Labeling Chinese Predicates with Semantic Roles

Nianwen Xue*
University of Colorado at Boulder

In this article we report work on Chinese semantic role labeling, taking advantage of two recently completed corpora, the Chinese PropBank, a semantically annotated corpus of Chinese verbs, and the Chinese Nombank, a companion corpus that annotates the predicate–argument structure of nominalized predicates. Because the semantic role labels are assigned to the constituents in a parse tree, we first report experiments in which semantic role labels are automatically assigned to hand-crafted parses in the Chinese Treebank. This gives us a measure of the extent to which semantic role labels can be bootstrapped from the syntactic annotation provided in the treebank. We then report experiments using automatic parses with decreasing levels of human annotation in the input to the syntactic parser: parses that use gold-standard segmentation and POS-tagging, parses that use only gold-standard segmentation, and fully automatic parses. These experiments gauge how successful semantic role labeling for Chinese can be in more realistic situations. Our results show that when hand-crafted parses are used, semantic role labeling accuracy for Chinese is comparable to what has been reported for the state-of-the-art English semantic role labeling systems trained and tested on the English PropBank, even though the Chinese PropBank is significantly smaller in size. When an automatic parser is used, however, the accuracy of our system is significantly lower than the English state of the art. This indicates that an improvement in Chinese parsing is critical to high-performance semantic role labeling for Chinese.

1. Introduction

Semantic role labeling (SRL) is the task of identifying arguments for a predicate and assigning semantically meaningful labels to them. A semantic role represents a semantic relation between a predicate and one of its arguments. Typical semantic roles include agent, patient, source, goal, and so forth, that are core to a predicate, as well as location, time, manner, cause, and so on, that are peripheral. Such semantic information is important in answering *who*, *what*, *when*, *where*, and *why* questions therefore is crucial to natural language processing (NLP) tasks such as question-answering (Narayanan and Harabagiu 2004), information extraction (Surdeanu et al. 2005), summarization (Melli et al. 2005), and machine translation (Boas 2002). Any NLP task that requires some form of semantic interpretation could potentially benefit from a high performance semantic role labeling system.

For an automatic system, a semantic role labeling task involves locating the linguistic units, typically words or phrases, in natural language text that serve as arguments

* 1777 Exposition Drive, Boulder, CO 80309. E-mail: Nianwen.Xue@colorado.edu.

Submission received: 15 July 2006; revised submission received: 4 May 2007; accepted for publication: 19 June 2007.

to a predicate and assigning semantic role labels to them based on the context in which they occur. Since the seminal work of Gildea and Jurafsky (2002), statistical and machine learning approaches have been the predominant research paradigm in semantic role labeling, like most of the subfields in natural language processing and computational linguistics. A prerequisite for statistical and machine learning approaches to semantic role labeling is the availability of a significant amount of semantically interpreted corpora from which automatic systems can learn. The recent activities in semantic role labeling (Carreras and Màrquez 2004b, 2005; Litkowski 2004) have in large part been driven by the availability of semantically annotated corpora such as the FrameNet (Baker, Fillmore, and Lowe 1998), Proposition Bank (Palmer, Gildea, and Kingsbury 2005), and Nombank (Meyers et al. 2004) projects for English; the tectogrammatical layer annotation of the Prague Dependency Treebank (Sgall, Panevová, and Hajičová 2004) for Czech; and the Salsa Project for German (Burchardt et al. 2006). These semantically annotated corpora not only provide the training and test material for the development of machine learning systems, but also effectively define semantic role labeling as a task.

PropBank and FrameNet have been the two most widely used corpora in developing automatic semantic role labeling systems. Although both corpora provide predicate–argument structure annotation, they use very different semantic role labels, especially for the core arguments of each predicate. In FrameNet, the semantic roles of a predicate (called a **Lexical Unit (LU)**) are organized by **semantic frames**, which are conceptual structures that describe a particular situation or event along with their participants, which are called **frame elements (FEs)**. All LUs in the same semantic frame share one set of semantic roles. For example, the verbs *buy* and *sell* both belong to the semantic frame *Commercial_transaction*, which involves a *Buyer* and *Seller* exchanging *Money* and *Goods*. In addition to these four *core* FEs, there are also three Non-Core FEs: *Means*, the manner in which the transaction takes place; *Rate*, the price of payment per unit of *Goods*; and *Unit*, the unit of measure for the *Goods*. Semantic role labeling based on FrameNet annotation attempts to identify the syntactic constituents in a sentence and assign FEs to them (1a). Notice that for any given sentence, not all FEs have to be realized and they do not have to be realized in the same syntactic position.

(1) a. FrameNet
[**Buyer** We] always [**LU** bought] [**Goods** a few dark-red carnations] [**Seller** from her]
During the later part of the nineteenth century, [**Seller** the landowners] [**LU** sold] [**Goods** the land] [**Buyer** to developers] in very small lots.

b. PropBank
[**Arg0** We] always [**Rel** bought] [**Arg1** a few dark-red carnations] [**Arg2** from her]
During the later part of the nineteenth century, [**Arg0** the landowners] [**Rel** sold] [**Arg1** the land] [**Arg2** to developers] in very small lots.

Like FrameNet, PropBank also assigns semantic role labels to syntactic constituents (rather than to the heads in a dependency structure) in a sentence. Unlike FrameNet, there is no reference ontology like the semantic frame that provides a general set of semantic roles. Instead, for the core arguments, the PropBank uses a set of predicate-specific semantic role labels represented by an integer prefixed by *Arg*: *Arg0* through *Arg5*. Predicates vary on the number of core arguments they take, but generally the total number of core arguments does not exceed six. These core arguments are defined in **frame files**, with one frame file for each predicate. Within a frame file, the core

arguments are organized by **framesets**, which are the major senses of a predicate. A new frameset is postulated only when it takes a different set of core arguments from existing framesets. In addition to the core arguments, there is also a finite set of roles reserved for adjunct-like arguments. Each adjunct-like argument is represented as *ArgM*, indicating that it is a modifier argument, followed by a secondary tag indicating the type of modifier. Secondary tags are for semantic information such as location, manner, and time that are not specific to a particular verb or even a particular class of verbs and they are defined based on a general set of guidelines. There is thus a dichotomy in the representation of the semantic roles for the core and peripheral arguments in the PropBank annotation.

The predicate-specific nature of the PropBank semantic roles is clear when compared with the FrameNet FE. In (1b), for example, the seller is always labeled *Seller* and the buyer is always labeled *Buyer* in the FrameNet annotation whether the predicate is *buy* or *sell*. In contrast, in the PropBank annotation, the buyer is *Arg0* when the predicate is *buy* and *Arg2* when the predicate is *sell*. Conversely, the seller is *Arg0* when the predicate is *sell* and *Arg2* when the predicate is *buy*. While FrameNet annotates the semantic roles of both verbal and nominal predicates, the annotation of PropBank is limited to verbs, with the nominal predicates annotated in a separate but related project, the Nombank Project (Meyers et al. 2004). The NomBank Project adopted the same predicate-specific approach in representing the core arguments of a predicate as PropBank, with special treatment for noun-specific phenomena such as support verbs.

There is considerable work on English semantic role labeling with both annotation conventions. Gildea and Jurafsky (2002) did their seminal work using data from FrameNet. The Senseval-3 international competition on semantic role labeling (Litkowski 2004) also used the FrameNet annotation. There is an even larger body of work using PropBank because it has a larger amount of annotated data on a well-established data set, the Penn Treebank (Marcus, Santorini, and Marcinkiewicz 1993). Using the standard training and test sets in the Penn Treebank, there has been a rapid improvement in performance due to the use of more advanced machine-learning techniques and more informative linguistic features. The performance using automatic parses on Section 23 of the Penn Treebank has approached 0.81 F-score (Pradhan, Ward et al. 2005). There have also been two consecutive CoNLL competitions (Carreras and Màrquez 2004b, 2005) on semantic role labeling using the PropBank data.

Research on Chinese semantic role labeling is still in its infancy. Work on Chinese semantic role labeling has been scant and sporadic, mostly due to the lack of a publicly available semantically annotated corpus of significant size. Although most of the machine-learning techniques used in English semantic role labeling are readily transferable to Chinese, such technological transfer is only possible with similarly annotated data. To our knowledge, there are only two such data sets, which all used a small corpus that the authors created on their own. Sun and Jurafsky (2004) did preliminary work on Chinese semantic role labeling on 10 selected verbs using Support Vector Machines and reported promising early results.[1] Noting that Chinese syntactic parsing is an especially challenging task, Kwong and T'sou (2005) reformulated semantic role labeling as a task of detecting and classifying the heads of arguments to avoid the hard problem of getting the correct text spans for the arguments. In this article, we

1 They restated their results in Chen, Sun, and Jurafsky (2005) due to an error in retraining the Collins parser (Collins 1999) on Chinese, which led to inflated Chinese syntactic parsing and therefore semantic role labeling results.

report work on the semantic role labeling of Chinese predicates for both verbs and their nominalizations, exploiting two recently completed corpora, the Chinese PropBank (Xue and Palmer 2003), a corpus that annotates the predicate–argument structure of verbs, and the Chinese NomBank (Xue 2006a), a companion corpus that annotates the predicate-argument structure of nominalized predicates in Chinese. Both corpora are built on top of the Chinese Treebank (Xue et al. 2005), in the sense that the semantic role labels are assigned to constituents in the parse tree.

The Chinese PropBank and Nombank adopted the English PropBank predicate-specific approach in representing the semantic roles of the core arguments. In the absence of a Chinese linguistic ontology like the semantic frames developed for the English FrameNet Project, using the PropBank-style of semantic roles allows faster development. The predicate-specific approach of the PropBank annotation builds a solid foundation for making high-level generalizations in a bottom-up manner, if broader generalizations are needed. The Chinese PropBank focuses on the context-sensitive component of the semantic role annotation, using frame files to guide its annotation. The semantic roles defined in the frame files are for *expected* arguments, that is, all possible arguments for each frameset of a predicate. In a particular sentence, an expected argument may not always be realized, and when it is, it may not always be realized in the same syntactic position as a result of syntactic alternations (Levin 1993) or other syntactic processes. In addition, different framesets of a verb take different sets of arguments that demonstrate different syntactic patterns. Thus, predicate–argument structure analysis at the PropBank annotation level represents a crucial leap towards proper representation of semantic structure from the syntactic structure. Should the need for more general semantic roles arise, these predicate-specific semantic roles can be mapped (Yi, Loper, and Palmer 2007) to FrameNet-style or even VerbNet-style (Kipper et al. 2006) labels.

