The Design of Chinese HPSG for Data-Oriented Parsing

Xiangli Wang[1], Shunya Iwasawa[1], Yusuke Miyao[1], Takuya Matsuzaki[1], Jun'ichi Tsujii[123]

[1]Department of Computer Science, University of Tokyo
[2]School of Computer Science, University of Manchester
[3]National Center for Text Mining
Hongo 7-3-1, Bunkyo-ku, Tokyo, 113-0033, Japan
{xiangli, iwasawa, yusuke, matuzaki, tsujii}@is.s.u-tokyo.ac.jp

1 Introduction

Data-oriented methods have been a main stream for broad-coverage and high-accuracy syntactic parsing (Charniak and Johnson, 2005; McDonald and Pereira, 2006; Miyao and Tsujii, 2005). Rather than hand-crafting linguistically motivated grammar rules, these methods automatically acquire (a part of) grammar rules and statistical models from treebanks. Following the great success of the research on English parsing, the same methodology has been applied to Chinese parsing (Levy and Manning, 2003; Wang et al., 2005; Guo et al., 2007), although, for some reason, the accuracy of Chinese parsing has been unsatisfactory to date.

We explore data-oriented Chinese parsing in the framework of Head-Driven Phrase Structure Grammar (HPSG) (Sag et al., 2003). The development process of our Chinese parser basically follows the work on English (Miyao and Tsujii, 2005): we convert an existing Chinese treebank into an HPSG treebank, and obtain a large lexicon and a statistical model from the HPSG treebank.

As a first step towards data-oriented HPSG parsing for Chinese, this paper presents our design of Chinese HPSG. Assuming that we will obtain a large lexicon from a treebank, this work focuses on the establishment of a theoretical framework, i.e., HPSG signs, schemas, etc. Since our goal is wide-coverage parsing, our framework is intended to cover syntactic constructions that appear frequently in real-world text, rather than minor distinctions among rare constructions. Following our formalization of the structure of Chinese sentences, we introduce our design of HPSG signs and schemas, and describe how we explain some particular and essential phenomena in Chinese, such as the *ba*/*bei* constructions and the topic-sentence construction.

The research on Chinese HPSG has been limited to linguistics (Gao 2000; Li 2001), and few computational resources are available. In the framework of Lexical Functional Grammar (LFG), both hand-crafted and data-oriented Chinese parsers have been developed recently (Fang and King 2007; Guo et al. 2007).

2 The Structure of Chinese Sentences

Our formalization of the Chinese sentence structure is based on Sentence Structure Grammar (SSG) (Wang and Miyazaki, 2007). We represent the structure of Chinese sentences in three levels: predicative parts, sentences, and topic-sentences. Figure 2.1 provides the model of predicative parts. The predicate (P) is the head, which may be modified by optional complements[1] (C). The predicate subcategorizes for up to two obejects (O1, O2). Adverbials (Z) may modify the predicate from its left side. Predicates are verbs as well as adjectives, as adjectives in Chinese can be a predicate without copulas.

Figure 2.2 and 2.3 present the model of sentences. Sentences consist of a subject (S) and one or more predicative parts, followed by modal particles (Y) (Dexi, 1982). The predicate subcategorizes for a subject, while modal particles are optional.

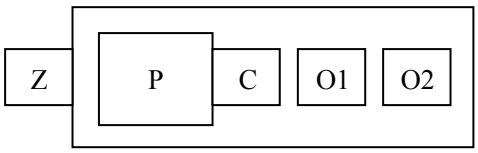

Figure 2.1: The model of predicative parts

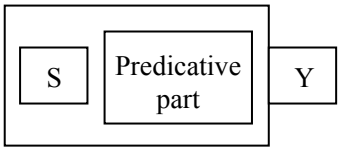

Figure 2.2: The model of sentences with one predicative part

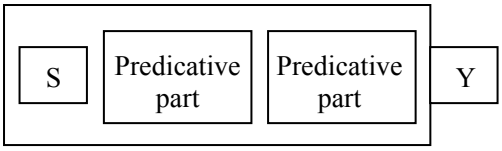

Figure 2.3: The model of sentences with multiple predicative parts

[1] The term *complement* in the Chinese grammar is a different concept from that in HPSG: *complement* refers to a grammatical unit that appears right after a predicate and adds a meaning to it.

