# Semantics-Driven Shallow Parsing for Chinese Semantic Role Labeling

**Weiwei Sun**
Department of Computational Linguistics, Saarland University
German Research Center for Artificial Intelligence (DFKI)
D-66123, Saarbrücken, Germany
wsun@coli.uni-saarland.de

## Abstract

One deficiency of current shallow parsing based Semantic Role Labeling (SRL) methods is that syntactic chunks are too small to effectively group words. To partially resolve this problem, we propose semantics-driven shallow parsing, which takes into account both syntactic structures and predicate-argument structures. We also introduce several new "path" features to improve shallow parsing based SRL method. Experiments indicate that our new method obtains a significant improvement over the best reported Chinese SRL result.

## 1 Introduction

In the last few years, there has been an increasing interest in Semantic Role Labeling (SRL) on several languages, which consists of recognizing arguments involved by predicates of a given sentence and labeling their semantic types. Both full parsing based and shallow parsing based SRL methods have been discussed for English and Chinese. In Chinese SRL, shallow parsing based methods that cast SRL as the classification of syntactic chunks into semantic labels has gained promising results. The performance reported in (Sun et al., 2009) outperforms the best published performance of full parsing based SRL systems.

Previously proposed shallow parsing takes into account only syntactic information and basic chunks are usually too small to group words into argument candidates. This causes one main deficiency of Chinese SRL. To partially resolve this problem, we propose a new shallow parsing. The new chunk definition takes into account both syntactic structure and predicate-argument structures of a given sentence. Because of the semantic information it contains, we call it semantics-driven shallow parsing. The key idea is to make basic chunks as large as possible but not overlap with arguments. Additionally, we introduce several new "path" features to express more structural information, which is important for SRL.

We present encouraging SRL results on Chinese PropBank (CPB) data. With semantics-driven shallow parsing, our SRL system achieves 76.10 F-measure, with gold segmentation and POS tagging. The performance further achieves 76.46 with the help of new "path" features. These results obtain significant improvements over the best reported SRL performance (74.12) in the literature (Sun et al., 2009).

## 2 Related Work

CPB is a project to add predicate-argument relations to the syntactic trees of the Chinese TreeBank (CTB). Similar to English PropBank, the arguments of a predicate are labeled with a contiguous sequence of integers, in the form of AN (N is a natural number); the adjuncts are annotated as such with the label AM followed by a secondary tag that represents the semantic classification of the adjunct. The assignment of argument labels is illustrated in Figure 1, where the predicate is the verb "提供/provide" For example, the noun phrase "保险公司/the insurance company" is labeled as *A0*, meaning that it is the *proto-Agent* of "提供".

Sun et al. (2009) explore the Chinese SRL problem on the basis of shallow syntactic information at the level of phrase chunks. They present a semantic chunking method to resolve SRL on basis of shallow parsing. Their method casts SRL as the classification of syntactic chunks with IOB2 representation for semantic roles (i.e. semantic

*Proceedings of the ACL 2010 Conference Short Papers*, pages 103–108,
Uppsala, Sweden, 11-16 July 2010. ©2010 Association for Computational Linguistics

|  |  |  |  |  |  |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|
| WORD: | 保险 | 公司 | 已 | 为 | 三峡 | 工程 | 提供 | 保险 | 服务 |
|  | insurance | company | already | for | Sanxia | Project | provide | insurance | service |
| POS: | [NN | NN] | [AD] | [P] | [NR] | [NN] | [VV] | [NN | NN] |
| SYN CH: | [NP] |  | [ADVP] | [PP | NP | NP ] | [VP] | [NP] |  |
| SEM CH: | B-A0 |  | B-AM-ADV | B-A2 | I-A2 | I-A2 | B-V | B-A1 |  |
|  | The insurance company has provided insurance services for the Sanxia Project. |  |  |  |  |  |  |  |  |

Figure 1: An example from Chinese PropBank.

