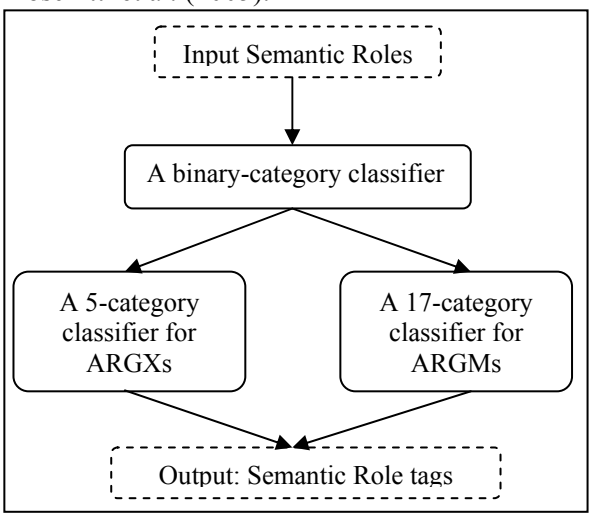

Figure 2. The architecture of our hierarchical SRC system

As what has been shown in figure 2, a semantic role will first be determined whether it is a numbered argument or an ARGM by a binary-category classifier. And, then if the semantic role is a numbered argument, it will be determined by a 5-category classifier designed for ARGX, i.e. the numbered arguments. If it is an ARGM, the functional tag will be assigned by a 17-category classifier built for ARGMs. Accordingly, with this hierarchical architecture, the SRC problem is divided into 3 sub tasks, each of which has an independent classifier.

3.3 Integrating the Idea of Exploiting Argument Interdependence

Jiang et al. (2005) has built a semantic role classifier exploiting the interdependence of semantic roles. It has turned the single point classification problem into the sequence labeling problem with the introduction of semantic context features. Semantic context features indicates the features extracted from the arguments around the current one. We can use window size to represent the scope of the context. Window size [-m, n] means that, in the sequence that all the arguments has constructed, the features of previous m and following n arguments will be utilized for the classification of current semantic role. There are two kinds of argument sequences in Jiang et al. (2005), and we only test the linear sequence. Take the sentence in figure 1 as an example. The linear sequence of the arguments in this sentence is: 截止目前(until then),

[2] Extra features e.g. predicate may be still useful because that the information, provided by the high-level description of self-descriptive features, e.g. phrase type, are limited.

保险公司 (the insurance company), 已 (has), 为三峡工程 (for the Sanxia Project), 保险服务 (insurance services). For the argument 已 (has), if the semantic context window size is [-1,2], the semantic context features e.g. headword, phrase type and etc. of 保险公司 (the insurance company), 为三峡工程 (for the Sanxia Project) and 保险服务 (insurance services) will be utilized to serve the classification task of 已 (has).

While their paper has improved the SRC performance on English, it also has one potential disadvantage, which is that they didn't separate the core arguments and ARGMs. The influence and explanations of this defect are presented in Section 6. But in our hierarchical system, this problem can be solved. Since in the first step, we have separated the numbered arguments and ARGMs. We suppose that with the separation of the two semantic role groups, the system performance will be further improved.

4 Feature Selection Strategy

Due to what we have discussed in the section 3.1, we need to select different features for the three sub task of SRC. In this paper, we did not make the selection manually; however, we make a simple greedy strategy for feature selection to do it automatically. Although the best solution may not be found, automatic selection of features can try far more combinations of feature templates than manual selection. Because of this, this strategy possibly can produce a better local optional solution.

First, we built a pool of feature templates which has proven to be useful on the SRC. Most of the feature templates are standard, so only the new ones will be explained. The candidate feature templates include:

Voice from Sun and Jurafsky (2004).

Head word POS, Head Word of Prepositional Phrases, Constituent tree distance, from Pradhan etc. (2004).

Position, subcat frame, phrase type, first word, last word, subcat frame+, predicate, path, head word and its POS, predicate + head word, predicate + phrase type, path to BA and BEI, verb class[3]*, verb class + head word, verb class + phrase type,* from Xue (2008).

predicate POS, first word + last word, phrase type of the sibling to the left, phrase type of the sibling to the right, verb + subcate frame+, verb POS + subcat frame+, the amount of VPs in path, phrase type + phrase type of parent node, which can be easily understood by name.

voice position, indicates whether the voice marker (BA, BEI) is before or after the constituent in focus.

subcat frame,* the rule that expands the parent node of the constituent in focus.

subcat frame@, the rule that expands the constituent in focus.

layer of the constituent in focus, the number of constituents in the ascending part of the path subtracted by the number of those in the descending part of path, e.g. if the path is PP-BNF ↑ VP ↓ VP ↓ VV, the feature extracted by this template will be -1.

