Multi-Predicate Semantic Role Labeling

Haitong Yang and Chengqing Zong
National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, Beijing, 100190, China
{htyang, cqzong}@nlpr.ia.ac.cn

Abstract

The current approaches to Semantic Role Labeling (SRL) usually perform role classification for each predicate separately and the interaction among individual predicate's role labeling is ignored if there is more than one predicate in a sentence. In this paper, we prove that different predicates in a sentence could help each other during SRL. In multi-predicate role labeling, there are mainly two key points: argument identification and role labeling of the arguments shared by multiple predicates. To address these issues, in the stage of argument identification, we propose novel predicate-related features which help remove many argument identification errors; in the stage of argument classification, we adopt a discriminative reranking approach to perform role classification of the shared arguments, in which a large set of global features are proposed. We conducted experiments on two standard benchmarks: Chinese PropBank and English PropBank. The experimental results show that our approach can significantly improve SRL performance, especially in Chinese PropBank.

1 Introduction

Semantic Role Labeling (SRL) is a kind of shallow semantic parsing task and its goal is to recognize some related phrases and assign a joint structure (WHO did WHAT to WHOM, WHEN, WHERE, WHY, HOW) to each predicate of a sentence (Gildea and Jurafsky, 2002). Because of the ability of encoding semantic information, SRL has been applied in many tasks of NLP, such as question and answering (Narayanan and Harabagir, 2004), information extraction (Surdeanu et al., 2003; Christensen et al., 2005), and machine translation (Wu and Fung, 2009; Liu and Gildea, 2010; Xiong et al., 2012; Zhai et al., 2012).

The justices will be **forced** to **reconsider** the questions.

[A1] [Pred]

[A0] [Pred] [A1]

Figure 1: A sentence from English PropBank, with an argument shared by multiple predicates

Currently, an SRL system works as follows: first identify argument candidates and then perform classification for each argument candidate. However, this process only focuses on one independent predicate without considering the internal relations of multiple predicates in a sentence. According to our statistics, more than 80% sentences in Propbank carry multiple predicates. One example is shown in Figure 1, in which there are two predicates 'Force' and 'Reconsider'. Moreover, the constituent 'the justices' is shared by the two predicates and is labeled as A1 for 'Force' but as A0 for 'Reconsider'. We call this phenomenon of the shared arguments **Role Transition** . Intuitively, all predicates in a sentence are closely related to each other and the internal relations between them would be helpful for SRL.

This paper has made deep investigation on multi-predicate semantic role labeling. We think there are mainly two key points: argument identification and role labeling of the arguments shared by multiple predicates. We adopt different strategies to address these two issues.

During argument identification, there are a large number of identification errors caused by the poor performance of auto syntax trees. However, many of these errors can be removed, if we take other predicates into consideration. To achieve this purpose, we propose novel predicates-related features which have been proved to be effective to recog-

Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 363–373,
October 25-29, 2014, Doha, Qatar.

nize many identification errors. After these features added, the precision of argument identification improves significantly by 1.6 points and 0.9 points in experiments on Chinese PropBank and English PropBank respectively, with a slight loss in recall.

Role labeling of the shared arguments is another key point. The predicates and their shared argument could be considered as a joint structure, with strong dependencies between the shared argument's roles. If we consider linguistic basis for joint modeling of the shared argument's roles, there are at least two types of information to be captured. The first type of information is the compatibility of Role Transition among the shared argument's roles. A noun phrase may be labeled as A0 for a predicate and at the same time, it can be labeled as A1 for another predicate. However, there are few cases that a noun phrase is labeled as A0 for a predicate and as AM-ADV for another predicate at the same time. Secondly, joint modeling the shared arguments could explore global information. For example, in '*The columbia mall is expected to open*', there are two predicates 'expect' and 'open' and a shared argument 'the columbia mall'. Because this shared argument is before 'open' and the predicate 'open' is in active voice, a base classifier often incorrectly label this argument A0 for 'open'. But if we observe that the argument is also an argument of 'expect', it should be labeled as A1 for 'expect' and 'open'.

Motivated by the above observations, we attempt to jointly model the shared arguments' roles. Specifically, we utilize the discriminative reranking approach that has been successfully employed in many NLP tasks. Typically, this method first creates a list of n-best candidates from a base system, and then reranks them with arbitrary features (both local and global), which are either not computable or are computationally intractable within the base model.

We conducted experiments on Chinese PropBank and English PropBank. Results show that compared with a state-of-the-art base model, the accuracy of our joint model improves significantly by 2.4 points and 1.5 points on Chinese PropBank and English PropBank respectively, which suggests that there are substantial gains to be made by jointly modeling the shared arguments of multiple predicates.