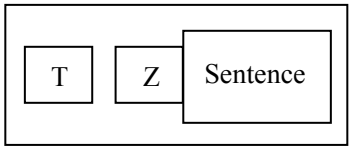

Figure 2.4: The model of topic-sentences

Examples of simple sentence structures are given below:

1) S P

1a. 约翰 吃
John eat
John eats.

1b. 做 饭 很 有趣
cook meal very interesting
Cooking is very interesting.

2) S P O1

2a. 约翰 吃 苹果
John eat apple
John eats apples.

2b. 约翰 学 做 饭
John learn cook meal
John learns cooking.

3) S P O1 O2

3a. 约翰 送 麦克 苹果
John give Mike apple
John gives Mike apples.

The predicate is the syntactic head of the whole sentence, and determines the number and the types of arguments it can take. For example, "吃" (eat) takes one nominal object, while "学" (learn) takes one verb phrase as in sentence 2b. Generally, a subject is a noun phrase, but some predicates take a clause or a verb phrase as a subject, as shown in 1b.

Figure 2.4 presents the model of topic-sentences. A topic (T) appears in the beginning of a sentence, and adverbials may appear between the topic and the main part of the sentence. Examples are given below:

4a. 大象 鼻子 长
elephant nose long
Elephant's nose is long.

4b. 约翰 他 吃 苹果
John he eat apple
John, he eats apples.

3 The Design of Chinese HPSG

In this section, we first describe the definition of signs and schemas in our Chinese HPSG. We then discuss some constructions that are particular and essential in Chinese.

3.1 Signs and Schemas

We define signs and ID schemas basically following the definition by Sag et al. (2003). Figure 3.1 shows a lexical sign for the transitive verb "吃" (eat). PHON is a feature for a surface string of a word. HEAD expresses the characteristics of the head word of a constituent. MOD, SPR, and COMPS represent selectional constraints of a modifiee, left arguments and right arguments. GAP expresses constraints of moved arguments. INDEX and RESTR are the features for expressing semantic structures. INDEX represents the predicate argument structure of the main predicate, and RESTR provides semantic restrictions to the main predicate.

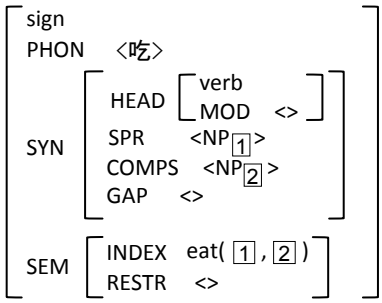

Figure 3.1: A lexical sign for "吃" (eat)

We define the following schemas: Specifier-Head, Head-Complement, Head-Modifier (left-head), Modifier-Head (right-head), Filler-Head, Coordination, and Topic-Sentence. The Topic-Sentence schema is peculiar to our Chinese grammar (the details will be discussed in Section 3.2.8), while the other schemas are the same as the original HPSG theory for English.

In the level of the syntactic structure, several constructions are the same as in English, such as the predicate-object and subject-predicate constructions, although some constructions look similar but have slight differences; for example, adverbial constituents appear only in the left side of the predicate.

3.2 Essential Constructions in Chinese

3.2.1 Predicate-Argument Construction

The syntactic relationships between a predicate and its arguments, i.e., subjects and objects, are very similar to English. We therefore describe these constructions using the Specifier-Head schema for subjects and the Head-Complement schema for objects without any modifications.

3.2.2 Predicate-Complement Construction

As shown in Figure 2.1, complements appear directly after the predicate. Since complements are optional, we analyze the predicate-complement construction with the Head-Modifier schema. The

predicate-complement construction in our grammar includes the following types (examples are given in 5a to 5e):

1) V + complement word
2) V + 得/不 + complement word
3) V + 得 + clause
4) V + aspect particle
5) V + prepositional phrase

5a. 吃 完
eat finish
finish eating

5b. 吃 得 完
eat complement-de finish
can finish eating

5c. 玩 得 很 愉快
play complement-de very happy
play and be happy

5d. 吃 着 饭
eat progressive meal
be eating a meal

5e. 住 在 东京
live in Tokyo
live in Tokyo

Type 1, 2 and 3 are typically explained as complements in Chinese grammatical textbooks. In our grammar, we additionally classify type 4 and 5 into this construction as they have similar syntactic characteristics: aspect particles (type 4), such as "着" (progressive aspect), "了" (perfect aspect) and "过" (past experience), and prepositional phrases that appear after the predicate (type 5).