SemCat (semantic category) of predicate, SemCat of first word, SemCat of head word, SemCat of last word, SemCat of predicate + SemCat of first word, SemCat of predicate + SemCat of last word, predicate + SemCat of head word, SemCat of predicate + head word. The semantic categories of verbs and other words are extracted from the Semantic Knowledge-base of Contemporary Chinese (Wang et al. 2003).

verb AllFrameSets, the combination of all the framesets of a predicate.

verb class + verb AllFrameSets, verb AllFrameSets + head word, verb AllFrameSets + phrase type.

There are more than 40 feature templates, and it is quite difficult to traverse all the possible combinations and get the best one. So we use a greedy algorithm to get an approximate optimal solution.

The feature selection algorithm is as follows. Each time we choose one of the feature templates and add it into the system. The one, after which is added, the performance is the highest, will be chosen. Then we continue to choose feature templates until there are no one left. In the end, there are a series of feature sets, which recorded the process of feature selection. Then we choose the feature set which can perform the best on development set. The code of feature selection algorithm is designed in Figure 3.

[3] It is a bit different from Xue (2008), since we didn't use the syntactic alternation information.

1. add all feature templates to set **S** ,the set of selected feature templates C_0 is null
2. for i = 0 to n-1, n is the number of elements in **S**
3. P_i =0
4. for each feature template ft_j in set **S**
5. $C'_i = C_i + ft_j$
6. train a model with features extracted by C'_i and test on development set
7. if the result $P' > P_i$
8. $P_i = P'$, k= j
9. end for
10. $C_{i+1} = C_i + ft_k$
11. $S = S - ft_k$
12. end for
13. the set C_m correspondent to P_m, which is the highest, will be chosen.

Figure 3. the greedy feature selection algorithm

To make a comparison, we also built a traditional 1-step semantic role classifier based on this feature selection strategy. We will take this classifier as the baseline system.

5 Experiment Settings

5.1 Classifier

In our SRL system, we use a Maximum Entropy toolkit with tunable Gaussian Prior and L-BFGS parameter estimation, which is implemented by Zhang Le. This toolkit is available at `http://homepages.inf.ed.ac.uk/s0450736/maxent_toolkit.html`. It can well handle the multi-category classification problem and it is quite efficient.

5.2 Data

We use Chinese PropBank 1.0 (LDC number: LDC2005T23) in our experiments. PropBank 1.0 includes the annotations for files chtb_001.fid to chtb_931.fid, or the first 250K words of the Chinese TreeBank 5.1. For the experiments, the data of PropBank is divided into three parts. 648 files (from chtb_081 to chtb_899.fid) are used as the training set. The development set includes 40 files, from chtb_041.fid to chtb_080.fid. The test set includes 72 files, which are chtb_001 to chtb_041, and chtb_900 to chtb_931. We use the same data setting with Xue (2008), however a bit different from Xue and Palmer (2005).

6 Results and Discussion

The results of the feature selection are presented in table1. In this table, "Baseline" indicates the 1-step architecture, and "Hierarchical" indicates the "hierarchical feature selection architecture" implemented in this paper. "X_M", "ARGX" and "ARGM" indicate the three sub-procedures of the hierarchical architecture, which are "ARGX and ARGM separation", "ARGX classification", "ARGM classification" respectively. "Y" in the table indicates that the feature template has been selected for the sub task.

According to table 1, we can find some interesting facts, which in turn prove what we found about semantic role groups in section 3.1.

In table 1, feature templates related to the syntactic structure includes: voice-related group (voice, voice information, path to BA and BEI), frame-related group (verb class, verb class + head word, verb class + phrase type, all frames of verb, verb class + all frames of verb), the layer of argument, position and 4 kinds of subcat frames. As we assumed before, these features are crucial to core arguments but of little use to adjuncts. The results have proven this assumption. Of the entire 14 syntactic structure-related feature templates, 8 were selected by the ARGX process but only 2 was selected by the ARGM process. The two exceptions should be viewed as the result of random impact, which cannot be avoided in automatic feature selection.