Our contributions can be summarized as follows:

- To the best of our knowledge, this is the first work to investigate the mutual effect of multiple predicates' semantic role labeling.
- We present a rich set of features for argument identification and shared arguments' classification that yield promising performance.
- We evaluate our method on two standard benchmarks: Chinese PropBank and English PropBank. Our approach performs well in both, which suggests its good universality.

The remainder of this paper is organized as follows. Section 2 gives an overview of our approach. We discuss the mutual effect of multi-predicate' argument identification and argument classification in Section 3 and Section 4 respectively. The experiments and results are presented in Section 5. Some discussion and analysis can be found in Section 6. Section 7 discusses the related works. Finally, the conclusion and future work are in Section 8.

2 Approach Overview

As illustrated in Figure 2, our approach follows the standard separation of the task of semantic role labeling into two phases: **Argument Identification** and **Argument Classification** . We investigate the effect of multiple predicates in Argument Identification and Argument Classification respectively. Specifically, in the stage of Argument Identification, we introduce new features related to predicates which are effective to recognize many argument identification errors. In the stage of Argument Classification, we concentrate on the classification of the arguments shared by multiple predicates. We first use a base model to generate n-best candidates for the shared arguments and then construct a joint model to rerank the n-best list, in which a rich set of global features are proposed.

3 Argument Identification

In this section, we investigate multi-predicate' mutual effects in Argument Identification. Argument Identification is to recognize the arguments from all candidates of each predicate. Here, we use the Maximum Entropy (ME) classifier to perform binary classification. As a discriminative model, ME can easily incorporate arbitrary features and

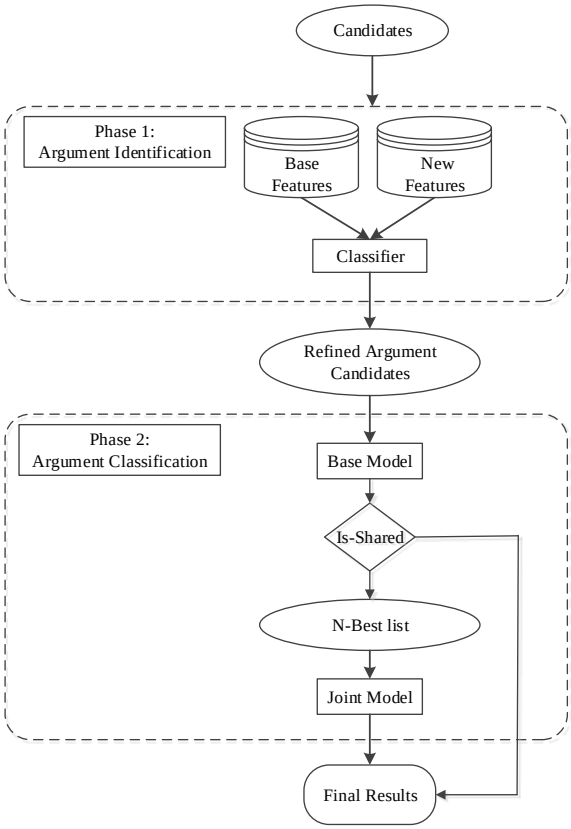

Figure 2: The overview of our approach

achieve good performance. The model is formulated as follows:

$$p(y|x) = \frac{1}{Z(x)} \exp(\sum_i \theta_i f_i(x, y)) \quad (1)$$

in which x is the input sample, y(0 or 1) is the output label, $f(x, y)$ are feature functions and $Z(x)$ is a normalization term as follows:

$$Z(x) = \sum_y \exp(\sum_i \theta_i f_i(x, y))$$

3.1 Base Features

Xue (2008) took a critical look at the features used in SRL and achieved good performance. So, we use the same features in Xue (2008) as the base features:

- Predicate lemma
- Path from node to predicate
- Head word
- Head word's part-of-speech
- Verb class (Xue, 2008)
- Predicate and Head word combination
- Predicate and Phrase type combination
- Verb class and Head word combination
- Verb class and Phrase type combination

3.2 Additional Features

In the SRL community, it is widely recognized that the overall performance of a system is largely determined by the quality of syntactic parsers (Gildea and Palmer, 2002), which is particularly notable in the identification stage. Unfortunately, the state-of-the-art auto parsers fall short of the demands of applications. Moreover, when there are multiple predicates, or even multiple clauses in a sentence, the problem of syntactic ambiguity increases drastically (Kim et al., 2000). For example, in Figure 3, there is a sentence with two consecutive predicates '是' (is) and '开发' (develop). Compared with the gold tree, the auto tree is less preferable, which makes the classifier easily mistake '建筑' (building) as an argument of '开发' (develop) with base features. But this identification error can be removed if we note that there is another predicate '是' (is) before '开发' (devel-