Using the semantic annotation of the Chinese PropBank and NomBank as training and test material, we were able for the first time to develop a Chinese semantic role labeling system that is trained and tested on semantically annotated Chinese corpora of significant sizes. Using parses produced with different levels of automation (a fully automatic parser, a parser with correct segmentation, a parser with both correct segmentation and POS-tagging, and treebank gold-standard parses), we were able to quantify the impact of different Chinese language processing components on the semantic analysis of Chinese predicates. Using a Maximum Entropy–based (McCallum 2002) machine learning system, our experimental results show that just by using the features reported in the English semantic role labeling literature, our baseline system achieved a very high accuracy on Chinese verbs when the gold-standard treebank parses are used. This suggests that these features port very well between English and Chinese. There is a gradual degradation in semantic role labeling performance with a decreasing level of human annotation (from gold-standard treebank parses to fully automatic parses). We were able to achieve a modest improvement with additional features tailored to the Chinese language, bringing the overall accuracy to F1-scores of 0.92 and 0.67, respectively, when using treebank and fully automatic parses. We were able to achieve a larger improvement on the semantic role labeling of nominalized predicates by using noun-specific features (F-scores of 0.70 and 0.57, respectively, for treebank and fully automatic parses), but our results still show that the semantic role labeling of nominalized predicates is a much more challenging task than that of verbs. This is partly due to the smaller training set for nominalized predicates, with the number of nominalized predicates being less than one third of the number of verb instances in the same corpus. More importantly, the arguments of nominalized predicates have a

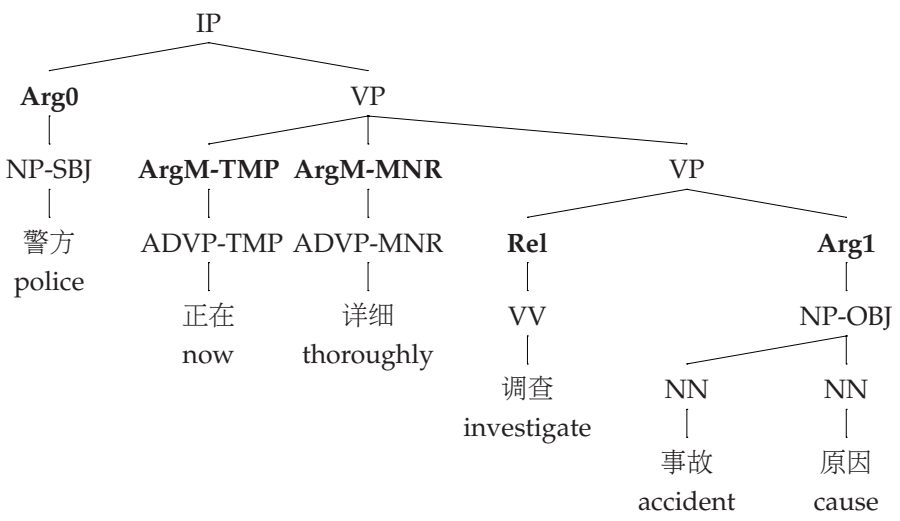
much more uneven distribution: Arguments of a nominalized predicate can occur either inside the NP headed by the predicate or outside when a support verb is present (see Section 2 for examples). This makes it particularly challenging to determine whether a constituent in the parse tree is an argument or not. This observation is supported by the large margin in performance between the semantic role labeling results we achieved when the constituents are known and unknown.

This article is structured as follows. In Section 2, we discuss the semantic annotation of the Chinese PropBank and NomBank in greater detail. In Section 3, we describe the general architecture of our system, focusing on shared components for both verbs and their nominalizations. In Sections 4 and 5 we present our experiments on verbs and nouns, respectively. Section 6 discusses related work and Section 7 concludes this article and discusses future work.

2. The Chinese PropBank and NomBank

The Chinese PropBank and the Chinese NomBank adopt the descriptive framework of the English Proposition Bank in which semantic arguments and adjuncts are treated differently. The semantic arguments of a predicate are labeled with a contiguous sequence of integers, in the form of *ArgN*, where *N* is an integer between 0 and 5. These labels can only be interpreted in the context of a specific predicate. In other words, these argument labels are not meaningful without knowing what the predicate is. In fact, as we will show later in this section, these numbered labels are meaningful only within a particular sense of a predicate. In general, like English, a Chinese predicate takes fewer than 6 arguments. The assignment of numbered argument labels is illustrated in Example (2), where the predicate is the verb 调查 ("investigate"). Its subject 警方 ("the police") is labeled *Arg0* and its object 事故 ("accident") 原因 ("cause") *Arg1*. The semantic role labels added to the parse tree are in bold.

(2)



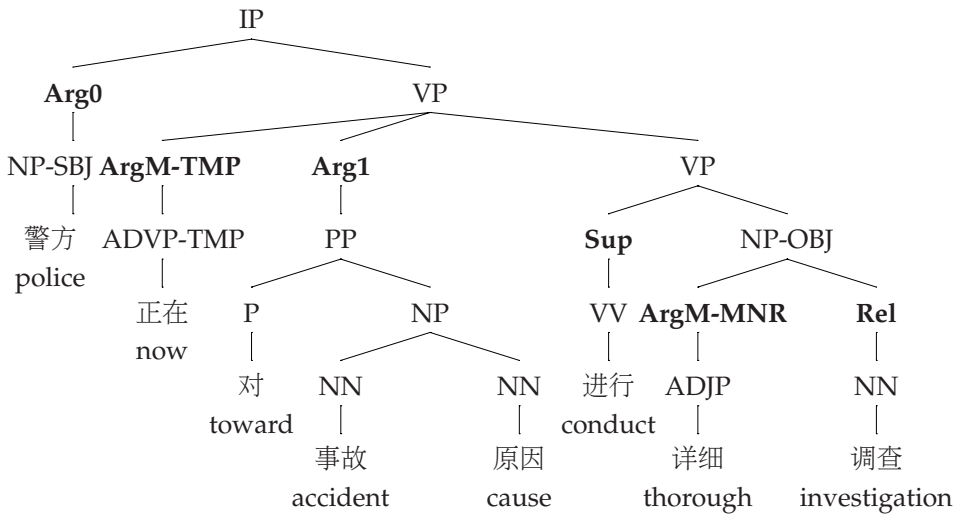
"The police are thoroughly investigating the cause of the accident."

The semantic adjuncts, on the other hand, are annotated as such with the label *ArgM* followed by a secondary tag that represents the semantic classification of the adjunct. Unlike the numbered argument labels for semantic arguments, the secondary

tags represent information that is not predicate-specific. For instance, the adverbial modifiers 正在 ("right now") and 详细 ("thorough") in Example (2) are labeled *ArgM-TMP* and *ArgM-MNR* respectively, where the secondary tag *TMP* indicates a temporal modifier and *MNR* indicates manner. The secondary tags are not predicate-specific in the sense that they are not required by this particular predicate and they are not selective with regard to the predicate they can occur with. There is a limited set of such secondary tags that are used in the Chinese PropBank and the Chinese NomBank and the complete list of such secondary tags is presented in Table 1.

The same approach is taken to annotate the nominalized predicates in the Chinese NomBank. This is illustrated in Example (3), a close paraphrase of Example (2), where the nominalized predicate 调查 ("investigation") takes the same arguments as its verbal counterpart. 警方 ("the police") is again *Arg0* and 对 ("toward") 事故 ("accident") 原因 ("cause"), despite its categorial change from a noun phrase to a prepositional phrase, remains *Arg1*. There are also two semantic adjuncts: *ArgM-TMP* 正在 ("now") and *ArgM-MNR* 详细 ("thorough"). It is worth noting, however, that in this particular case, the nominalized predicate needs a support verb 进行 ("conduct") to satisfy the grammatical constraint that there be a verb in the sentence, and it is explicitly marked as *Sup*. In addition, despite the categorial change of 详细 from adverb to adjective, the semantic role label remains unchanged. In this sense, the semantic annotation abstracts away from the underlying syntactic annotation.

(3)



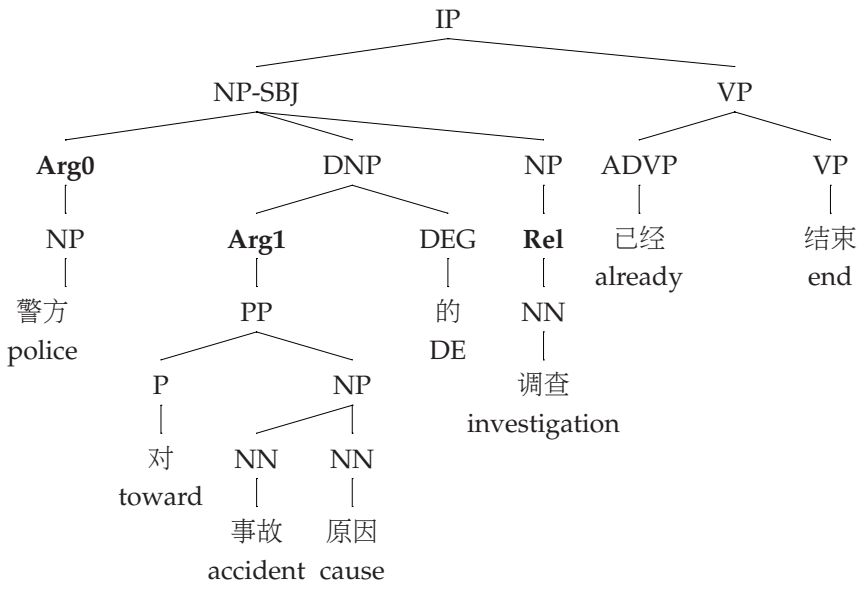
"The police are conducting a thorough investigation of the cause of the accident."

Table 1
The complete list of functional tags defined in the Chinese Propbank and NomBank.

ADV	adverbial	FRQ	frequency
BNF	beneficiary	LOC	locative
CND	condition	MNR	manner
DIR	direction	PRP	purpose or reason
DIS	discourse marker	TMP	temporal
DGR	degree	TPC	topic
EXT	extent		

Not all occurrences of a nominalized predicate need to be accompanied by a support verb. In fact, it is often the case that all arguments of a nominalized predicate occur in a noun phrase headed by the nominalized predicate. For example, in Example (4), both *Arg0* 警方 ("police") and *Arg1* 对 ("toward") 事故 ("accident") 原因 ("cause") are syntactic modifiers of the nominalized predicate 调查 ("investigation").

(4)



"The police investigation of the cause of the accident has ended."

The PropBank-style annotation is designed to account for syntactic variations, that is, the different ways in which the same predicate–argument structure is realized. In Examples (2) and (3), we have already seen where essentially the same predicate is realized as a verb or a noun, and its arguments are realized as different syntactic categories in different syntactic positions. Syntactic variations occur even without the categorial change of the predicate. Levin (1993) demonstrates extensively how the argument structure of English verbs can be realized differently through diathesis alternations. Similar alternations can also be observed in Chinese, and Example (5) shows this:

(5) a. **[Arg1** 中 美 交往 的 大门 **]** **[rel** 打开 **]** 了 。
China the U.S. contact DE door open ASP .
"The door of contact between China and the U.S. has opened."

b. **[ArgM-TMP** ７０年代初 **]**，**[Arg0** 中 美 两 国 领导人
70s beginning , China the U.S. two country leader
] **[ArgM-ADV** 果断 **]** **[rel** 打开 **]** 了 **[Arg1** 中 美 交往 的
decisively open ASP China the U.S. contact DE
大门 **]**。
door .
"In the beginning of the 1970s, the leaders of China and the U.S. decisively opened the door of contact between China and the U.S."

Note that even though 中 ("China") 美 ("the U.S.") 交往 ("contact") 的 ("DE") 大门 ("door") occurs in different syntactic positions in (5a) and (5b), it is labeled *Arg1* in both cases. The semantic role label an argument receives is independent of its syntactic

realizations. The semantic roles or expected arguments can be realized syntactically in different ways. It should also be pointed out that the line drawn between arguments and adjuncts here is not based on the obligatory/optional dichotomy. In some cases, some constituents are clearly arguments but they are also clearly optional. For example, in the unaccusative (or pseudo-passive) construction, the agent is clearly optional syntactically and it is equally clear that it is an argument. In Example (5a), for example, the "door-opener" is optional but is clearly an argument.

The Chinese PropBank also adds a coarse-grained sense tag to the predicate. The senses of a predicate are motivated by the argument structure of this predicate and are thus an integral part of the predicate–argument structure annotation. Sense disambiguation is performed only when different senses of a predicate require different sets of arguments. For example, the "evolve" sense of the verb "发展" expects five arguments: The cause of the evolution, which is often not realized, the entity evolving, the starting point of the evolution, the end point of the evolution, and the range of the evolution. When it means "recruit," two arguments are expected: the recruiter and the entity recruited. Because each of these senses can be realized in different subcategorization frames related through syntactic alternations, in the PropBank annotation convention, these senses are formally called **framesets**, meaning sets of subcategorization frames that realize a particular sense. The examples in (6) illustrate two of the framesets of "发展".

(6) **Frameset 1**: "evolve"
Semantic roles:
Arg0: cause of evolution
Arg1: entity evolving
Arg2: evolving from
Arg3: evolving to
Arg4: range of evolution

a. [**Arg1** 俄罗斯 国内 对 工业品 需求] [**Arg3** 向
Russia domestic for industrial product demand in
中高档 方向] [**Rel** 发展] 。
mid- and upper scale direction develop .
"Russia's domestic demand for industrial products is evolving in the direction of mid- and upper scale products."