chunks). Two labeling strategies are presented: 1) directly tagging semantic chunks in one-stage, and 2) identifying argument boundaries as a chunking task and labeling their semantic types as a classification task. On the basis of syntactic chunks, they define semantic chunks which do not overlap nor embed using IOB2 representation. Syntactic chunks outside a chunk receive the tag O (Outside). For syntactic chunks forming a chunk of type *A**, the first chunk receives the *B-A** tag (Begin), and the remaining ones receive the tag *I-A** (Inside). Then a SRL system can work directly by using sequence tagging technique. Shallow chunk definition presented in (Chen et al., 2006) is used in their experiments. The definition of syntactic and semantic chunks is illustrated Figure 1. For example, "保险公司/the insurance company", consisting of two nouns, is a noun phrase; in the syntactic chunking stage, its two components "保险" and "公司" should be labeled as *B-NP* and *I-NP*. Because this phrase is the *Agent* of the predicate "提供/provide", it takes a semantic chunk label *B-A0*. In the semantic chunking stage, this phrase should be labeled as *B-A0*.

Their experiments on CPB indicate that according to current state-of-the-art of Chinese parsing, SRL systems on basis of full parsing do not perform better than systems based on shallow parsing. They report the best SRL performance with gold segmentation and POS tagging as inputs. This is very different from English SRL. In English SRL, previous work shows that full parsing, both constituency parsing and dependency parsing, is necessary.

Ding and Chang (2009) discuss semantic chunking methods without any parsing information. Different from (Sun et al., 2009), their method formulates SRL as the classification of words with semantic chunks. Comparison of experimental results in their work shows that parsing is necessary for Chinese SRL, and the semantic chunking methods on the basis of shallow parsing outperform the ones without any parsing.

Joint learning of syntactic and semantic structures is another hot topic in dependency parsing research. Some models have been well evaluated in CoNLL 2008 and 2009 shared tasks (Surdeanu et al., 2008; Hajič et al., 2009). The CoNLL 2008/2009 shared tasks propose a unified dependency-based formalism to model both syntactic dependencies and semantic roles for multiple languages. Several joint parsing models are presented in the shared tasks. Our focus is different from the shared tasks. In this paper, we hope to find better syntactic representation for semantic role labeling.

## 3 Semantics-Driven Shallow Parsing

### 3.1 Motivation

There are two main jobs of semantic chunking: 1) grouping words as argument candidate and 2) classifying semantic types of possible arguments. Previously proposed shallow parsing only considers syntactic information and basic chunks are usually too small to effectively group words. This causes one main deficiency of semantic chunking. E.g. the argument "为三峡工程/for the Sanxia Project" consists of three chunks, each of which only consists of one word. To rightly recognize this *A2*, Semantic chunker should rightly predict three chunk labels. Small chunks also make the important "path" feature sparse, since there are more chunks between a target chunk and the predicate in focus. In this section, we introduce a new chunk definition to improve shallow parsing based SRL, which takes both syntactic and predicate-argument structures into account. The key idea is to make syntactic chunks as large as possible for semantic chunking. The formal definition is as follows.

### 3.2 Chunk Bracketing

Given a sentence $s = w_1, ..., w_n$, let $c[i:j]$ denote a constituent that is made up of words between $w_i$ and $w_j$ (including $w_i$ and $w_j$); let $p_v = \{c[i:j] | c[i:j] \text{ is an argument of } v\}$