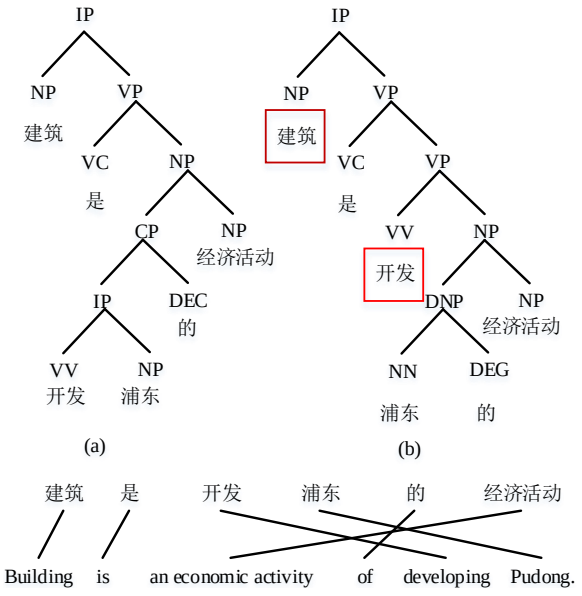

Figure 3: An example from Chinese PropBank. Tree (a) is the gold syntax tree and (b) is parsed by a state of-the-art parser Berkeley parser. On tree (b), '建筑' (building) is mistaken as an argument of '开发' (develop) with base features.

op). Similar examples with the pattern 'NP + 是 + VV' can be found in PropBank, in which the subject NP of the sentence is usually not an argument of the latter predicate. Thus, '是' (is) is an effective clue to detect this kind of identification error.

It is challenging to obtain a fully correct syntax tree for a complex sentence with multiple predicates. Therefore, base features that heavily rely on syntax trees often fail in discriminating arguments from candidates as demonstrated in Figure 3. However, by considering the elements of neighboring predicates, we could capture useful clues like in the above example and remove many identification errors. Below, we define novel predicaterelated features to encode these 'clues' to refine candidates.

There are mainly five kinds of features as follows.

- **Is the given predicate the nearest one?**

  This is a binary feature that indicates whether the predicate is the nearest one to the candidate.

- **Local adjunct**

  This is a binary feature designed for adjective and adverbial phrases. Some adjunct phrases, such as '仅' (only), have a limited sphere of influence. If the candidate is 'local' but the given predicate is not the nearest one, the candidate is often not an argument for the given predicate. To collect local adjuncts, we traverse the whole training set to get the initial lexicon of adjuncts and refine it manually.

- **Cut-Clause**

  This type of feature is a binary feature designed to distinguish identification errors of noun phrase candidates. If a noun phrase candidate is separated from the given predicate by a clause consisting of a NP and VP, the candidate is usually not the argument of the given predicate.

- **Different Relative Positions with Conjunctions**

  This is a binary feature that describes whether the candidate and the predicate are located in different positions as separated by conjunctions such as '但是' (but). Conjunctions are often used to concatenate two clauses, but the first clause commonly describes one proposition and the second clause describes another one. Thus, if the candidate and the given predicate have different positions relative to the conjunctions, the candidate is usually not the argument of the given predicate.

- **Consecutive Predicates Sequence**

  When multiple predicates appear in a sentence consecutively, parse errors frequently occurs due to the problems of syntactic ambiguity as demonstrated in Figure 2. To indicate such errors, sequence features of the candidates and consecutive predicates are defined specifically. For instance, for the candidate '建筑' (building) of '开发' (develop), the features are 'cand-是-开发' and 'cand-是-VV', in which we use 'cand' to represent the position of the candidate.

4 Argument Classification

In this section, we investigate multi-predicate' mutual effects in Argument Classification. Argument Classification is to assign a label to each argument candidate recognized by the phase of Argument Identification.

4.1 Base Model

A conventional method in Argument Classification is to assign a label to each argument candidate by a classifier independently. We call this kind of method Base Model. In the base model, we still adopt ME (1) as our classifier; all base features of Argument Identification are contained (shown in subsection 3.1). In addition, there are some other features:

- Position: the relative position of the candidate argument compared to the predicate
- Subcat frame: the syntactic rule that expands the parent of the verb
- The first and the last word of the candidate
- Phrase type: the syntactic tag of the candidate argument
- Subcat frame+: the frame that consists of the NPs (Xue, 2008).