Compared with the different features selected by these tasks, we can find other interesting results. Few of the features selected by the X_M process also have related with the verb or the syntactic structures, which is quite similar with the ARGM process. This is probably because most of ARGMs are easy to be identified without syntactic structure information, which makes the opposite of ARGMs, i.e. the ARGXs easy to be filtered. Besides, the features selected by the baseline system have much in common with those selected by the ARGX process. This can be explained by the fact that both in the development and test set, the amount of core arguments outperforms that of adjuncts. The proportions between core arguments and adjuncts are 1.79:1 on the development set, and 1.63:1 on the test set. Because of the bias, the baseline system will tend to choose more syntactic structure-related features to label core arguments precisely.

Baseline	Hierarchical			Feature Name
	X_M	ARGX	ARGM	
		Y		predicate
Y		Y		predicate POS
	Y		Y	first word
			Y	first word + last word
Y		Y		head word
Y				head word POS
Y	Y			phrase type
	Y		Y	phrase type + phrase type of parent node
			Y	phrase type of the sibling to the left
Y		Y		phrase type of the sibling to the right
Y	Y			position
		Y		voice
Y				voice position
Y		Y		path to BA and BEI
Y	Y	Y		verb class
			Y	verb class + head word
Y	Y			verb class + phrase type
Y		Y		verb AllFrameSets
Y		Y		verb class + verb AllFrameSets
		Y		subcat frame
	Y			subcat frame*
		Y		subcate frame@
			Y	subcat frame+
Y		Y		layer of the constituent in focus
	Y	Y	Y	predicate + head word
Y	Y	Y	Y	predicate + phrase type
Y	Y	Y		SemCat of predicate
Y				SemCat of first word
Y		Y		SemCat of last word
			Y	SemCat of predicate + SemCat of last word
Y		Y		SemCat of head word

Table 1. Feature selection results for the baseline and the hierarchical system

	Baseline	Hierarchical
DEV	95.15%	**95.94%**
TEST	93.38%	**94.31%**

Table 2. Comparison of the performance between the baseline and hierarchical system

With this new architecture, we have achieved improvement on the performance of the semantic role classification, which can be found in table 2. Our classifier performs better both on the development and the test set. The labeled precision on the development set is from 95.15% to 95.94%, and the test set is from 93.38% to 94.31%, with an ERR (error reduction rate) of 14.05%. Both of the improvements are statistically significant ($\chi 2$ test with p= 0.05). The errors of SRC have three origins, which are correspondent to the three sub tasks of the hierarchical architecture. Detailed information of the comparison between the two systems can be found in table 3, which can tell us where the improvements come from.

	Baseline	Hierarchical
ARGX/ARGM errors	1.66%	1.75%
inner ARGX errors	3.59%	2.75%
inner ARGM errors	1.37%	1.19%
TOTAL	6.62%	5.69%

Table 3 Error rate analysis on the test set

In table 3, the percentages are calculated the way that the number of the errors was divided by the number of the arguments in the test set. ARGX/ARGM errors represent the errors that the semantic roles are classified into wrong group, e.g. ARGXs are labeled as ARGMs and vice versa. The inner errors represent the errors in a group, e.g. ARG0 are labeled as ARG1. From this table, we can find that ARGX is the most difficult task. X-M and ARGM are less challenging. Besides the relatively little error reduction in the ARGM process, the greatest part of improvement comes from the process of the most difficult sub task: the ARGX sub task. It is a bit surprising that the first step of the X_M in the hierarchical system process did not perform better than that in the baseline system.

	Baseline	Hierarchical	Sum
ARG0	96.14%	96.58%	2046
ARG1	92.75%	94.60%	2428
ARG2	78.46%	78.85%	260
ARG3	60.00%	76.00%	25
ARG4	40.00%	100.00%	5
ARGM-ADV	96.64%	96.85%	1490
ARGM-ASP	100.00%	0.00%	1
ARGM-BNF	91.30%	86.96%	23
ARGM-CND	77.78%	77.78%	9
ARGM-CRD	N/A	N/A	0
ARGM-DGR	N/A	N/A	0
ARGM-DIR	54.84%	58.06%	31
ARGM-DIS	79.38%	79.38%	97
ARGM-EXT	50.00%	25.00%	8
ARGM-FRQ	N/A	N/A	0
ARGM-LOC	90.91%	92.21%	308
ARGM-MNR	89.92%	91.13%	248
ARGM-PRD	N/A	N/A	0
ARGM-PRP	97.83%	97.83%	46
ARGM-TMP	95.41%	96.30%	675
ARGM-TPC	33.33%	8.33%	12
TBERR[4]	0.00%	0.00%	2