|  | WORD | POS | TARGET | PROPOSITION |  |  |  | CHUNK 1 | CHUNK 2 |
|---|---|---|---|---|---|---|---|---|---|
| China | 中国 | NR | - | (A0* | * | * | * | B-NP | B-NP^S |
| tax | 税务 | NN | - | * | * | * | * | I-NP | I-NP^S |
| department | 部门 | NN | - | *) | * | * | * | I-NP | I-NP^S |
| stipulate | 规定 | VV | 规定 | (V*) | * | * | * | O | O |
| : | ： | PU | - | * | * | * | * | O | O |
| owing | 欠缴 | VV | 欠缴 | (A1* | (V*) | * | (A0* | O | O |
| tax payment | 税款 | NN | - | * | (A1*) | * | * | B-NP | B-NP^VP |
| company | 企业 | NN | - | * | (A0*) | * | * | B-NP | B-NP^NP |
| *Function Word* | 的 | DEG | - | * | * | * | * | O | O |
| leaders | 领导人 | NN | - | * | * | * | *) | B-NP | B-NP^NP |
| not | 不 | AD | - | * | * | * | (AM-ADV*) | B-ADVP | B-ADVP^VP |
| can | 得 | VV | 得 | * | * | (V*) | * | O | O |
| leave the country | 出境 | VV | 出境 | *) | * | * | (V*) | B-VP | B-VP^VP |

Figure 2: An example for definition of semantics-driven chunks with IOB2 representation.

denote one predicate-argument structure where $v$ is the predicate in focus. Given a syntactic tree $\mathcal{T}_s = \{c[i:j] | c[i:j]$ is a constituent of $s\}$, and its all argument structures $\mathcal{P}_s = \{p_v |$ v is a verbal predicate in $s\}$, there is one and only one chunk set $\mathcal{C} = \{c[i:j]\}$ s.t.

1. $\forall c[i:j] \in \mathcal{C}, c[i:j] \in \mathcal{T}_s$;
2. $\forall c[i:j] \in \mathcal{C}, \forall c[i^v : j^v] \in \cup\mathcal{P}_s, j < i^v$ or $i > j^v$ or $i^v \leq i \leq j \leq j^v$;
3. $\forall c[i:j] \in \mathcal{C}$, the parent of $c[i:j]$ does not satisfy the condition 2.
4. $\forall \mathcal{C}'$ satisfies above conditions, $\mathcal{C}' \subset \mathcal{C}$.

The first condition guarantees that every chunk is a constituent. The second condition means that chunks do not overlap with arguments, and further guarantees that semantic chunking can recover all arguments with the last condition. The third condition makes new chunks as big as possible. The last one makes sure that $\mathcal{C}$ contains all sub-components of all arguments. Figure 2 is an example to illustrate our new chunk definition. For example, "中国/Chinese 税务/tax 部分/department" is a constituent of current sentence, and is also an argument of "规定/stipulate". If we take it as a chunk, it does not conflict with any other arguments, so it is a reasonable syntactic chunk. For the phrase "欠缴/owing 税款/tax payment", though it does not overlap with the first, third and fourth propositions, it is bigger than the argument "税款" (conflicting with condition 2) while labeling the predicate "欠缴", so it has to be separated into two chunks. Note that the third condition also guarantees the constituents in $\mathcal{C}$ does not overlap with each other since each one is as large as possible. So we can still formulate our new shallow parsing as an "IOB" sequence labeling problem.

## 3.3 Chunk Type

We introduce two types of chunks. The first is simply the phrase type, such as *NP*, *PP*, of current chunk. The column *CHUNK 1* illustrates this kind of chunk type definition. The second is more complicated. Inspired by (Klein and Manning, 2003), we split one phrase type into several subsymbols, which contain category information of current constituent's parent. For example, an *NP* immediately dominated by a *S*, will be substituted by *NP^S*. This strategy severely increases the number of chunk types and make it hard to train chunking models. To shrink this number, we linguistically use a cluster of CTB phrasal types, which was introduced in (Sun and Sui, 2009). The column *CHUNK 2* illustrates this definition. E.g., *NP^S* implicitly represents *Subject* while *NP^VP* represents *Object*.

## 3.4 New Path Features

The *Path* feature is defined as a chain of base phrases between the token and the predicate. At both ends, the chain is terminated with the POS tags of the predicate and the headword of the token. For example, the path feature of "保险公司" in Figure 1 is "公司-ADVP-PP-NP-NP-VV". Among all features, the "path" feature contains more structural information, which is very important for SRL. To better capture structural information, we introduce several new "path" features. They include:

- NP|PP|VP path: only syntactic chunks that takes tag *NP*, *PP* or *VP* are kept.