4.2 Joint Model

As discussed briefly in Section 1, there are many dependencies between the shared arguments' labeling for different predicates, but the base model completely ignores such useful information. To incorporate these dependencies, we employ the discriminative reranking method. Here, we first establish a unified framework for reranking. For an input x, the generic reranker selects the best output y^* among the set of candidates $GEN(x)$ according to the scoring function:

$$y^* = \underset{y \in GEN(x)}{\operatorname{argmax}} \; score(y) \tag{2}$$

In our task, $GEN(x)$ is a set of the n-best candidates generated from the base model. As usual, we calculate the score of a candidate by the dot product between a high dimensional feature and a weight W:

$$score(y) = W \cdot f(y) \tag{3}$$

We estimate the weight W using the averaged perceptron algorithm (Collins, 2002a) which is well known for its fast speed and good performance in similar large-parameter NLP tasks (Huang, 2008). The training algorithm of the generic averaged perceptron is shown in Table 1. In line 5, the algorithm updates W with the difference (if any) between the feature representations of the best scoring candidate and the gold candidate. We also use a refinement called "averaged parameters" that the final weight vector W is the average of weight vectors over T iterations and N samples. This averaging effect has been shown to reduce overfitting and produces more stable results (Collins, 2002a).

Pseudocode: Averaged Structured Perceptron
1: **Input**: training data(x_t, y_t^*) for $t = 1, ..., T$;
2: $\bar{w}^{(0)} \leftarrow 0$; $v \leftarrow 0$; $i \leftarrow 0$
3: **for** n in 1,...,N **do**
4: **for** t in $1, ..., T$ **do**
5: $\bar{w}^{(i+1)} \leftarrow$ update $\bar{w}^{(i)}$ according to (x_t, y_t^*)
6: $v \leftarrow v + \bar{w}^{i+1}$
7: $i \leftarrow i + 1$
8: $\bar{w} \leftarrow v//(N * T)$
9: **return** $\bar{w}$

Table 1: The perceptron training algorithm

4.3 Features for Joint Model

Here, we introduce features used in the joint model. For clear illustration, we describe these features in the context of the example in Figure 1.

Role Transition (RT): a binary feature to indicate whether the transitions among roles of the candidate are reasonable. Because all roles are assigned to the same candidate, all role transitions should be compatible. For instance, if an argument is labeled as AM-TMP for one predicate, it cannot be labeled as AM-LOC for another predicate. This feature is constructed by traversing the training data to ascertain whether transitions between all roles are reasonable. In Table 2, we list some role transitions which are obtained from the training data of experiments on Chinese PropBank.

Roles and Predicates' Sequence (RPS): a joint feature template that concatenates roles and the given predicates. For the gold candidate 'Arg1, Arg0', the feature is 'Arg1-force, Arg0-reconsider'.

Roles and Predicates' Sequence with Relative Orders (RPSWR): the template is similar to the above one except that relative orders between roles and predicates are added. If the shared argument is before the given predicate, the feature is described as 'Role-Predicate'; otherwise, the feature is 'Predicate-Role'. And, if the predicate's voice is passive, the order is reversed. Thus, for the gold candidate 'Arg1, Arg0', this feature is 'force-Arg1, Arg0-reconsider'.

Roles and Phrase Type Sequence (RPTS)

Roles and Head Word Sequence (RHWS)

Roles and Head Word's POS Sequence (RHWPS)

These three features are utilized to explore the shared argument's relations with roles.

Time and Location Class (TLC): We find there are much confusions between AM-TMP and AM-LOC in the base model. To fix these errors, we add two features: Time and Location Class. For these features, we just collect phrases labeled as AM-TMP and AM-LOC from the training data. When the argument belongs to Time or Location Class, we add a sequence template consisting of 'Role-Time' for Time Class or 'Role-Location' for Location Class. For the gold candidate 'Arg1, Arg0', the feature is 'Arg1-none, Arg0-none' because 'the justices' belongs neither to Time Class nor to Location Class.

Role	Arg0	Arg1	Arg2	AM-LOC	AM-TMP	AM-ADV	AM-MNR	AM-TPC
Arg0	+	+	+	+	+	+	+	+
Arg1	+	+	+	+	-	+	+	+
Arg2	+	+	+	+	-	-	-	+
AM-LOC	+	+	+	+	-	+	-	+
AM-TMP	+	-	-	-	+	+	-	-
AM-ADV	+	-	+	-	+	+	-	-
AM-MNR	+	+	-	-	-	-	+	-
AM-TPC	+	+	+	+	+	+	-	+

Table 2: Some role transitons from Chinese PropBank. "+" means reasonable role transition and "-" means illegal.

5 Experiments

5.1 Experimental Setting

To evaluate the performance of our approach, we have conducted on two standard benchmarks: Chinese PropBank and English PropBank. The experimental setting is as follows:

Chinese:

We use Chinese Proposition Bank 1.0. All data are divided into three parts. 648 files (from chtb_081.fid to chtb_899.fid) are used as the training set. 40 files (from chtb_041.fid to chtb_080.fid) constitutes the development set. The test set consists of 72 files (chtb_001.fid to chtb_040.fid and chtb_900.fid to chtb_931.fid). This data setting is the same as in (Xue, 2008; Sun et al., 2009). We adopt Berkeley Parser[1] to carry out auto parsing for SRL and the parser is retrained on the training set. We used n =10 joint assignments for training the joint model and testing.