Frameset 2: "recruit"
Arg0: recruiter
Arg1: entity recruited

b. [**Arg0** ＣＥＣ] [**ArgM-TMP** 目前] [**ArgM-LOC** 在世界 一百六十个
CEC presently in the world 160 CL
国家 和 地区] [**Rel** 发展] 了 [**Arg1** 八千多万 个
country and region recruited ASP more than eight thousand CL
用户] 。
subscriber .
"CEC presently have recruited over eight thousand subscribers from 160 countries and regions in the world."

The Chinese NomBank uses the same framesets as defined for verbs because its annotation is guided by the same frame files. However, typically only a subset of the framesets for verbs have corresponding nominalized forms. For example, Frameset 1 in

Example (6) has a corresponding nominalized form as illustrated in Example (7), but Frameset 2 does not.

(7) 海峡 两 岸 今后 可 共同 规划 [**Arg1** 两岸
Taiwan Straits two side from now on can together plan cross-Strait
关系] 的 [**Rel** 发展] 。
relations DE development .
"The two sides of the Taiwan Straits can plan the development of the cross-Strait relations hereafter."

3. System Overview

Assuming the availability of a parse tree (either hand-crafted parses in a treebank or parses generated by an automatic parser) as input, to assign the semantic role labels described in Section 2 automatically involves first of all identifying which constituents in the parse tree are semantic arguments to the predicate in question and then assigning appropriate semantic role labels to them. The predominant approach to the semantic role labeling task is to formulate it as a classification problem (Pradhan, Ward et al. 2004; Xue and Palmer 2004) that can be solved with machine-learning techniques. One can imagine a classification task in which each constituent in the parse tree is labeled either with one of the numbered argument labels (*Arg0* through *Arg5*), or with one of the semantic adjunct labels *ArgM-TMP*, *ArgM-MNR*, and so on, or with the *NULL* label, indicating the constituent is neither an argument nor adjunct to the predicate. This simple formulation of the classification problem is rarely practiced in the semantic role labeling literature for the simple reason that the majority of the constituents in a parse tree are generally not related to the predicate in question. For machine-learning approaches, this means that the negative samples, constituents that are labeled *NULL*, would far outweigh the positive samples, constituents that are actual semantic arguments or adjuncts. Such an imbalance would lead to poor performance for machine-learning systems, so in practice, most semantic role labeling systems work in stages, which minimally consist of an **argument detection** stage and an **argument classification** stage. Argument detection is generally formulated as a binary classification task that separates constituents that are arguments or adjuncts to a predicate from those that are not related to the predicate in question. By lumping together argument and adjunct labels, the positive and negative sample imbalance is alleviated somewhat. In addition, it has been shown that argument detection and argument classification need different sets of features (Xue and Palmer 2004). A system cannot take advantage of this if both are done in one fell swoop. With a powerful machine-learning algorithm, argument detection can be done with high accuracy (Hacioglu et al. 2004; Pradhan, Ward et al. 2004), provided that the appropriate features are used.

The positive and negative sample imbalance can only be partially addressed by having a separate argument detection stage. Even with a binary classification task, the number of negative samples is still overwhelmingly larger than the positive samples. In addition, it does not take advantage of the fact that the arguments and adjuncts of a predicate, verbal or nominalized, are related to the predicate itself in linguistically well-understood structural configurations. The overwhelming majority of the arguments and adjuncts are populated along the *spine* of the parse tree that the predicate projects. A substantial number of the constituents can be eliminated from further consideration as negative samples with a high degree of certainty. (See Sections 4.1 and 5.1 for an

evaluation of how effectively the pruning algorithm works for verbs and nouns.) This was proved to be a successful strategy by Xue and Palmer (2004) for the semantic role labeling of English verbs; they use a heuristic algorithm to first prune out irrelevant constituents before the remaining candidates are fed into an argument detection algorithm. This strategy has also been effectively adopted by others (Cohn and Blunsom 2005; Punyakanok, Roth, and Yih 2005) and is used here. As we will show in later sections, the pruning algorithm needs to be slightly different for verbal and nominal predicates and they do not work equally well for all experimental conditions. Generally, the efficacy of the pruning algorithm correlates with the quality of syntactic parses that the semantic role labeling system takes as input. That is, it works much better with treebank parses than with automatic parses. It also works more effectively for verbs than for nouns, the arguments of which have a more diverse distribution.

Argument classification, which classifies the constituents into a category that corresponds to one of the semantic role labels, is a natural multi-category classification problem. It has been generally shown in the literature (Pradhan et al. 2003) that it is a good idea to bias the argument detection stage toward high recall so that reasonably good candidates can be passed along to the argument classification stage, and this means the tag set for argument classification also includes the *NULL* label. Many classification techniques—SVM (Pradhan, Ward et al. 2004), perceptrons (Carreras and Màrquez 2004a), Maximum Entropy (Xue and Palmer 2004), and so forth—have been successfully used to solve the semantic role labeling problem. In the work we report here, for both argument detection and classification tasks, we used a Maximum Entropy classifier with a tunable Gaussian prior in the Mallet Toolkit (McCallum 2002). The Maximum Entropy classifier does multi-category classification and thus can be straightforwardly applied to the problem here. The classifier can be tuned to minimize overfitting by adjusting the Gaussian prior.

In summary, in our system, the semantic role labeling is done in three stages, as illustrated in Figure 1: pruning, argument detection, and argument classification.

4. Semantic Role Labeling of Verbs

In this section we report our experiments on the semantic role labeling of Chinese verbs, using the Chinese PropBank as training and test material. There are two variables in our experimental settings. The first variable is the level of human annotation in the syntactic parses that serve as input to our semantic role labeling system. We used parses that are fully automatic, parses that assume correct segmentation, parses that assume correct segmentation as well as POS-tagging, and hand-crafted treebank parses. The second experimental variable is whether it is known which constituents in the parse

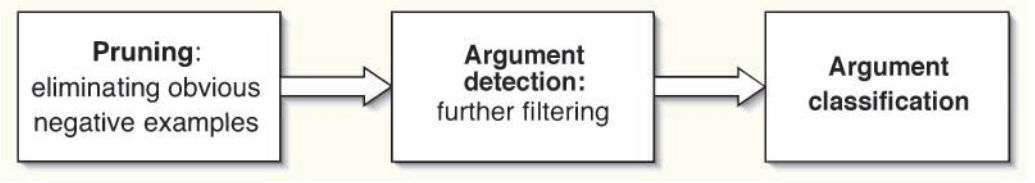


Figure 1
System architecture.

tree are arguments. If the constituents are known, the semantic role labeling reduces to a pure classification problem where each class is one of the semantic role labels. The semantic role labeling system only needs to determine what the correct semantic role should be. If it is unknown which constituents are arguments or adjuncts and which ones are irrelevant to the predicate in question, the system then needs to first figure out which constituents, out of all the constituents in the parse tree, are arguments to the predicate and then decide what the correct semantic role labels should be. We did not experiment with all combinations of these two variables and the known constituent experiment is only done for treebank parses.

In this section, we start by describing our pruning procedure for verbs in Section 4.1. We then present the features for our experiments on verbs in Section 4.2. In Section 4.3 we briefly describe the parsers we used for our experiments and we discuss our experimental results in Section 4.4.

4.1 Pruning for Verbs

Section 3 demonstrated the need for and the feasibility of using a heuristic algorithm to address the imbalance of positive and negative samples and in this section we show how this algorithm is implemented for verbal predicates. The algorithm starts from the predicate that anchors the annotation, and first collects all the syntactic complements of this predicate, which are represented as sisters to the predicate. It then iteratively moves one level up to the parent of the current node till it reaches the root of the tree. At each level, the system tries to determine whether or not that level is a coordination structure. The system only considers a constituent to be a potential candidate if it is a modifier or a complement to the current node. In the case of a coordination structure, the conjunct that the predicate does not occur in and all its children are eliminated as possible arguments to the predicate in question. Punctuation marks at all levels are ignored. It is worth pointing out that the functional tags and traces, which would have been useful for this purpose, are not used to determine the candidates to allow for a fair comparison between experiments on hand-crafted parses and parses generated by an automatic parser. Typically, current parsers do not predict functional tags and traces.[2] To use Example (8) as a walk-through example, assuming the predicate we are interested in is 调查 ("investigate"), the algorithm starts from this predicate and adds the NP 事故 ("accident") 原因 ("cause") to the candidate list because it is a complement to the predicate. Then it goes one level up to the VP and adds its two sisters, the ADVPs, 正在 ("right now") and 详细 ("carefully") to the candidate list because they are modifiers. Then the algorithm goes another level up to another VP, and determines that the two VPs at this level are conjoined by the punctuation mark, and no candidate is added at this level because it is a coordination structure. The algorithm then goes up to the highest VP level, and adds its sister, the NP 警方 ("police") to the list of candidates. The algorithm terminates at the highest IP[3] level. The candidates collected by this algorithm are in circles. The nodes traversed are linked by dotted lines.

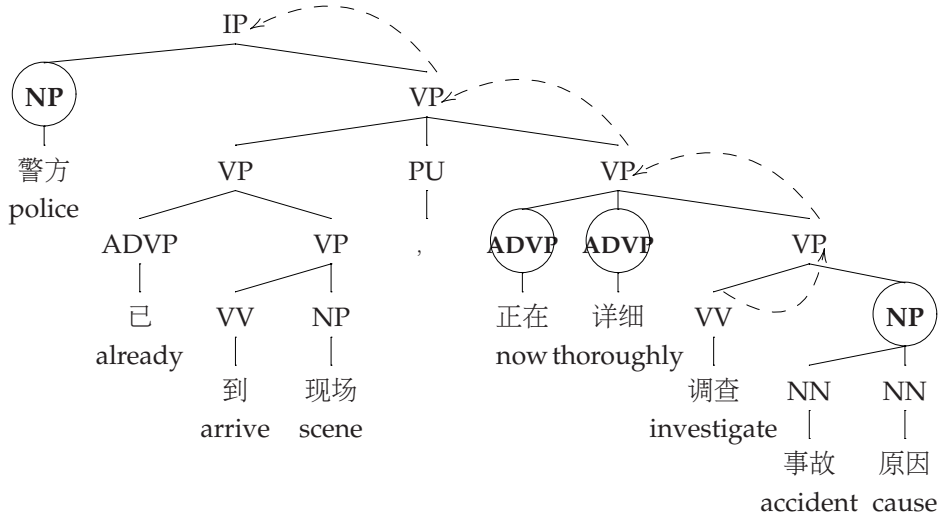
It is perhaps not surprising that the pruning algorithm works better with treebank parses than with automatic parses. When the treebank parses are used, our pruning

2 There are some ongoing efforts to develop parsers that produce functional tags and traces (Gabbard, Kulick, and Marcus 2006).

3 IP in the Chinese Treebank roughly corresponds to S in the Penn Treebank.

algorithm can recall over 99% (8,052 out of 8,121 arguments in the test data) of the arguments while pruning out over 93% (258,959 out of 276,734) nodes in the parse trees. When automatic parses (Maxent segmentation + Bikel parser) are used as input to our semantic role labeling system, out of 87% of the arguments that have a corresponding constituent in the parse tree, our pruning algorithm can recall 74% of the arguments while pruning out over 92% (247,530 out of 267,381) of the nodes in the parses. Our experiments show that even when the automatic parses are used, the results are far better when the pruning algorithm is used than when it is not used. If the pruning algorithm is not used, the recall improves somewhat, but the precision plummets. The less-than-expected drop in recall when the pruning algorithm is used is perhaps due to the fact that the arguments that are pruned out also happen to be the hardest for the semantic role labeling system to get right.

(8)



"The police have arrived at the scene and are thoroughly investigating the cause of the accident."