When labeling the predicate “出境/leave the country” in Figure 2, this feature of “中国税务部门/Chinese tax departments” is *NP+NP+NP+NP+VP*.

- V|的 path: a sequential container of POS tags of verbal words and “的”; This feature of “中国税务部门” is *NP+VV+VV+的+VV+VP*.

- O2POS path: if a word occupies a chunk label *O*, use its POS in the path feature. This feature of “中国税务部门” is *NP+VV+PU+VV+NP+NP+DEG+ADVP+VV+VP*.

## 4 Experiments and Analysis

### 4.1 Experimental Setting

Experiments in previous work are mainly based on CPB 1.0 and CTB 5.0. We use CoNLL-2005 shared task software to process CPB and CTB. To facilitate comparison with previous work, we use the same data setting with (Xue, 2008). Nearly all previous research on Chinese SRL evaluation use this setting, also including (Ding and Chang, 2008, 2009; Sun et al., 2009; Sun, 2010). The data is divided into three parts: files from chtb_081 to chtb_899 are used as training set; files from chtb_041 to chtb_080 as development set; files from chtb_001 to chtb_040, and chtb_900 to chtb_931 as test set. Both syntactic chunkers and semantic chunkers are trained and evaluated by using the same data set. By using CPB and CTB, we can extract gold standard semantics-driven shallow chunks according to our definition. We use this kind of gold chunks automatically generated from training data to train syntactic chunkers.

For both syntactic and semantic chunking, we used conditional random field model. Crfsgd[1], is used for experiments. Crfsgd provides a feature template that defines a set of strong word and POS features to do syntactic chunking. We use this feature template to resolve shallow parsing. For semantic chunking, we implement a similar *one-stage* shallow parsing based SRL system described in (Sun et al., 2009). There are two differences between our system and Sun et al.’s system. First, our system uses Start/End method to represent semantic chunks (Kudo and Matsumoto, 2001). Second, word formation features are not used.

| Test | P(%) | R(%) | $F_{\beta=1}$ |
|---|---|---|---|
| (Chen et al., 2006) | 93.51 | 92.81 | 93.16 |
| Overall (C1) | 91.66 | 89.13 | 90.38 |
| Bracketing (C1) | 92.31 | 89.72 | 91.00 |
| Overall (C2) | 88.77 | 86.71 | 87.73 |
| Bracketing (C2) | 92.71 | 90.55 | 91.62 |

Table 1: Shallow parsing performance.

### 4.2 Syntactic Chunking Performance

Table 1 shows the performance of shallow syntactic parsing. Line *Chen et al., 2006* is the chunking performance evaluated on syntactic chunk definition proposed in (Chen et al., 2006). The second and third blocks present the chunking performance with new semantics-driven shallow parsing. The second block shows the overall performance when the first kind of chunks type is used, while the last block shows the performance when the more complex chunk type definition is used. For the semantic-driven parsing experiments, we add the *path* from current word to the first verb before or after as two new features. Line *Bracketing* evaluates the word grouping ability of these two kinds of chunks. In other words, detailed phrase types are not considered. Because the two new chunk definitions use the same chunk boundaries, the fourth and sixth lines are comparable. There is a clear decrease between the traditional shallow parsing (Chen et al., 2006) and ours. We think one main reason is that syntactic chunks in our new definition are larger than the traditional ones. An interesting phenomenon is that though the second kind of chunk type definition increases the complexity of the parsing job, it achieves better bracketing performance.