English:

We choose English Propbank as the evaluation corpus. According to the traditional partition, the training set consists of the annotations in Sections 2 to 21, the development set is Section 24, and the test set is Section 23. This data setting is the same as in (Xue and Palmer, 2004; Toutanova et al., 2005). We adopt Charniak Parser[2] to carry out auto parsing for SRL and the parser is retrained on the training set. We used n =10 joint assignments for training the joint model and testing.

5.2 Experiment on Argument Identification

We first investigate the performance of our approach in Argument Identification.

For the task of Argument Identification (AI), we adopt auto parser to produce auto parsing trees for SRL. The results are shown in Table 3. We can see that in the experiment of Chinese, the F1 score reaches to 78.79% with base features. While after additional predicates-related features are added, the precision has improved by 1.6 points with slight loss in recall, which leads to the improvement of 0.6 points in F1. The similar effect occurred in the experiment of English. After additional features added in the identification module, the precision is improved by about 0.9 points with a slight loss in recall, leading to an improvement of 0.3 points in F1. However, the improvement in English is slight smaller than in Chinese. We think the main reason is that there are less parse errors in English than in Chinese. All results demonstrate that the novel predicted-related features are effective in recognizing many identification errors which are difficult to discriminate with base features.

		P(%)	R(%)	F_1(%)
Ch	Base	84.36	73.90	78.79
	+Additional	85.97	73.72	79.38*
En	Base	82.86	76.83	79.73
	+Additional	83.75	76.69	80.06

Table 3: Comparison with Base Features in Argument Identification. Scores marked by "*" are significantly better ($p < 0.05$) than base features.

5.3 Experiment on Argument Classification

5.3.1 Results

Errors produced in AI will influenced the evaluation of Argument Classification (AC). So, to evaluate fairly we assume that the argument constituents of a predicate are already known, and the

[1] http://code.google.com/p/berkeleyparser/

[2] https://github.com/BLLIP/bllip-parser

		Num	Acc(%)
Ch	Shared	2060	91.36
	All	8462	92.77
En	Shared	2015	93.85
	All	14061	92.30

Table 4: Performance of the Base Model in Argument Classification

	Methods	Acc(%)
Ch	Base	91.36
	Joint	**93.74***
En	Base	93.85
	Joint	**95.33***

Table 5: Comparison with Base Model on shared arguments. Scores marked by "*" are significantly better ($p < 0.05$) than base model.

task is only to assign the correct labels to the constituents. The evaluation criterion is Accuracy.

The results of the base model are shown in Table 4. We first note that in testing set, there are a large number of shared arguments, which weighs about one quarter of all arguments in Chinese and 14% in English. Therefore, the fine processing of these arguments is essential for argument classification. However, the base model cannot handle these shared arguments so well in Chinese that the accuracy of the shared arguments is lower by about 1.4 points than the average value of all arguments. Nevertheless, from Table 5 we can see that our joint model's accuracy on the shared arguments reaches 93.74%, 2.4 points higher than the base model in Chinese. Although the base model obtain good performance on shared arguments of English, our joint model's performance reaches 95.33%, 1.5 points higher than the base model. This indicates that even though the base model is optimized to utilize a large set of features and achieves the state-of-the-art performance, it is still advantageous to model the joint information of the shared arguments.

Another point to note is that our joint model in resolving English SRL task is not so good as in Chinese SRL. There are mainly two reasons. The first reason is that the shared arguments occur less in English than in Chinese so that training samples are insufficient for our discriminative model. The second reason is the annotation of some intransitive verbs. In English PropBank, there is a class of intransitive verbs such as "land" (known as verbs of variable behavior), for which the argument can be tagged as either ARG0 or ARG1. Here, we take examples from the guideline[3] of English PropBank to explain.

"A bullet (ARG1) landed at his feet"

"He (ARG0) landed"

In the above examples, the two arguments and the predicate 'land' have the same relative order and voice but the arguments have different labels for their respective predicates. In fact, according to the intention of the annotator, ARG0 and ARG1 are both correct. Unfortunately, in English PropBank, there is only one gold label for each argument, which leads to much noise for our joint model. Moreover, such situations are not rare in the corpus.

5.3.2 Feature Performance

We investigate effects of the features of joint model to the performance and results are shown in Table 6. Each row shows the improvement over the baseline when that feature is used in the joint model. We can see that features proposed are beneficial to the performance of the joint model. But some features like 'RPS' and 'RPSRO' play a more important role.