Table 4 Detailed labeled precision on the test set

Table 4 presented the labeled precision of each type of semantic role. It demonstrates that with respect to ARGMs and ARGXs, the hierarchical system outperforms the baseline system. Furthermore, the improvement on ARGXs is greater than that of ARGMs. All types of numbered arguments get improvement in the hierarchical architecture, especially ARG1, ARG4 and ARG3. Although the performances of some types of the ARGMs decreased, the performances of all types of the ARGMs which occurs more than 100 times increased, including ADV (adverbials), LOC (locatives), MNR (manner markers) and TMP (temporal markers).

After the hierarchical system was built, we tried to integrate the idea of exploiting argument interdependence into our system. We extract the semantic context features in a linear order, with the window size from [0,0] to [-3,3]. Larger window sizes are of little value since too few arguments have more than 6 other arguments in context. The results are presented in table 5.

	Baseline	Hierarchical
Base	93.38%	94.31%
+window selection	93.38%	**94.68%**

Table 5 integrating window selection into our system

"Base" stands for the hierarchical system built above, without semantic context features. "+window selection" indicates the new system which has utilized the idea of exploiting argument interdependence. The best window sizes for the baseline system, ARGX and ARGM processes in the hierarchical system are [0,0], [-1,1], [0,0] respectively, which were determined by testing on the development set. We can find that only for the ARGX process, the semantic context features are useful. For the baseline system and the ARGM process, exploiting argument interdependence does not help improve the performance. This conclusion is different from Jiang et al. (2004), but it can be explained in the following way.

In fact, the interdependence only exists among core arguments. For ARGMs, it is a different thing. An ARGM cannot provide any information about the type of the arguments close to it and the semantic context features does not help the classification of ARGMs. Also, take the sentence in section 2 as an example, the fact that 截止目前 (until now) is ARGM-TMP cannot raise the probability that 保险公司 (the insurance company) is ARG0 or 已 (has) is ARGM-ADV and vice versa. However, if we know that 保险公司 (the insurance company) is ARG0, at least the phrase 保险服务 (insurance services) can never be ARG0. The semantic context features extracted from or for ARGMs will do

[4] From the name, TBERR possibly indicates the labeled errors in Chinese PropBank. However, we did not find any explanations, so we just put it here and group it to ARGM.

harm to the improvement of the system, since they are irrelative information. Because of the same reason, the performance of base system also decreased when semantic context features were extracted, since the core arguments and the ARGMs are mixed together in the baseline system.

But for the ARGX sub task of our hierarchical system, since we have separated the numbered arguments and ARGMs first, the influences of ARGMs can be eliminated. This made the interdependence of core arguments can be directly explored from the extraction of semantic context features. So the ARGX sub task is improved.

To prove that our method is effective, we also make a comparison between the performances of our system and Xue and Palmer (2005), Xue (2008). Xue (2008) is the best SRL system until now and it has the same data setting with ours. The results are presented in Table 6.

X & P (2005)	Xue(2008)	Ours
93.9%	94.1%	**94.68%**

Table 6. Comparison with previous systems

We have to point out that all the three systems are based on Gold standard parsing. From the table 6, we can find that our system is better than both of the related systems. Our system has outperformed Xue (2008) with a relative error reduction rate of 9.8%.

7 Conclusions and Future Work

In this paper, we have divided all the semantic roles into two groups according to their semantic relations with the verb. After the grouping of the semantic roles was made, we designed a hierarchical semantic role classifier. To capture the accurate information of different semantic role groups, we designed a simple feature selection algorithm to calibrate features for each sub task of SRC. It was very encouraging that the hierarchical SRC system outperformed the strong baseline built with traditional methods. And the selected features could be explained, which in turn proves that the linguistic discrepancy of semantic role groups not only exists but also can be captured. Then we integrated the idea of exploiting argument interdependence to further improve the performance of our system and explained linguistically why the results of our system were different from the ones in previous research.

Although we make discriminations of arguments and adjuncts, the analysis is still coarse-grained. Yi et al. (2007) has made the first attempt working on the single semantic role level to make further improvement. However, the impact of this idea is limited due to that the amount of the research target, ARG2, is few in PropBank. What if we could extend the idea of hierarchical architecture to the single semantic role level? Would that help the improvement of SRC?