4.2 Features

One characteristic of feature-based semantic role modeling is that the feature space is generally large. This is in contrast with a low-level NLP task such as POS tagging, which generally has a small feature space. A wide range of features have been shown to be useful in previous work on semantic role labeling (Gildea and Jurafsky 2002; Pradhan, Ward et al. 2004; Xue and Palmer 2004) and we suspect that many more will be tested before the field will settle down to a core set of features. In their preliminary work on Chinese semantic role labeling, Sun and Jurafsky (2004) successfully ported a number of the features used in Gildea and Jurafsky (2002) to Chinese. In our experiments we adapted more features that have been described in recent work on English semantic role labeling to Chinese. We used a combination of features from Gildea and Jurafsky (G&J) (2002), Pradhan, Ward et al. (P et al.) (2004), and Xue and Palmer (X&P) (2004), and these are used as baseline features. In addition, we proposed a set of new features that used verb class information induced from the frame files of the Chinese PropBank, as well as features that were designed to exploit the grammatical constructions that

are unique to Chinese, specifically the BA (Bender 2000) and BEI (Huang 1999) constructions. We briefly discuss these features and where necessary explain at an intuitive level why they are useful for semantic role labeling. It has been well-established in the semantic role labeling literature that features are not equally effective for argument detection and argument classification (Xue and Palmer 2004). Our experimental results on Chinese semantic role labeling generally support this observation. The features we used for the semantic role labeling of verbs are listed below. Features that are marked as "C" are used only in the argument classification task. Features that are marked as "D" are for argument detection only. Features that are used in both argument detection and argument classification stages are marked as "C,D".

I. **Baseline Features:**

C *Position*: The position is defined in relation to the predicate verb and the values are *before* and *after*. (G&J)

C *Subcat frame*: The rule that expands the parent of the verb, for example, *VP→VV+NP*. (G&J)

C *Phrase type*: The syntactic category of the constituent in focus, for example, *NP*, *PP*. (G&J)

C *First and last word of the constituent in focus*. (P et al.)

C *Phrase type of the sibling to the left*. (P et al.)

C *Subcat frame+*: The subcat frame that consists of the NPs that surround the predicate verb. This feature is defined by the position of the constituent in focus in relation to this syntactic frame. (X&P)

C,D *Predicate*: The verb itself. (G&J)

C,D *Path*: The path between the constituent in focus and the predicate. (G&J)

C,D *Head word and its part of speech*: The head word and its part of speech are often good indicators of the semantic role of a constituent. (G&J)

C,D *Combination features*: Predicate head word combination, predicate phrase type combination. (X&P)

II. **New features**

C,D Path to BA and BEI: BA and BEI are function words that impact the order of the arguments. BA words are a closed set and in the Chinese Treebank they have the POS tag *BA*. Similarly, BEI words are also a closed set and they are POS-tagged *SB* (for short BEI) and *LB* (for long BEI).

C,D *Verb class*: Verb class itself, verb class + head word combination, verb class + phrase type combination.

The position feature is useful because constituents receiving a particular semantic role label may occur in some typical positions. For example, the majority of the adjuncts, *ARGM*s, occur before the verb in Chinese. The path feature, defined as the route from the constituent in focus to the predicate, represents a more "fine-grained" position. Whereas the values for the simple position feature are just *BEFORE* or *AFTER*, the values for the path feature can represent syntactic notions like *subject* or *object*. For example, a

subject may be represented as "NP↑IP↓VP↓VV" and an object may be represented as "VV↑VP↓NP." Intuitively, path features are more informative than simple position features but they are also sparse because they are more specific. The path feature proves to be particularly effective for the argument detection task, which is perhaps not unexpected. As we have shown in Section 4.1, the arguments and adjuncts of a predicate tend to be populated along the spine of a parse tree anchored by the predicate, and this information is captured very nicely by the path from the predicate to the constituent in question.

The head word and its part of speech are clearly informative for semantic role labeling. For example, a noun phrase headed by 今天 ("today") is very likely to be a temporal element; so is a prepositional phrase with the head word 在 ("at"). However, for prepositional phrases, the preposition is not always the most informative element. Sometimes the head word of its NP complement is more predictive of the semantic category. For example, in the prepositional phrase 在 ("at") 北京 ("Beijing"), the NP head 北京 ("Beijing") is more telling of the fact that it indicates a location. So for prepositional phrases we use both the preposition and the head noun as features in our system. As has been discussed by Sun and Jurafsky, the head word feature also tends to be sparse, especially given the smaller size of the Chinese Treebank. The chance of seeing a word in the test data that also occurs in the training data is small. The POS tag serves as one form of backoff: Constituents headed by words that have the same part-of-speech are likely to receive the same semantic role labels as well.

The subcat feature, as implemented in previous work (Gildea and Jurafsky 2002), is defined as the rule that expands the VP dominating the verbal predicate. By definition, it does not vary with the constituents in a parse tree. In other words, all constituents in the parse tree share the same subcat feature. To make up for the weakness of this feature, we implemented another feature called subcat+, which is the *syntactic frame* feature in Xue and Palmer (2004). The subcat+ feature heuristically identifies the key NP arguments for a given predicate, and the feature value of a given constituent is determined by its position in relation to these NPs and the predicate. In this way this feature varies with the constituent being classified and it also partially addresses the issue that the semantic role of one constituent is not independent of other arguments for this predicate.

As we pointed out in Section 2, the argument labels in the PropBank annotation are verb-specific. Given a head word or phrase type, the system will be more certain of the semantic role label when it also knows what the predicate is. The same head word or phrase type may be associated with different semantic role labels for different predicates. The head word + predicate and the phase type + predicate features are designed to capture this linguistic intuition. The other type of combination features are verb class + head word and verb class + phrase type. We will discuss the use of verb classes as features in detail in Section 4.2.1.

The first-word-in-the-constituent and the phrase label of the left sibling features are from Pradhan, Ward et al. (2004) and the interested reader is referred to their work for an explanation of why these are useful features. Because Chinese is a language with mixed headedness, namely, some phrases are left-headed and some phrases are right-headed, the first word and last word are a more robust but sloppier way of finding the head when the head-finding heuristics fail.

Some of the linguistic phenomena that impact the syntactic realization of argument structures in Chinese are the BA and BEI constructions. In the Chinese Treebank, BA and BEI represent closed sets of light verbs that take clausal complements. The subject of the clausal complement of BA tends to be *Arg1* instead of *Arg0* in a canonical clause structure. The BEI construction is the Chinese passive construction in which the subject

of the clause headed by BEI is typically *Arg1*. In order to capture this information, we added as features the path from the BEI and BA words to the constituent in focus. BA and BEI are not predicates themselves, so these features are only invoked for the predicate in the complement clause of the BA and BEI.

4.2.1 Using Verb Classes to Improve Semantic Role Labeling. With the current experimental setup, as is also the case in most of the work on semantic role labeling, training data and test data are not divided by verb instances but by the number of articles. As a result, it is expected that the verb instances are not evenly divided. It is entirely possible that some verbs can only be found in the training data and others can only be found in the test data. By our count, there are 4,526 verb types in the training data and 1,038 verb types in the test data. One hundred seventy-six verb types that occur in the test data are absent from the training data. Because the semantic role labels are defined with regard to the individual verbs, this can be a real problem because the model learned in the training process does not optimally fit with the test data if different verbs are involved. Fortunately, many verbs have similar argument structures and therefore are annotated with similar semantic role labels in the Chinese PropBank. For example, verbs like 加大 ("enlarge"), 加剧 ("make more drastic"), 加快 ("accelerate"), 加强 ("strengthen"), 加深 ("deepen"), 加速 ("accelerate"), 加重 ("give more weight"), 加高 ("make higher") all take two arguments, a theme that undergoes a change of state and an external force or agent that brings about the change of state. These verbs are uniformly annotated and they all have two numbered arguments with *Arg0* denoting the cause and *Arg1* denoting the theme. It would make sense to group these verbs together into a class and use this information in the features as has been done for English using VerbNet (Yi, Loper, and Palmer 2007). Having a membership in a particular class says something about the predicate–argument structure of a verb. When a verb is absent in the training data, which is a familiar sparse data problem, the class information may tell the system how to label the semantic roles of this verb based on its semantic class.

Although to our knowledge no such classification exists for Chinese verbs based on the predicate–argument structure, a rough classification can be automatically derived from the frame files, which are created to guide the PropBank annotation. We classified the verbs along three dimensions: the number of arguments, the number of framesets, and selected syntactic alternations.

Number of arguments Verbs in the Chinese PropBank can have one to five arguments, with the majority of them having one, two, or three arguments. Verbs with zero arguments are auxiliary verbs[4] like 必 ("will"), 得以 ("be able to"), 该 ("should"), 敢 ("dare"), 可 ("may"), 肯 ("be willing to"), 能 ("can"), 能够 ("can"), 须 ("must"), 应当 ("should"), and some other light verbs. Verbs that have five arguments are change of state verbs like 延长 ("lengthen"), 缩短 ("shorten"), 降低 ("lower"), 提高 ("increase"), 扩大 ("enlarge"), 缩小 ("make smaller"). These verbs generally take as arguments a theme that undergoes the change of state, the original state, the new state, the range of the change, and the cause or agent that brings about the change.

Number of framesets A frameset roughly corresponds to a major sense. This information is used because it is common that the different framesets of a verb can have different numbers of arguments. For example, verbs like 平衡 ("balance") can be used either as a non-stative verb, in which case it means "balance," or a stative verb, in

4 One could say that the argument of the auxiliary verb is the entire proposition, but in this phase of the Chinese PropBank, auxiliary verbs are not annotated.

which case it means "balanced." When it is used as a non-stative verb, it takes two arguments, the thing or situation that is balanced and the balancer, the entity that maintains the balance. When it is used as a stative verb, obviously it only takes a single argument.

Syntactic alternations We also represent certain types of syntactic alternations. One salient type of syntactic alternation is the well-known "subject of intransitive / object of transitive" alternation described in detail in Levin (1993). Chinese verbs that demonstrate this alternation pattern include 出版 ("publish"). For example, 这 ("this") 本 (CL) 书 ("book") plays the same semantic role even though it is the subject in "这/this 本/CL 书/book 出版/publish 了/AS" and the object in "这/this 家/CL 出版/publishing 社/house 出版/publish 了/ASP 这/this 本/CL 书/book."

Thus each verb will belong to a class with a symbol representing each of the three dimensions. For example, a given verb may belong to the class "C1C2a," which means that this verb has two framesets, with the first frameset having one argument and the second having two arguments. The "a" in the second frameset represents a type of syntactic alternation. Forty classes were semi-automatically derived in this manner.

Such a classification scheme will undoubtedly prove to be linguistically unsophisticated. Verbs that have the same number of arguments may have different types of arguments, and the current classification system does not pick up these distinctions. However, our experiments show that even such a simple classification can be used to provide features that improve the semantic role labeling performance.

4.3 Using Automatic Parses

Previous work (Sun and Jurafsky 2004) on Chinese semantic role labeling uses a parser that assumes correct (hand-crafted) segmentation. As word segmentation is a very challenging problem that has attracted a large body of research by itself, it is still unclear how well semantic role tagging in Chinese can be performed in realistic situations. In our experiments, we implemented a Maximum Entropy–based parser similar to Luo (2003). The parser performs Chinese word segmentation, POS tagging, and parsing in one integrated system. The parser is trained on the Xinhua news and Broadcast news portion of the Chinese Treebank, which has 498K words. Tested on the held-out test data, the parser achieved an unlabeled precision and recall of 0.889 and 0.868, respectively, for the combined word segmentation and parsing accuracy. When the word segmentation is singled out for evaluation, the parser achieved an F-score of 0.969. It is important to point out that these results cannot be directly compared with most of the results reported in the literature, where correct segmentation is assumed. In addition, in order to account for the differences in segmentation, each character has to be treated as a leaf of the parse tree. This is in contrast with word-based parsers where words are terminals. For comparison purposes, we also used the Bikel parser (Bikel 2004). Because the Bikel parser assumes segmented sentences as input, we extracted the segmentation from the output of our parser and fed it into the Bikel parser. We also experimented with using gold-standard segmentation and POS from the Chinese Treebank as input to the Bikel parser to measure the effect of segmentation and POS tagging on the performance on the semantic role labeling. Because semantic role labeling is performed on the output of a syntactic parser, only constituents in the parse tree are candidates. If there is no constituent in the parse tree that shares the same text span with an argument in the manual annotation, the system cannot possibly get

Table 2
Semantic role labeling results for verbal predicates.

parse	constituents	feature set	precision	recall	F1 measure
treebank	known	baseline	n/a	n/a	.931 (acc)
treebank	known	all	n/a	n/a	.941 (acc)
treebank	unknown	baseline	.920	.900	.910
treebank	unknown	all	.930	.910	.920
maxent parser	unknown	baseline	.689	.597	.639
maxent parser	unknown	all	.694	.602	.645
Bikel parser (auto seg)	unknown	baseline	.745	.596	.662
Bikel parser (auto seg)	unknown	all	.748	.603	.668
Bikel parser (gold seg)	unknown	all	.768	.625	.689
Bikel parser (gold pos)	unknown	all	.795	.656	.719

a correct annotation. In other words, the best the system can do is to correctly label all arguments that have a constituent with the same text span as in the parse tree.