Features	Chinese	English
base	91.36	93.85
RT	91.70	94.10
RPS	92.30	94.70
RPSRO	92.24	94.50
RPTS	91.80	94.18
RHWS	91.63	93.95
RHWPS	91.43	94.23
TCL	91.93	94.23
All	**93.74**	**95.33**

Table 6: Features performance in the Joint Model. We use first letter of words to represent features.

5.4 SRL Results

We also conducted the complete experiment on the auto parse trees. The results are shown in Table 7. In experiments on Chinese PropBank, we can see that after novel predicate-related features are added in the stage of Argument Identification, our model outperforms the base model by 0.5 points

[3] http://verbs.colorado.edu/propbank/EPB-AnnotationGuidelines.pdf

		$F_1(\%)$
Chinese	Base	74.04
	Base + AI	74.50
	Base + AI + AC	75.31
English	Base	76.44
	Base + AI	76.70
	Base + AI + AC	77.00

Table 7: Results on auto parse trees. Base means the baseline system, +AI meaning predcates-related features added in AI, + AC meaning joint module added.

	Methods	$F_1(\%)$
Chinese	Xue(2008)	71.90
	Sun et al.(2009)	74.12
	Ours	75.31
English	Surdeanu and Turmo(2005)	76.46
	Ours	77.00

Table 8: Comparison with Other Methods

in F1. Furthermore, after incorporating the joint module, the performance goes up to 75.31%, 1.3 points higher than the base model. We obtain similar observations in experiments on English PropBank, but due to reasons illustrated in Subsection 5.3, the performance of our method is slight better than the base model.

We compare our method with others and the results are shown in Table 8. In Chinese, Xue (2008) and Sun et al. (2009) are pioneer works in Chinese SRL. Our approach outperforms these approaches by about 3.4 and 1.9 F1 points respectively. In English SRL, we compare out method with Surdeanu and Turmo (2005) which is best result obtained with single parse tree as the input in CONLL 2005 SRL evaluation. Our approach is better than their approach which ignores the relation of multiple predicates' SRL.

6 Discussion and Analysis

In this section, we discuss some case studies that illustrate the advantages of our model. Some examples from our experiments are shown in Table 9. In example (1), the argument is a prepositional phrase '在 普及 九 年 义务 教育 的 同时' (at the same time of compulsory education) and shared by two predicates '得到' (witness) and '扩大' (expand). In the corpus, a prepositional phrase is commonly labeled as ARGM-LOC and ARGM-TMP. Thus, the base model labeled the argument into these classes but one as ARGM-LOC, another as ARGM-TMP. Unfortunately, ARGM-LOC for '得到' (witness) is wrong while our joint model outputs both correct answers, which benefits from the role transition feature. From Table 1, we can see that the role transition between ARGM-TMP and ARGM-LOC is impossible, which lowers the score of candidates containing both ARGM-LOC and ARGM-TMP in the joint model. Thus, the joint model is more likely to output the gold candidate.

In example (2), the argument is '海拉尔 飞机场' (Hailar Airport) and shared by two predicates '扩建' (expand) and '成为' (become). Because of the high similarity of the features in the base model, the argument for both predicates is classified into the same class ARG0, but the label for '扩建' (expand) is wrong. Nevertheless, our joint model obtains both correct labels, which benefits from the global features. After searching the training data, we find some similar examples to this one, such as '地铁 运行 里程 已 扩建 至 1 2 0 公里' (The railway operation mileage is expanded to 120 kilometers), in which '地铁 运行 里程' (the railway operation mileage) is labeled as ARG1 for '扩建' (expand) but ARG0 for '至' (to). We think these samples provide evidence for our joint model while these information has not been captured by the base model.

In example (3), the argument is '国内外 知名度很 高 的 大 集团' (a large group with high reputation) and shared by predicates '发展' (develop) and '为' (become). Different from the above cases in which only one label is wrong in the base model, both labels for '发展' (develop) and '为' (become) are misclassified by the base model. However, our method still gets correct answers for both predicates, which also benefits from the global features.

7 Related work

Our work is related to semantic role labeling and discriminative reranking. In this section, we briefly review these two types of work.

On Semantic Role Labeling

Gildea and Jurafsky (2002) first presented a system based on a statistical classifier which is trained on a hand-annotated corpora FrameNet. In their pioneering work, they used a gold or autoparsed syntax tree as the input and then extracted various lexical and syntactic features to identify the