Acknowledgements

This work was supported by National Natural Science Foundation of China under Grant No. 60303003 and National Social Science Foundation of China under Grant No. 06BYY048.

We want to thank Nianwen Xue, for his generous help at the beginning of this work. And thanks to the anonymous reviewers, for their valuable comments on this paper.

References

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of the 17th international conference on Computational linguistics*, Montreal, Canada.

Boas, Hans C. 2002. Bilingual FrameNet dictionaries for machine translation. In *Proceedings of LREC 2002*, Las Palmas, Spain.

Carreras, Xavier and Lluís Màrquez. 2004. Introduction to the conll-2004 shared task: Semantic role labeling. In *Proceedings of the Eighth Conference on Natural Language Learning,* Boston, Massachusetts.

Carreras, Xavier and Lluís Màrquez. 2005. Introduction to the conll-2005 shared task: Semantic role labeling. In *Proceedings of the Nineth Conference on Natural Language Learning,* Ann Arbor, Michigan.

Màrquez, Lluís, Xavier Carreras, Kenneth C. Litkowski, Suzanne Stevenson. 2008. Semantic Role Labeling: An Introduction to the Special Issue, *Computational linguistics.* 34(2):146-159..

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

Jiang, Zheng Ping, Jia Li, Hwee Tou Ng. 2005. Semantic Argument Classification Exploiting Argument Interdependence. In *19th International Joint Conference on Artificial Intelligence.* Edinburgh, Scotland.

Kingsbury, Paul and Martha Palmer. 2002. From TreeBank to PropBank. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*, Las Palmas, Spain.

Kipper, Karin, Hoa Trang Dang, and Martha Palmer. 2000. Class-based construction of a verb lexicon. In *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence*, Austin, Texas, USA.

Moschitti. Alessandro. 2004. A Study on Convolution Kernels for Shallow Statistic Parsing. In *Proceedings of the 42nd Meeting of the Association for Computational Linguistics*, Barcelona, Spain.

Moschitti, Alessandro, Ana-Maria Giuglea, Bonaventura Coppola, and Roberto Basili. 2005. *Hierarchical semantic role labeling*. In *Proceedings of the Nineth Conference on Natural Language Learning,* Ann Arbor, Michigan.

Narayanan, Srini and Sanda Harabagiu. 2004. Question answering based on semantic structures. In *Proceedings of the 20th International Conference on Computational Linguistics*, Geneva, Switzerland.

Pradhan, Sameer, Kadri Hacioglu, Valerie Kruglery, Wayne Ward, James H. Martin, Daniel Jurafskyz. 2004. Support vector learning for semantic argument classification. *Machine Learning Journal*, 60(1-3):11-39.

Sun, Honglin and Daniel Jurafsky. 2004. Shallow Semantic Parsing of Chinese. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, Boston, Massachusetts.

Surdeanu, Mihai, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, Ann Arbor, Michigan.

Wang, Hui, Weidong Zhan, Shiwen Yu. 2003. The Specification of The Semantic Knowledge-base of Contemporary Chinese, In *Journal of Chinese Language and Computing*, 13(2):159-176.

Xue, Nianwen and Martha Palmer. 2003. Annotating the Propositions in the Penn Chinese Treebank. In *Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing*, Sapporo, Japan.

Xue, Nianwen and Martha Palmer. 2004. Calibrating features for semantic role labeling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain.

Xue, Nianwen and Martha Palmer. 2005. Automatic semantic role labeling for Chinese verbs. In *19th International Joint Conference on Artificial Intelligence.* Edinburgh, Scotland.

Xue, Nianwen, Fei Xia, Fu dong Chiou, and Martha Palmer. 2005. The Penn Chinese TreeBank: Phrase Structure Annotation of a Large Corpus. Natural Language Engineering, 11(2):207–238.

Xue, Nianwen. 2008. Labeling Chinese predicates with semantic roles. *Computational linguistics.* 34(2):225-255.

Yi, Szu-ting, Edward Loper, Martha Palmer. 2007. Can Semantic Roles Generalize Across Genres? In *Proceedings of Human Language Technologies and The Conference of the North American Chapter of the Association for Computational Linguistics*, Rochester, NY, USA.

Zhang, Min, Wanxiang Che, Ai Ti Aw, Chew Lim Tan, Guodong Zhou, Ting Liu, Sheng Li. 2007. A Grammar-driven Convolution Tree Kernel for Semantic Role Classification, in *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, Prague, Czech Republic.