4.4 Results and Discussion

4.4.1 Data. In all our experiments we use the Chinese Proposition Bank Version 1.0.[5] This version of the Chinese PropBank (Xue and Palmer 2003) consists of standoff annotation on the first 760 articles (`chtb_001.fid` to `chtb_931.fid`) of the Chinese Treebank. This chunk of the data has 250K words and 10,364 sentences. The total number of verb types in this chunk of the data is 4,854.[6] Following the convention of the English semantic role labeling experiments, we divide the training and test data by the number of articles, not by the verb instances. For all our experiments on semantic role labeling of verbs, 72 files (`chtb_001.fid` to `chtb_040.fid` and `chtb_900.fid` to `chtb_931.fid`) are held out as test data,[7] 40 files (`chtb_041.fid` to `chtb_080.fid`) are used as development set, and the remaining 648 files (`chtb_081.fid`to `chtb_899.fid`) are used as training data. The training, development, and test sets have 30,280; 1,971; and 3,454 propositions, respectively. Our parser is trained on the training and development set plus 275K words of broadcast news that have been recently annotated as part of the Chinese Treebank Project.[8] That is, in addition to the training data for the semantic role labeling experiments, it also uses a portion of the treebank which has not yet been propbanked.

4.4.2 Results. The results of the semantic role labeling for both hand-crafted and automatic parses are presented in Table 2. These results represent an improvement over what has been reported in Xue and Palmer (2005) due to the improved parsing results and new features. To be used in real-world natural language applications, a semantic

5 This data is publicly available through the Linguistic Data Consortium.
6 These include the so-called stative verbs, which roughly correspond to adjectives in English.
7 This chunk of data is chosen as test data because it is double annotated and adjudicated.
8 We did not use the Sinorama portion of the Chinese Treebank because it is a very different genre and adding it to the training data hurts parser performance (Bikel 2004).

role tagger has to use automatically produced constituent boundaries either from a parser or by some other means, but experiments with hand-crafted parses will help us evaluate how much of a challenge it is to map a syntactic representation to a semantic representation, which may very well vary from language to language. When hand-crafted parses in the Chinese Treebank are used as input, our system achieved an F-score of 0.92 for combined argument detection and classification. This accuracy is achieved when the new features are added. Without the new features, the accuracy drops about one percentage point. When the arguments are known, the accuracy is at 94.1% when the new features are used, up one percentage point from the baseline. This accuracy is fairly high considering the fact that the state-of-the-art for semantic role labeling systems trained on the English PropBank (Palmer, Gildea, and Kingsbury 2005) is about 93% percent (Pradhan, Ward et al. 2004; Xue and Palmer 2004) when the arguments are known and the English PropBank is a much larger corpus that has one million words. The high baseline accuracy also suggests that the features used for the English semantic role labeling port very well to Chinese. In addition, there are several facilitating factors for Chinese semantic role labeling when hand-crafted parses are provided as input. First of all, Chinese verbs appear to be less polysemous, at least the ones that occur in the Chinese Treebank. Of the 4,854 verbs in this version of the Chinese Proposition bank, only 62 verbs have three or more framesets. In contrast, 294 verbs out of the 3,300 verbs in the Penn English PropBank have three or more framesets. When a verb is less polysemous, the arguments of the verb tend to be realized in a more uniform manner in syntax. As a result, the argument labels are easier to predict from their structure. Chinese seems to compensate for this fact by using a larger number of verbs. This becomes obvious when we consider the fact that the 4,854 verbs are from just 250K words and the 3,300 verbs in the English PropBank are from one million words. A related fact is that adjectives in Chinese are traditionally counted as verbs and they generally have only one argument with a much simpler syntactic realization. For example, 便宜 (”inexpensive”) and 薄 (”thin”) are considered stative verbs in the Chinese Treebank.

We also believe that a more subtle explanation for the higher semantic role labeling accuracy given the annotation of the Chinese Treebank is the fact that the Chinese Treebank has richer structure (see Xue et al. [2005] for a comparison of the Penn English Treebank and the Chinese Treebank). By using less flat and more hierarchical structures, the Chinese Treebank resolves some of the attachment ambiguities that impair semantic role labeling. For example, the complement and adjunct in a VP in the Chinese Treebank are attached in different syntactic configurations with regard to the verb. Because complements are generally numbered arguments and adjuncts are generally *ARGM*s, the semantic role labeler can take advantage of this information when it tries to determine when a constituent is a numbered argument or an adjunct.

This apparent advantage in Chinese semantic role labeling is diminished when an automatic parser is used. First of all, the hierarchical structures in the hand-crafted parses that aid semantic role labeling are hard to recover with an automatic parser. Resolving the many attachment ambiguities caused by the hierarchical structures in language is one of the most difficult problems in the parsing literature. Parsing Chinese in a realistic scenario is especially difficult given that it has to build structures from characters rather than words, and Chinese also has few morphological clues to help in making parsing decisions. Our results show that the semantic role labeling accuracy improves by 2.1% in F-score when the correct segmentation is used as input to the Bikel parser. When the correct POS tags are used, the semantic role labeling accuracy improves another 3%. At present, improvement in Chinese parsing is also hindered by the

smaller training set. Although the Chinese Treebank 5.1 has a decent size of 500K words, it consists of data from very different sources. Due to their very different styles, training on one portion of the data does not help or may even hurt the parsing accuracy on the other portion (see Bikel [2004] for a discussion of this issue). The situation improves somewhat with the addition of the 275K words from broadcast news,[9] which leads to an improvement in parsing accuracy. We believe further improvement in semantic role labeling accuracy will be to a large extent contingent on the parsing accuracy, which requires more training materials that are similar in style.

5. Semantic Role Labeling of Nominalized Predicates

In this section, we describe our experiments on nominalized predicates in Chinese, using the Chinese NomBank as training and test data. In Section 5.1 we show that the pruning algorithm for nominalized predicates needs to account for two disjoint cases. When a support verb is present, the pruning algorithm needs to go outside the NP headed by the predicate to search for potential arguments. When there is no support verb, the arguments can generally be found inside the NP headed by the predicate. In Section 5.2, we describe the features used in the semantic role labeling of nominalized predicates. There are three groups of features: features used in the semantic role labeling of verbs minus a few features that do not carry over to nouns, features used for verbs that need to be substantially adapted, and new features we designed specifically for nominalized predicates. The experiments we conducted on nominalized predicates largely parallel those of verbs, for an easier comparison. Again there are two experimental variables, the level of human annotation in the input to the semantic role labeling system and whether the constituents for the arguments are known. The experimental results are presented in Section 5.3.

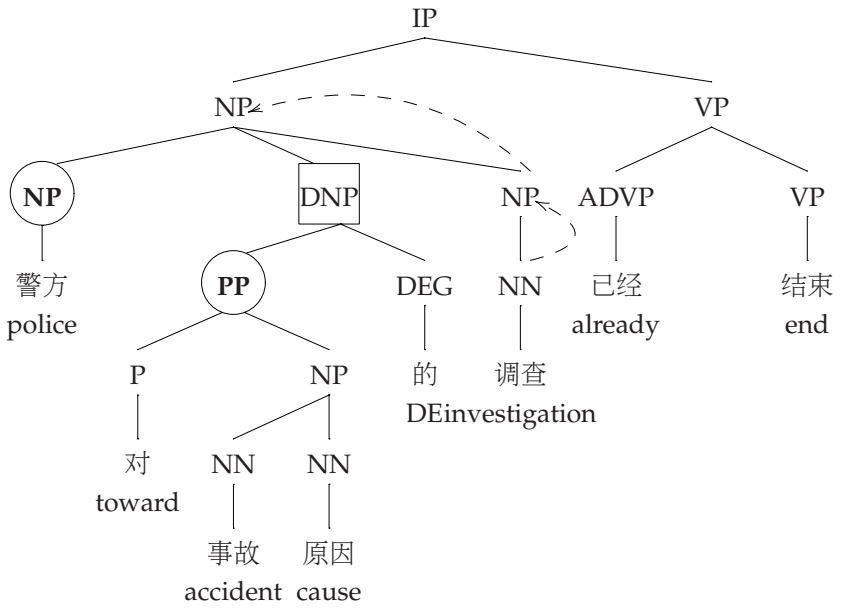
5.1 Pruning for Nominalizations

Like verbal predicates, the arguments and adjuncts of a nominalized predicate are related to the predicate itself in linguistically well-understood structural configurations. As we pointed out in Section 2, most of the arguments for nominalized predicates are inside the NP headed by the predicate unless the NP is the object of a support verb, in which case its arguments can occur outside the NP. Typically the subject of the support verb is also an argument of the nominalized predicate, as illustrated in Example (3). The majority of the constituents are not related to the predicate in question, especially because the sentences in the treebank tend to be very long. There are two distinct cases that need to be handled differently, depending on the presence or absence of a support verb for the nominalized predicate. When the nominalized predicate does not occur with a support verb, generally all of its arguments are realized within the NP of which it is the head. The pruning algorithm starts from the predicate, collects its sisters, and adds them to the candidate list. It then iteratively goes up one level and collects the sisters of that constituent until it reaches the top-level NP of which it is the head. An exception is made when the constituent is DNP, in which case the candidate added is the first daughter of the DNP, not the DNP itself. This is illustrated in Example (9),

9 This new data set will soon be available via the LDC in another Chinese Treebank release.

where the algorithm starts from the nominalized predicate 调查 (”investigation”), and, because it does not have any sisters, it does not add anything to the candidate list at this level. It then goes up to its parent NP, and collects its sisters NP (警方 ”police”) and DNP (对 ”toward” 事故 ”accident” 原因 ”cause” 的 ”DE”). In the case of DNP, the candidate added is actually its first daughter, the PP.

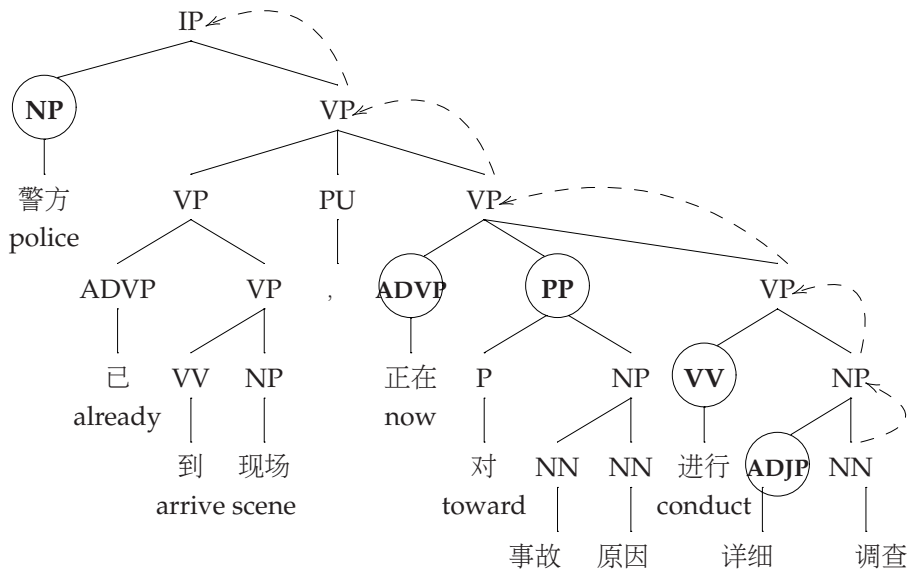
(9)



”The police investigation of the cause of the accident has ended.”