Examples	Base	Ours
1. **在 普及 九 年 义务 教育 的 同时**，中等 职业 教育 也 **得到**[1]发展规模，不断**扩大**[2] (**At the same time of compulsory education**, secondary vocational education have **achieved**[1] significant development and constant **expanding**[2])	ARGM-LOC | 得到 ARGM-TMP | 扩大	ARGM-TMP | 得到 ARGM-TMP | 扩大
2. **海拉尔飞机场 扩建**[1] **成为**[2] 国际航空港 (**Hailar Airport** had been **expanded**[1] and **became**[2] an international airport)	ARG0 | 扩建 ARG0 | 成为	ARG1 | 扩建 ARG0 | 成为
3. 海尔 集团 的 销售 收入 已 突破 六十亿元, **发展**[1] **为**[2] **国内外知名度 很 高 的 大 集团** (Haier Group's sales revenue has exceeded six billion yuan and it has **developed**[1] to **be**[2] **a large group with high reputation**)	ARG1 | 发展 ARG0 | 为	ARG3 | 发展 ARG1 | 为

Table 9: Some examples in our experiments

semantic roles for a given predicate. After Gildea and Jurafsky (2002), there have been a large number of works on automatic semantic role labeling. Based on a basic discriminative model, Punyakanok et al. (2004) constructed an integer linear programming architecture, in which the dependency relations among arguments are implied in the constraint conditions. Toutanova et al. (2008) proposed a joint model to explore relations of all arguments of the same predicate. Unlike them, this paper focus on mining relations of different predicates' semantic roles in one sentence. And, there have been many extensions in machine learning models (Moschitti et al., 2008), feature engineering (Xue and Palmer, 2004), and inference procedures (Toutanova et al., 2005; Punyakanok et al., 2008; Zhuang and Zong, 2010a; Zhuang and Zong, 2010b).

Sun and Jurafsky (2004) did the preliminary work on Chinese SRL without employing 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 utilized Support Vector Machine to identify and classify the arguments. They made the first attempt on Chinese SRL and produced promising results. After the PropBank (Xue and Palmer, 2003) was built, Xue and Palmer (2004) and Xue (2008) took a critical look at features of argument detection and argument classification. Unlike others' using syntax trees as the input of SRL, Sun et al. (2009) performed Chinese semantic role labeling with shallow parsing. Li et al. (2010) explored joint syntactic and semantic parsing of Chinese to further improve the performance of both syntactic parsing and SRL.

However, to the best of our knowledge, in the literatures, there is no work related to multi-predicate semantic role labeling.

On Discriminative Reranking

Discriminative reranking is a common approach in the NLP community. Its general procedure is that a base system first generates n-best candidates and with the help of global features, we obtain better performance through reranking the n-best candidates. It has been shown to be effective for various natural language processing tasks,such as syntactic parsing (Collins, 2000; Collins, 2002b; Collins and Koo, 2005; Charniak and Johnson, 2005; Huang, 2008), semantic parsing (Lu et al., 2008; Ge and Mooney, 2006), part-of-speech tagging (Collins, 2002a), named entity recognition (Collins, 2002c), machine translation (Shen et al., 2004) and surface realization in generation (Konstas and Lapata, 2012).

8 Conclusion and Feature Work

This paper investigates the interaction effect among multi-predicate's SRL. Our investigation has shown that there is much interaction effect of multi-predicate's SRL both in Argument Identification and in Argument Classification. In the stage of argument identification, we proposed novel features related to predicates and successfully removed many argument identification errors. In the stage of argument classification, we concentrate on the classification of the arguments shared by multiple predicates. Experiments have shown

that the base model often fails in classifying the shared arguments. To perform the classification of the shared arguments, we propose a joint model and with the help of the global features, our joint model yields better performance than the base model. To the best of our knowledge, this is the first work of investigating the interaction effect of multi-predicate's SRL.

In the future, we will explore more effective features for multi-predicate's identification and classification. Since we adopt reranking approach in the shared arguments' classification, the performance is limited by n-best list. Also, we would like to explore whether there is another method to resolve the problem.

Acknowledgments

We thank the three anonymous reviewers for their helpful comments and suggestions. We would also like to thank Nianwen Xue for help in baseline system, and Tao Zhuang, Feifei Zhai, Lu Xiang and Junjie Li for insightful discussions. The research work has been partially funded by the Natural Science Foundation of China under Grant No.61333018 and the Hi-Tech Research and Development Program ("863" Program) of China under Grant No.2012AA011101, and also the High New Technology Research and Development Program of Xinjiang Uyghur Autonomous Region under Grant No.201312103 as well.

References

Collin F. Baker, Charles j. Fillmore, and John B. Lowe. 1998. The berkeley framenet project. In *Proceedings of COLING-ACL 1998*.

Janara Christensen, Mausam, Stephen Soderland and Oren Etzioni. 2010. Semantic Role Labeling for Open Information Extraction. In *Proceedings of ACL 2010*.

Eugene Charniak and Mark Johnson. 2005. Coarse tofine n-best parsing and maxent discriminative reranking. In *Proceedings of ACL 1998*.

Michael Collins and Terry Koo. 2005. Discriminative reranking for natural language parsing. *Computational Linguistics*, 31(1):25 ‑ 69.