### 4.3 SRL Performance

Table 2 summarizes the SRL performance. Line *Sun et al., 2009* is the SRL performance reported in (Sun et al., 2009). To the author’s knowledge, this is the best published SRL result in the literature. Line *SRL (Chen et al., 2006)* is the SRL performance of our system. These two systems are both evaluated by using syntactic chunking defined in (Chen et al., 2006). From the first block we can see that our semantic chunking system reaches the state-of-the-art. The second and third blocks in Table 2 present the performance with

[1] http://leon.bottou.org/projects/sgd

new shallow parsing. Line *SRL (C1)* and *SRL (C2)* show the overall performances with the first and second chunk definition. The lines following are the SRL performance when new “path” features are added. We can see that new “path” features are useful for semantic chunking.

| Test | P(%) | R(%) | F$_{\beta=1}$ |
|---|---|---|---|
| (Sun et al., 2009) | 79.25 | **69.61** | 74.12 |
| SRL [(Chen et al., 2006)] | **80.87** | 68.74 | **74.31** |
| SRL [C1] | 80.23 | 71.00 | 75.33 |
| + NP\|PP\|VP path | 80.25 | 71.19 | 75.45 |
| + V\|的 path | 80.78 | 71.67 | 75.96 |
| + O2POS path | 80.44 | 71.59 | 75.76 |
| + All new path | 80.73 | 72.08 | 76.16 |
| SRL [C2] | 80.87 | 71.86 | 76.10 |
| + All new path | **81.03** | **72.38** | **76.46** |

Table 2: SRL performance on the test data. Items in the first column *SRL [(Chen et al., 2006)]*, *SRL [C1]* and *SRL [C2]* respetively denote the SRL systems based on shallow parsing defined in (Chen et al., 2006) and Section 3.

## 5 Conclusion

In this paper we propose a new syntactic shallow parsing for Chinese SRL. The new chunk definition contains both syntactic structure and predicate-argument structure information. To improve SRL, we also introduce several new “path” features. Experimental results show that our new chunk definition is more suitable for Chinese SRL. It is still an open question what kinds of syntactic information is most important for Chinese SRL. We suggest that our attempt at semantics-driven shallow parsing is a possible way to better exploit this problem.

## Acknowledgments

The author is funded both by German Academic Exchange Service (DAAD) and German Research Center for Artificial Intelligence (DFKI).

The author would like to thank the anonymous reviewers for their helpful comments.

## References

Wenliang Chen, Yujie Zhang, and Hitoshi Isahara. 2006. An empirical study of Chinese chunking. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 97–104. Association for Computational Linguistics, Sydney, Australia.

Weiwei Ding and Baobao Chang. 2008. Improving Chinese semantic role classification with hierarchical feature selection strategy. In *Proceedings of the EMNLP 2008*, pages 324–333. Association for Computational Linguistics, Honolulu, Hawaii.

Weiwei Ding and Baobao Chang. 2009. Fast semantic role labeling for Chinese based on semantic chunking. In *ICCPOL ’09: Proceedings of the 22nd International Conference on Computer Processing of Oriental Languages. Language Technology for the Knowledge-based Economy*, pages 79–90. Springer-Verlag, Berlin, Heidelberg.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL-2009), June 4-5*. Boulder, Colorado, USA.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 423–430. Association for Computational Linguistics, Sapporo, Japan.

Taku Kudo and Yuji Matsumoto. 2001. Chunking with support vector machines. In *NAACL ’01: Second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies 2001*, pages 1–8. Association for Computational Linguistics, Morristown, NJ, USA.

Weiwei Sun. 2010. Improving Chinese semantic role labeling with rich features. In *Proceedings of the ACL 2010*.

Weiwei Sun and Zhifang Sui. 2009. Chinese function tag labeling. In *Proceedings of the 23rd Pacific Asia Conference on Language, Information and Computation*. Hong Kong.

Weiwei Sun, Zhifang Sui, Meng Wang, and Xin Wang. 2009. Chinese semantic role labeling with shallow parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 1475–1483. Association for Computational Linguistics, Singapore.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The conll 2008 shared task on joint parsing of syntactic and semantic dependencies. In *CoNLL 2008: Proceedings of the Twelfth Conference on Computational Natural Language Learning*, pages 159–177. Coling 2008 Organizing Committee, Manchester, England.

Nianwen Xue. 2008. Labeling Chinese predicates with semantic roles. *Comput. Linguist.*, 34(2):225–255.