When a nominalized predicate occurs with a support verb, the NP headed by the nominalized predicate is generally the object of the support verb. Arguments can often be found both inside and outside this object NP. The pruning algorithm starts from the nominalized predicate and collects its sisters. It then iteratively goes one level up until it reaches the top-level IP node. At each level, the sisters of the current node are added to the list of candidates. Note that the algorithm does not stop at the top NP level, so that arguments outside the NP can also be captured. In practice, it is generally not known to the algorithm whether the governing verb, the verb that takes the NP headed by the nominalized predicate as object, is a support verb or not. Support verbs are often light verbs and they are only a subset of all governing verbs. The system simply assumes that all verbs taking the NP headed by a nominalized predicate as its object are support verbs, adds constituents outside the NP as candidates, and lets the machine-learning algorithm figure out whether they are arguments or not. This pruning process is illustrated in Example (10), where the algorithm starts from the nominalized predicate 调查 (”investigation”). It first collects its sister ADJP (详细 ”thorough”), and then it will go one level up to the NP, and adds the support verb (进行 ”conduct”) to the candidate list. It will go another level up to the VP and adds its sisters ADVP (正在 ”now”) and PP (对 ”toward” 事故 ”accident” 原因 ”cause”) to the candidate list. It then goes one more level up and decides this is a coordination structure; no candidate is added at this level. At the next VP level it adds 警方 (”police”) to the list of candidates. The algorithm terminates at the IP node.

(10)



"The police has arrived at the scene and is thoroughly investigating the cause of the accident."

Overall, pruning works less effectively for nouns than for verbs. When treebank parses are used, our pruning algorithm can recall over 94% of the arguments while pruning out 93% (87,724 out of 93,916) of the nodes. When automatic parses (maxent segmentation + Bikel parser) are used, our pruning algorithm can recall 73% of the arguments out of the 88% of arguments that have a constituent in the parse tree, while pruning out 93% (85,160 out of 91,356) of the nodes. However, although there is a small drop in recall (from 0.569 to 0.529) compared with when the pruning algorithm is not used, there is a huge gain in precision (from 0.146 to 0.623), a similar trend to that which we have observed for the semantic role labeling of verbs.

5.2 Features

The features we use for the semantic role labeling of nominalized predicates fall into three groups. The baseline features we used are the same features we used for the semantic role labeling of verbs. The second group of features are adapted from features used in the semantic role labeling of verbs. In particular, the path feature is redefined in the semantic role labeling of nominalized predicates. A significant number of NPs in the Chinese Treebank are flat and they consist of a sequence of nouns. When there are nouns on both sides of a predicate, which is a noun itself, the path from the predicate to the preceding or following noun has the same value. However, the preceding and following nouns do not have the same probability of being an argument. We therefore need to clearly mark the position of the predicate (e.g, P=NN↑NP↓NN is not the same as NN↑NP↓NN=P). Such a problem does not exist for the semantic role labeling of verbs because their arguments are rarely a verb as well. The third group of features are new features we added specifically for the semantic role labeling of nominalized predicates. Like the features for the semantic role labeling of verbal predicates, the features for argument detection only are marked as "D" and the features for argument

classification only are marked as "C." The features for both argument detection and argument classification are marked as "C,D." The complete list of features is listed here.

I. **Baseline Features:**

C *Position*: The position is defined in relation to the predicate and the values are *before* and *after*. Because most of the arguments for nominalized predicates in Chinese are before the predicates, this feature is not as discriminative as when it is used for verbal predicates where arguments can be both before or after the predicate. (G&J)

C *Phrase type*: The syntactic category of the constituent being classified. (G&J)

C *First and last word of the constituent being classified*. (P et al.)

C,D *Predicate*: The nominalized predicate itself. (G&J)

C,D *Predicate combination features*: Predicate + head word combination, predicate + phrase type combination. (X&P)

C,D *Predicate class*: The verb class the predicate belongs to; same predicate class as those used for verbs.

C,D *Predicate class combination features*. Predicate class + head combination, predicate class + phrase type combination.

C,D *Head word and its part of speech*: The head word and its part of speech. (G&J)

C,D *Path*: The path between the constituent being classified and the predicate. (G&J)

II. **Adapted features:**

C,D *Path*: The path between the constituent being classified and the predicate, with the predicate clearly identified.

III. **New Features:**

D *Topic NP*: A binary feature indicating whether the constituent is a topic if the predicate is the subject.

D *Inside NP headed by the predicate*: A binary feature indicating whether the constituent in focus is inside the NP headed by the predicate.

D *Position of the constituent in relation to the support verb*: The value can be before or after the support verb, or is the support verb itself.

C,D *Sisterhood with predicate*: A binary feature that indicates whether the constituent is a sister to the predicate.

C,D *Path + governing verb*. The path feature combined with the governing verb.

Several features that we used for the semantic role labeling of verbal predicates were dropped from our experiments with nominalized predicates. Specifically, the subcat feature and subcat+ features were not used because it is not clear how these features can be defined for a nominalized predicate. A couple of new features were added to the feature set for semantic role labeling of nominalized predicates. As we have demonstrated in Section 5.1, a support verb to a large extent determines whether or

not the arguments of a nominalized predicate can occur outside the NP of which it is the head. Therefore it is effective information for discriminating arguments from non-arguments. It is also indicative of the specific semantic role of an argument in the argument classification task. To capture this observation, we used a combined feature of path + governing verb that was only invoked when there was an intervening governing verb between the constituent being classified and the predicate. The governing verb is used as an approximation of the support verb for this feature because the system does not have prior knowledge of whether a verb is a support verb or not absent some external resource that provides a list of possible support verbs. The governing verb, on the other hand, can be approximated by looking at the syntactic configuration between the nominalized predicate and the verb. This feature is used for both argument detection and argument classification. Another feature we specifically used for the semantic role labeling of nominalized predicates is the sisterhood feature. When looking at the data, we found a substantial number of NPs headed by a nominalized predicate have a flat structure with their sisters as their arguments. The sisterhood feature is designed to capture this observation and it is also used for both argument detection and argument classification. The other three new features were used for argument detection only. When a nominalized predicate is in the subject position, the NP in the topic position tends to be its argument. A binary feature is invoked when the constituent in focus is an NP that is the left sister of the subject NP headed by the predicate. Whether an NP is a subject is also determined heuristically: An NP is considered to be subject if its parent is an IP and its right sister is a VP. Another binary feature used for argument detection is whether the constituent in focus is inside the NP headed by the predicate. Finally, the position of the constituent in relation to the support verb is also used as a feature for argument detection. The value for this feature can be before or after the support verb, or it can be the support verb itself.

5.3 Experiments

5.3.1 Data. Our system is trained and tested on a pre-release version of the Chinese NomBank. This version of the Chinese NomBank consists of standoff annotation on the first 760 articles (`chtb_001.fid` to `chtb_931.fid`) of the Chinese Treebank. This is the same chunk of treebank data as used in our experiments on verbs. It has 1,227 nominalized predicate types and 10,497 nominalized predicate instances, in comparison with the 4,854 verb predicate types and 37,183 verb predicate instances in the same chunk of data. By instance, the NomBank is between a quarter and one third of the size of the Chinese PropBank. Similarly to our experiments on verbs, we divide the training, development, and test data by the number of articles, not by the predicate instances. For all our experiments, we used the same data split as that of the verbs: 648 files (`chtb_081.fid` to `chtb_899.fid`) are used as training data, 40 files (`chtb_041.fid` to `chtb_080.fid`) are used as development data, and the other 72 files (`chtb_001.fid` to `chtb_040.fid` and `chtb_900.fid` to `chtb_931.fid`) are held out as test data. The same parsers are used for the semantic role labeling experiments for verbs and nouns.

5.3.2 Results and Discussion. Parallel to our experiments on verbs, we also present experiments using hand-crafted and automatic parses. The experimental results are presented in Table 3, which represents an improvement from what has been reported in Xue (2006b). The baseline results are obtained using the subset of features used in the semantic role labeling of verbs, minus the subcat and subcat+ features. We also report improved results by using additional new features and adapting the path feature. The

Table 3
Semantic role labeling results for nominalized predicates.

parse	constituents	feature set	precision	recall	F1 measure
treebank	known	baseline	n/a	n/a	.843 (acc)
treebank	known	all	n/a	n/a	.849 (acc)
treebank	unknown	baseline	.722	.608	.660
treebank	unknown	all	.734	.661	.696
maxent parser	unknown	baseline	.60	.471	.526
maxent parser	unknown	all	.60	.502	.547
Bikel parser (auto seg)	unknown	baseline	.611	.492	.545
Bikel parser (auto seg)	unknown	all	.623	.529	.573
Bikel parser (gold seg)	unknown	all	.629	.531	.576
Bikel parser (gold pos)	unknown	all	.657	.560	.604

use of adapted and new features leads to significant improvement in all experiment settings except when the constituents are already known and treebank parses are used. This is not surprising given that more new features were added to the argument detection task than the argument classification task.

Compared with the 94.1% for verbal predicates on the same data, the 84.3% the system achieved for nominalized predicates on treebank parses when the constituents are given is considerably lower, suggesting that the semantic role labeling for nominalized predicates is a much more challenging task. The difference between the semantic role labeling accuracy for verbal and nominalized predicates is even greater when the constituents are not given and the system has to identify the arguments to be classified. Our system achieves an F-score of 0.696 when treebank parses are used, and this is in contrast with the F-score of 0.92 for verbal predicates under similar experimental conditions.

For our experiments using automatic parses, we used the same parsers for nominalized and verbal predicates. The first parser is the character-based Maximum Entropy parser that we developed in-house; and it does word segmentation, POS-tagging, and syntactic parsing in one integrated system. The second parser is the Bikel parser that takes three different kinds of input. In its fully automatic mode, it uses the segmentation extracted from the output of our Maxent parser. We also experimented with using correct segmentation and correct segmentation plus correct POS-tagging as input to the Bikel parser to measure the degradation in performance with decreasing levels of human annotation. Our results show that the Bikel parser outperforms our Maxent parser 0.028 (F-score) in semantic role labeling accuracy when using fully automatic parses. When the Bikel parser is used, the system achieves an F-score of 0.573, in comparison with the 0.547 achieved by the Maxent parser. There is a gradual degradation in performance with less human annotation, consistent with our experiments on verbs. It is somewhat surprising that the segmentation does not affect the semantic role labeling for nominalized predicates as it does for verbs. Using correct POS tags as input to the Bikel parser, however, leads to a significant improvement of 0.028 in F-score over using correct segmentation only, from 0.576 to 0.604. Overall, there is a smaller gap between when treebank parses are used and when automatic parses are used. There are two possible explanations. One is that the NP structures are more local and less prone to parsing errors, so there is less of a difference between treebank and automatic parses. This is

consistent with the fact that 88% of the arguments for nominalized predicates were recovered by the parser, in contrast with the 87% of the arguments for verbal predicates. Another possible explanation is that argument detection is challenging even with gold-standard treebank parses, which makes the gap between treebank and automatic parses smaller.

5.3.3 Error Analysis. The much lower accuracy in the semantic role labeling of nominalized predicates warrants a closer examination. One thing we looked at is the fact that arguments of nominalized predicates can occur either inside or outside the NP headed by the predicate. Of the 1,124 predicate instances in the test data, 331 of them have arguments that occur outside the NP headed by the predicate. The remaining 793 instances have all their arguments inside the NP. We found a significant difference in the semantic role labeling accuracy for the two types of predicates in the experiment setting where the input to the semantic role labeling system is treebank parses and the constituents are unknown. For the predicates that have arguments outside the NP, the system achieved a precision and recall of 0.868 and 0.633, respectively. For the predicates that have all their arguments inside the NP, the precision and recall are 0.661 and 0.695, respectively. We believe the large difference in precision is the result of the system erroneously identifying arguments outside the NP when the predicate heads an NP that is the object of a verb, even if the verb is not a support verb. With a small data set, there is insufficient training data for the system to tell whether or not a verb is a support verb that licenses arguments outside the NP headed by the predicate. We also examined cases where the predicate heads an NP that is the head of a relative clause. Because the NP headed by the predicate is semantically associated with a trace inside the relative clause, its arguments can generally be found inside the relative clause. Out of the 1,124 predicate instances, 138 are the heads of relative clauses. The precision and recall for these predicates are 0.668 and 0.5, respectively, in comparison with the 0.749 and 0.696 for predicates that are not the head of a relative clause. The much lower recall suggests the arguments for the head of a relative clause are much harder to identify.