Michael Collins. 2000. Discriminative reranking fornatural language parsing. In *Proceedings of ICML 2000*.

Michael Collins. 2002a. Discriminative training methods for hidden Markov models: Theory and experiments with perceptron algorithms. In *Proceedings of EMNLP 2002*.

Michael Collins. 2002b. New ranking algorithms for parsing and tagging: Kernels over discrete structures, and the voted perceptron. In *Proceedings of ACL 2002*.

Michael Collins. 2002c. Ranking algorithms fornamed-entity extraction: Boosting and the voted perceptron. In *Proceedings of ACL 2002*.

Ruifang Ge and Raymond J. Mooney. 2006. Discriminative reranking for semantic parsing. In *Proceedings of COLING/ACL 2002*.

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

Daniel Gildea and Martha Palmer. 2002. The necessity of parsing for predicate argument recognition. In *ACL 2002*.

Liang Huang. 2008. Forest reranking: Discriminative parsing with non-local features. In *Proceedings of ACL 2008*.

Sung Dong Kim, Byoung-Tak Zhang and Yung Taek Kim. 2000. Reducing Parsing Complexity by IntraSentence Segmentation based on Maximum Entropy Model. In *Proceedings of SIGDAT 2000*.

Paul Kingsbury and Martha Palmer. 2002. From Treebank to PropBank. In *Proceedings of LREC 2002*.

Junhui Li, Guodong Zhou and Hwee Tou Ng. 2010. Joint Syntactic and Semantic Parsing of Chinese. In *Proceedings of ACL 2010*.

Ding Liu and Daniel Gildea. 2010. Semantic role features for machine translation. In *Proceedings of COLING 2010*.

Junhui Li, Guodong Zhou and Hwee Tou Ng. 2010. Wei Lu, Hwee Tou Ng, Wee Sun Lee, and Luke S. Zettlemoyer. 2008. A generative model for parsing natural language to meaning representations. In *Proceedings of EMNLP 2008*.

Alessandro Moschitti, Daniel Pighin and Roberto Basili. 2008. Tree Kernels for Semantic Role Labeling. In *Computational Linguistics*, 34(2): 193-224.

Srini Narayanan and Sanda Harabagiu. 2004. Question Answering based on Semantic Structures. In *Proceedings of COLING 2004*.

Vasin Punyakanok, Dan Roth, Wen-tau Yih and Dav Zimak. 2004. Semantic Role Labeling via Integer Linear Programming Inference. In *Proceedings of COLING 2004*.

Libin Shen, Anoop Sarkar, and Franz Josef Och. 2004. Discriminative reranking for machine translation. In *Proceedings of HLT/NAACL 2004*.

Mihai Surdeanu, Sanda Harabagiu, John Williams and Paul Aarseth. 2003. Using Predicate-Argument Structures for Information Extraction. In *Proceedings of ACL 2003*.

Mihai Surdeanu and Jordi Turmo. 2005. Semantic Role Labeling Using Complete Syntactic Analysis. In *Proceedings of CONLL 2005*.

Weiwei Sun, Zhifang Sui, Meng Wang and Xin Wang. 2009. Chinese Semantic Role Labeling with Shallow Parsing. In *Proceedings of ACL 2009*.

Ioannis Konstas and Mirella Lapata. 2012. Concept-totext generation via discriminative reranking. In *Proceedings of ACL 2012*.

Kristina Toutanova, Aria Haghighi and Christopher D.Manning. 2005. Joint learning improves semantic role labeling. In *Proceedings of ACL 2005*.

Kristina Toutanova, Aria Haghighi and Christopher D.Manning. 2008. A Global Joint Model for Semantic Role Labeling. In *Computational Linguistics*, 34(2): 161-191.

Dekai Wu and Pascale Fung. 2009. Can semantic role labeling improve smt. In *Proceedings of EAMT 2009*.

Deyi Xiong, Min Zhang and Haizhou Li. 2012. Modeling the Translation of Predicate-Argument Structure for SMT. In *Proceedings of ACL 2012*.

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

Nianwen Xue and Martha Palmer. 2004. Calibrating Features for Semantic Role Labeling. In *Proceedings of EMNLP 2004*.

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

Feifei Zhai, Jiajun Zhang, Yu Zhou and Chengqing Zong. 2012. Machine Translation by Modeling Predicate-Argument Structure Transformation. In *Proceedings of COLING 2012*.

Tao Zhuang and Chengqing Zong. 2010a. A Minimum Error Weighting Combination Strategy for Chinese Semantic Role Labeling. In *Proceedings of COLING 2010*.

Tao Zhuang and Chengqing Zong. 2010b. Joint Inference for Bilingual Semantic Role Labeling. In *Proceedings of EMNLP 2010*.