6. Related Work

Computational approaches to semantic interpretation have a long tradition, but the line of research that this work follows is relatively young. Gildea and Jurafsky (2002) provided the seminal work on the semantic role labeling, using the FrameNet corpus as training and test material. Since then, there has been rapid improvement in the semantic role labeling accuracy of English verbs, fueled by the development of PropBank (Palmer, Gildea, and Kingsbury 2005), which annotates the verbs in the one-million-word Penn Treebank with semantic role labels. A wide range of statistical and machine learning techniques have been applied to the semantic role labeling of verbs, using PropBank as training and test material. The machine-learning techniques used include Support Vector Machines (Pradhan, Ward et al. 2004; Tsai et al. 2005), Maximum Entropy (Xue and Palmer 2004; Haghighi, Toutanova, and Manning 2005; Liu et al. 2005; Yi and Palmer 2005), Conditional Random Fields (Cohn and Blunsom 2005), and many others. Because semantic role labeling is a complex task based on a wide range of lower level natural language techniques, many different preprocessing, integration, and combination techniques have been explored. The relative merits of using a full syntactic parser that provides hierarchical structures (Xue and Palmer 2004) vs. a shallow chunker (Pradhan, Hacioglu et al. 2005; Hacioglu et al. 2004) has been studied extensively. Noting that parsing errors are difficult or even impossible to recover at the semantic

role labeling stage, Yi and Palmer (2005) experimented with integrating semantic role labeling with a Maximum Entropy-based parser, effectively treating semantic role labels as function tags on the constituents in a parse tree. Koomen et al. (2005), Pradhan, Ward et al. (2005), Màrquez et al. (2005), and Tsai et al. (2005) pursued alternative approaches to make their semantic role labeling systems more robust by combining the output of multiple systems. Punyakanok, Roth, and Yi (2005), in particular, achieved the best performance (F1 = 0.794) on the WSJ test set in the 2005 CoNLL shared task by combining multiple semantic role labeling systems using an integer linear programming technique (Punyakanok et al. 2004). Pradhan, Hacioglu et al. (2005) reported the best result (F1 = 0.684) on the Brown test set using the WSJ data as the training set by combining the output of different semantic role labeling classifiers using a chunking procedure. They also reported the state-of-the-art result (F1 = 0.81) on the standard PropBank test set, using the same techniques. Most of the early systems consider each argument on its own when assigning the semantic role labels, allowing the theoretical possibility that more than one core argument may share the same semantic role label, violating the linguistic constraint that the same semantic role label cannot be assigned to more than one core argument. Toutanova, Haghighi, and Manning (2005) address this by using a joint-learning strategy to rule out such conflicting argument labels.

The semantic role labeling performance on the FrameNet data set has also improved significantly from Gildea and Jurafsky's (2002) early results, thanks mostly to the Senseval-3 semantic role labeling competition (Litkowski 2004). Participants of Senseval-3 have used a variety of machine learning algorithms to tackle the semantic role labeling problem: Maximum Entropy (Baldewein et al. 2004; Ngai et al. 2004; Kwon, Fleischman, and Hovy 2004); Boosting, SNOW, and Decision Lists (Ngai et al. 2004); SVM (Bejan et al. 2004; Moldovan et al. 2004; Ngai et al. 2004); Memory-based learning (Baldewein et al. 2004), as well as Generative models (Thompson, Patwardhan, and Arnold 2004). Bejan et al. (2004) achieved the best result using an SVM classifier combined with improved linguistic features. They achieved an F1 measure of 0.763 in their internal evaluation, and 0.831 using the more lenient official Senseval-3 scorer.

Compared with the large body of work on the semantic role labeling on verbs, the argument structure analysis of nominal predicates has so far received less attention. Jiang and Ng (2006) reports a semantic role labeling system on nominal predicates, also using the maximum entropy approach. Their system achieves F1 scores of 0.727 and 0.691, respectively, on gold-standard and automatic parses, indicating semantic role labeling of nominal predicates is a much more difficult problem than that of verbs for English as well. Outside the narrow domain of semantic role labeling, there has been a steady accumulation of work on semantic analysis of nouns and a gradual expansion of the domain in which the semantic analysis is performed. Lapata (2002) developed a probabilistic model for the interpretation of nominalizations, focusing on the semantic relation between the noun head and its prenominal modifier in a nominalized compound (i.e., whether the prenominal modifier is an underlying subject or direct object of the verb from which the nominalized head is derived). Their model achieved a very high accuracy of 0.861 when evaluated on data extracted from the British National Corpus. Girju et al. (2004) and Moldovan and Badulescu (2005) extended the domain of linguistic analysis to that of noun phrases. In particular, they focused on the study of four nominal constructions: **complex nominals** in which a head noun is modified by other nouns or adjectives derived from nouns, **genitives**, **adjective phrases**, and **adjective clauses**. In general, previous work on nominals, perhaps with the exception of Nakov and Hearst (2006) all attempt to specify a finite set of semantic relations between the nouns and their modifiers in the spirit of Levi (1979). The PropBank/NomBank approach to

semantic role labeling adopted here represents a departure from this tradition in that the semantic relations in PropBank and NomBank are predicate-specific. There is no serious attempt to induce cross-predicate semantic relations. In addition, the semantic relations represented by the PropBank/NomBank semantic roles are not pairwise relations between the predicate and one of its arguments as they are in previous work. A third difference is that the arguments of a nominalized predicate can be found outside the noun phrase headed by the predicate (Meyers, Reeves, and Macleod 2004; Meyers et al. 2004), making argument identification a much more challenging task. Whereas the English NomBank annotates both relational nouns and nominalizations, the Chinese NomBank only deals with nominalization, making it a more coherent task.

Work on Chinese semantic role labeling is still in its infancy. Lacking a Chinese corpus annotated with semantic roles, the few prior works generally relied on annotating a small corpus for their experiments. Sun and Jurasfky (2004) did preliminary work on the semantic role labeling of Chinese verbs by annotating 10 selected verbs that have a frequency ranging from 41 to 230, using the Chinese PropBank annotation guidelines. Pradhan, Sun et al. (2004) extended that work to Chinese nominalizations, and reported preliminary work for analyzing the predicate–argument structure of 630 propositions for 22 nominalizations taken from the Chinese Treebank. Noting the difficulty of Chinese parsing, Kwong and T'sou (2005) approached the semantic role labeling task as one of identifying and labeling the head word of the arguments. They annotated the semantic roles for 41 verbs in 980 sentences in a primary school textbook corpus and the same verbs in 2,122 sentences in a news corpus. Perhaps not surprisingly, they reported F1 scores of 0.529 and 0.444, respectively, for the textbook and news corpora when training and test data are from the same corpus, and 0.463 and 0.398, respectively, when the training and test data are from different corpora. As far as we know, the work reported here is the first to use sizable Chinese semantically annotated corpora. The approach adopted in the present work emphasizes the integration of linguistically informed heuristics and machine-learning approaches, and the exploration of the underlining linguistic insights behind the features used in machine-learning systems. We believe semantic role labeling provides an ideal stage where linguistic observations can be formalized as features and fed into a general machine-learning framework for testing and verification and natural language technologies can be advanced in the process.

7. Conclusions and Future Work

We have presented the first experimental results on Chinese semantic role labeling using the Chinese PropBank and the Chinese NomBank. We have shown that given gold-standard parses, Chinese semantic role labeling can be performed with considerable accuracy on Chinese verbs. In fact, even though the Chinese PropBank is a significantly smaller corpus than the English PropBank, we achieved results that are comparable with the state-of-the-art English semantic role labeling systems. We suggest three factors that are particularly conducive to the semantic role labeling of Chinese verbs when the hand-crafted treebank parses are used as input. One is that Chinese verbs tend to be less polysemous compared with English, which contributes to a more uniform mapping between the predicate–argument structure and its syntactic realization. Another facilitating factor is that stative verbs, which generally translate into adjectives in English, account for a large proportion of all the verbs in the Chinese PropBank and they tend to have simple argument structures. Finally, we suggest that the richer structure in the Chinese Treebank makes certain aspects of the semantic role labeling simpler.

One such example is that the clear structural distinction between syntactic arguments and adjuncts makes it easier for the semantic role labeling system to differentiate core arguments and adjuncts for Chinese verbs. These all translate into lower confusability along the lines of Erk and Padó (2005) in the mapping from the syntactic structure to the semantic role labels.

When the semantic role labeling takes raw text as input, it cannot take advantage of the rich syntactic structure of the treebank unless it can be reproduced with high accuracy by an automatic parser. Even though our experiments using fully automatic parses yield promising initial results, the accuracy is significantly lower than the English state of the art. Our parsing accuracy is hampered by a significantly smaller training set that is only half the size of the Penn Treebank. We also suggest that there are a few inherent linguistic properties of the Chinese language that make syntactic parsing a particularly challenging task. The first has to do with the fact that Chinese text does not come with word boundaries and our parser has to build structures from characters rather than words. The second has to do with the fact that Chinese has very little inflectional morphology that the parser can exploit when deciding the part-of-speech tags of the words. Both word segmentation and POS-tagging difficulties will lead to parsing errors when larger phrase structures are built.

Our experimental results also show a substantial gap between system performance on verbs and nominalized predicates. This difference can be partially attributed to the smaller corpus size of the Chinese Nombank, with fewer instances of nominalized predicates than verbs in the underlying Chinese Treebank, but we believe the main reason is that the semantic role labeling is more challenging for nominalized predicates than for verbs. This again can be explained in terms of confusability in the mapping from syntactic structure to the predicate–argument structure. In general, the NPs in the Chinese Treebank have flatter structures compared with verbs. For example, there is no clear structural distinction between arguments and adjuncts for nominalized predicates that are analogous to the argument/adjunct distinction for verbs. Another reason for the lower accuracy for nominalized predicates is the more diverse distribution of their arguments. Arguments can be found either inside or outside the NP headed by the predicate, or even in relative clauses that modify the NP headed by the predicate.

There are many directions we can go from here for future work. There are many proven techniques that can be implemented for Chinese, the most important of which is to make Chinese parsers more robust. One thing we plan to experiment with is the combination of multiple parsers and multiple semantic role labeling systems. We also believe that we have not settled on an "optimal" set of features for Chinese semantic role labeling and more language-specific customization is necessary. We believe that joint-learning is also a promising avenue to pursue, especially for verbs where generally more core arguments are realized.

Acknowledgments

We would like to thank Martha Palmer for her comments on this manuscript and early versions of the paper, and more importantly for her steadfast support for this line of research. We also would like to thank Scott Cotton for providing a PropBank library that greatly simplified our implementation. Thanks also to the anonymous reviewers for their invaluable comments. This work is supported by the NSF ITR via grant 130-1303-4-541984-XXXX-2000-1070.

References

Baker, Collin F., Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of COLING/ACL*, pages 86–90, Montreal, Canada.

Baldewein, Ulrike, Katrin Erk, Sebastian Padó, and Detlef Prescher. 2004. Semantic role labelling with similarity-based generalization using em-based clustering. In Rada Mihalcea and Phil Edmonds, editors, *Proceedings of Senseval-3*, pages 64–68, Barcelona, Spain.

Bejan, Cosmin Adrian, Alessandro Moschitti, Paul Morărescu, Gabriel Nicolae, and Sanda Harabagiu. 2004. Semantic parsing based on framenet. In Rada Mihalcea and Phil Edmonds, editors, *Proceedings of Senseval-3*, pages 73–76, Barcelona, Spain.

Bender, Emily. 2000. The syntax of Mandarin -ba. *Journal of East Asian Linguistics*, 9(2):105–145.

Bikel, Daniel M. 2004. *On the Parameter Space of Generative Lexicalized Statistical Parsing Models*. Ph.D. thesis, University of Pennsylvania.

Boas, Hans C. 2002. Bilingual FrameNet dictionaries for machine translation. In *Proceedings of LREC 2002*, pages 1364–1371, Las Palmas, Spain.

Burchardt, A., K. Erk, A. Frank, A. Kowalski, S. Pado, and M. Pinkal. 2006. The SALSA Corpus: a German corpus resource for lexical semantics. In *Proceedings of LREC 2006*, pages 969–974, Genoa, Italy.

Carreras, Xavier and Lluís Màrquez. 2004a. Hierarchical recognition of propositional arguments with perceptrons. In *Proceedings of the Eighth Conference on Natural Language Learning*, pages 106–109, Boston, MA.

Carreras, Xavier and Lluís Màrquez. 2004b. Introduction to the CoNLL-2004 shared task: Semantic role labeling. In *Proceedings of the Eighth Conference on Natural Language Learning*, pages 89–97, Boston, MA.

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*, pages 152–164, Ann Arbor, MI.

Chen, Ying, Hongling Sun, and Dan Jurafsky. 2005. A corrigendum to Sun and Jurafsky (2004) "Shallow Semantic Parsing of Chinese." Technical Report TR-CSLR-2005-01, University of Colorado at Boulder CSLR Tech Report.

Cohn, Trevor and Philip Blunsom. 2005. Semantic role labeling with tree conditional random fields. In *Proceedings of CoNLL2005*, pages 169–172, Ann Arbor, MI.

Collins, Michael. 1999. *Head-driven Statistical Models for Natural Language Parsing*. Ph.D. thesis, University of Pennsylvania.

Erk, K. and S Padó. 2005. Analyzing models for semantic role assignment using confusability. In *Proceedings of HLT-EMNLP*, pages 891–898, Vancouver, British Columbia, Canada.

Gabbard, Ryan, Seth Kulick, and Mitchell Marcus. 2006. Fully parsing the Penn treebank. In *Proceedings of HLT-NAACL 2006*, pages 184–191, New York, NY.

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

Girju, Roxana, Ana-Maria Giuglea, Marian Olteanu, Ovidiu Fortu, Orest Bolohan, and Dan Moldovan. 2004. Support vector machines applied to the classification of semantic relations. In *Proceedings of the HLT/NAACL Workshop on Computational Lexical Semantics*, pages 68–75, Boston, MA.

Hacioglu, Kadri, Sameer Pradhan, Wayne Ward, James H. Martin, and Daniel Jurafsky. 2004. Semantic role labeling by tagging syntactic chunks. In *Proceedings of CoNLL-2004*, pages 110–113, Ann Arbor, MI.

Haghighi, Aria, Kristina Toutanova, and Christopher Manning. 2005. A joint model for semantic role labeling. In *Proceedings of CoNLL*, pages 173–176, Ann Arbor, MI.

Huang, James C. T. 1999. Chinese passives in comparative perspective. *Tsing Hua Journal of Chinese Studies*, 29:423–509.

Jiang, Zheng Ping and Hwee Tou Ng. 2006. Semantic role labeling of NomBank: A maximum entropy approach. In *Proceedings of the EMNLP*, pages 138–145, Sydney, Australia.

Kipper, K., A. Korhonen, N. Bryant, and M. Palmer. 2006. Extending verbNet with novel verb classes. In *Proceedings of LREC*, Genoa, Italy.

Koomen, Peter, Vasin Punyakanok, Dan Roth, and Wen tau Yih. 2005. Generalized inference with multiple semantic role labeling systems. In *Proceedings of the Nineth Conference on Natural Language Learning*, pages 181–184, Ann Arbor, MI.

Kwon, Namhee, Michael Fleischman, and Eduard Hovy. 2004. Senseval automatic labeling of semantic roles using Maximum Entropy models. In Rada Mihalcea and Phil Edmonds, editors, *Proceedings of Senseval-3*, pages 129–132, Barcelona, Spain.

Kwong, Oi Yee and Benjamin K. T'sou. 2005. Data homogeneity and semantic role tagging in Chinese. In *Proceedings of the*

ACL-SIGLEX Workshop on Deep Lexical Acquisition, pages 1–9, Ann Arbor, MI.
Lapata, Maria. 2002. The disambiguation of nominalizations. *Computational Linguistics*, 28(3):357–388.
Levi, Judith. 1979. *The Syntax and Semantic of Complex Nominals*. New York: Academic Press.
Levin, Beth. 1993. *English Verbs and Alternations: A Preliminary Investigation*. Chicago: The Unversity of Chicago Press.
Litkowski, Ken. 2004. Senseval-3 task: Automatic labeling of semantic roles. In Rada Mihalcea and Phil Edmonds, editors, *Proceedings of Senseval-3*, pages 9–12, Barcelona, Spain.
Liu, T., W. Che, S. Li, Y. Hu, and H. Liu. 2005. Semantic role labeling with a maximum entropy classifier. In *Proceedings of CoNLL-2005*, pages 189–192, Ann Arbor, MI.
Luo, Xiaoqiang. 2003. A maximum entropy Chinese character-based parser. In *Proceedings of EMNLP*, Sapporo, Japan.
Marcus, Mitchell P., Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.
Màrquez, Lluís, Mihai Surdeanu, Pere Comas, and Jordi Turmo. 2005. A robust combination strategy for semantic role labeling. In *Proceedings of HLT/EMNLP 2005*, pages 644–651, Vancouver, Canada.
McCallum, Andrew Kachites. 2002. Mallet: A machine learning for language toolkit. Available at http://mallet.cs.umass.edu.
Melli, G., Y. Wang, Y. Liu, M. M. Kashani, Z. Shi, B. Gu, A. Sarkar, and F. Popowich. 2005. Description of SQUASH, the SFU question and summary handler for the DUC-2005 summarization task. In *Document Understanding Conference 2005*, Vancouver, Canada.
Meyers, A., R. Reeves, C. Macleod, R. Szekely, V. Zielinska, B. Young, and R. Grishman. 2004. The NomBank Project: An interim report. In *Proceedings of the NAACL/HLT Workshop on Frontiers in Corpus Annotation*, pages 24–31, Boston, MA.
Meyers, A., R. Reeves, and Catherine Macleod. 2004. NP-External arguments: A study of argument sharing in English. In *The ACL 2004 Workshop on Multiword Expressions: Integrating Processing*, pages 96–103, Barcelona, Spain.
Moldovan, Dan and Adriana Badulescu. 2005. A semantic scattering model for the automatic interpretation of genitives. In *Proceedings of HLT-EMNLP*, pages 891–898, Vancouver, Canada.
Moldovan, Dan, Roxana Girju, Marian Olteanu, and Ovidiu Fortu. 2004. SVM classification of FrameNet semantic roles. In *Proceedings of Senseval-3*, pages 167–170, Barcelona, Spain.
Nakov, Preslav and Marti Hearst. 2006. Using verbs to characterize noun-noun relations. In *Proceedings of AIMSA*, pages 233–244, Varna, Bulgaria.
Narayanan, Srini and Sanda Harabagiu. 2004. Question answering based on semantic structures. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 693–701, Geneva, Switzerland.
Ngai, Grace, Dekai Wu, Marine Carpuat, Chi-Shing Wang, and Chi-Yung Wang. 2004. Semantic role labeling with boosting, svms, maximum entropy, snow, and decision lists. In Rada Mihalcea and Phil Edmonds, editors, *Proceedings of Senseval-3*, pages 183–186, Barcelona, Spain.
Palmer, Martha, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.
Pradhan, Sameer, Kadri Hacioglu, Wayne Ward and James H. Martin, and Daniel Jurafsky. 2005. Semantic role chunking combining complementary syntactic views. In *Proceedings of CoNLL 2005*, pages 217–220, Ann Arbor, MI.
Pradhan, Sameer, Kadri Hacioglu, Wayne Ward, James H. Martin, and Daniel Jurafsky. 2003. Semantic role parsing: Adding semantic structure to unstructured text. In *Proceedings of the International Conference on Data Mining (ICDM-2003)*, pages 629–632, Melbourne, FL.
Pradhan, Sameer, Honglin Sun, Wayne Ward, James H. Martin, and Daniel Jurafsky. 2004. Parsing arguments of nominalizations in English and Chinese. In *Proceedings of NAACL-HLT 2004*, pages 141–144, Boston, MA.
Pradhan, Sameer, Wayne Ward, Kadri Hacioglu, James H. Martin, and Dan Jurafsky. 2005. Semantic role labeling using different syntactic views. In *Proceedings of ACL 2005*, pages 581–588, Ann Arbor, MI.
Pradhan, Sameer, Wayne Ward, Kadri Hacioglu, James H. Martin, and Daniel Jurafsky. 2004. Shallow semantic parsing using support vector machines. In

Proceedings of NAACL-HLT 2004, pages 233–240, Boston, MA.

Punyakanok, Vasin, Dan Roth, and W. Yih. 2005. The necessity of syntactic parsing for semantic role labeling. In *Proceedings of IJCAI-2005*, pages 1124–1129, Edinburgh, UK.

Punyakanok, Vasin, Dan Roth, Wen Tau. Yih, and Dav Zimak. 2004. Semantic role labeling via integer programming inference. In *Proceedings of COLING-2004*, pages 1346–1352, Geneva, Switzerland.

Sgall, Petr, Jarmila Panevová, and Eva Hajičová. 2004. Deep syntactic annotation: Tectogrammatical representation and beyond. In A. Meyers, editor, *Proceedings of the HLT-NAACL 2004 Workshop: Frontiers in Corpus Annotation*, pages 32–38, Boston, MA.

Sun, Honglin and Daniel Jurafsky. 2004. Shallow semantic parsing of Chinese. In *Proceedings of NAACL 2004*, pages 249–256, Boston, MA.

Surdeanu, Mihai, Sanda Harabagiu, John Williams, and Paul Aarseth. 2005. Using predicate-argument structures for information extraction. In *Proceedings of ACL 2003*, Ann Arbor, MI.

Thompson, Cynthia, Siddharth Patwardhan, and Carolin Arnold. 2004. Generative models for semantic role labeling. In Rada Mihalcea and Phil Edmonds, editors, *Proceedings of Senseval-3*, pages 235–238, Barcelona, Spain.

Toutanova, Kristina, Aria Haghighi, and Christopher Manning. 2005. Joint learning improves semantic role labeling. In *Proceedings of ACL-2005*, pages 589–596, Ann Arbor, MI.

Tsai, Tzong-Han, Chia-Wei Wu, Yu-Chun Lin, and Wen lian Hsu. 2005. Exploiting full parsing information to label semantic roles using an ensemble of ME and SVM via integer linear programming. In *Proceedings of CoNLL-2005*, pages 233–236, Ann Arbor, MI.

Xue, Nianwen. 2006a. Annotating the predicate-argument structure of Chinese nominalizations. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation*, pages 1382–1387, Genoa, Italy.

Xue, Nianwen. 2006b. Semantic role labeling of nominalized predicates in Chinese. In *Proceedings of HLT-NAACL 2006*, pages 431–438, New York, NY.

Xue, Nianwen and Martha Palmer. 2003. Annotating the propositions in the Penn Chinese Treebank. In *The 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 2004 Conference on Empirical Methods in Natural Language Processing*, pages 88–94, Barcelona, Spain.

Xue, Nianwen and Martha Palmer. 2005. Automatic semantic role labeling for Chinese verbs. In *Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence*, pages 1160–1165, 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.

Yi, Szuting, Edward Loper, and Martha Palmer. 2007. Can semantic roles generalize across genres? In *Proceedings of NAACL-2007*, pages 548–555, Rochester, NY.

Yi, Szuting and Martha Palmer. 2005. The integration of syntactic parsing and semantic role labeling. In *Proceedings of CoNLL-2005*, pages 237–240, Ann Arbor, MI.