3.2.3 Adverbial-Predicate Construction

Adverbials are another modifying part of the predicate, which appears before the predicate, including temporal phrases (6a), prepositional phrases (6b), adverbs and "地" phrases (6c). We describe this construction with the Modifier-Head schema.

6a. 约翰 今天 吃 苹果
John today eat apple
John eats apples today.

6b. 约翰 在 家 吃 苹果
John at home eat apple
John eats apples at home.

6c. 约翰 慢慢 地 吃 苹果
John slow adverbial-de eat apple
John eats apples slowly.

Auxiliary verbs are also treated in the same manner with the Modifier-Head schema, because they appear in a similar syntactic position.

3.2.4 Sentence-Particle Construction

In Chinese, one or more modal particles may be added to the end of a sentence, to modify the whole sentence. For example, "了" in 7a expresses past tense, and "吗" in 7b indicates interrogative.

7a. 你 做 饭 了
you cook meal past-tense
You have cooked a meal.

7b. 你 做 饭 了 吗
you cook meal past-tense interrogative
Have you cooked a meal?

This construction is analyzed with the Head-Modifier schema in a similar way to the predicate-complement construction.

3.2.5 *Ba*/*Bei* Constructions

The *ba* and *bei* constructions are particular in Chinese. *Ba* and *bei* can be considered as case-markers, where a constituent taken by *ba* fills the object of the predicate as in 8a, and a constituent taken by *bei* is the subject of the predicate as in 8b.

8a. 约翰 把 苹果 吃 了
John Ba apple eat past-tense
John ate apples.

8b. 苹果 被 约翰 吃 了
apple Bei John eat past-tense
Apples were eaten by John.

As those are arguments of the predicate, and appear in the left side of the predicate, these two types of constituents are treated as the specifiers of the predicate, and we use the Specifier-Head schema to describe the syntactic relationships between the predicate and the *ba*/*bei* phrases. Hence, distinct lexical entries are assigned to predicates with/without *ba* or *bei* phrases.

3.2.6 Pre-Object Construction

The pre-object construction is the construction in which an object of the predicate appears in the beginning of a sentence as a topic. For handling the relationship between the topic and the predicate, we analyze this construction as a movement, and use GAP and the Filler-Head schema as in the analysis of the movement in English.

9a. 苹果 约翰 吃 了
apple John eat past-tense
Apples, John ate.

3.2.7 Multi-Predicate Constructions

The multi-predicate constructions refer to the construction given in Figure 2.3, which can be further classified into three types: the coordinative construction, the serial-verb construction, and the duplicate-predicate construction. In the coordinative construction, multiple predicates represent concurrent events (10a), while in the serial-verb construc-

tion, the events represented by the predicates occur in a time series (10b). The duplicate-predicate construction refers to sentences in which the predicate word is duplicated (10c).

10a. 约翰 唱 歌 跳 舞
John sing song dance dance
John sings and dances.

10b. 约翰 去 学校 看 书
John go school read book
John goes to a school to read books.

10c. 约翰 唱 歌 唱 得 好
John sing song sing complement-marker good
John is good at singing.

At the present stage, we do not distinguish among these three types of constructions, because they are almost identical in the syntactic level. We explain all of these constructions with the Coordination schema. A more proper analysis of these constructions is left as future work.

3.2.8 Topic-Sentence Construction

The topic-sentence construction given in Figure 2.4 is a particular construction in Chinese (4a and 4b). We design the Topic-Sentence schema for analyzing this construction (Figure 3.2). This schema attaches a topic phrase to a sentence, where the right child is restricted to be a completed sentence with an adjectival predicate.

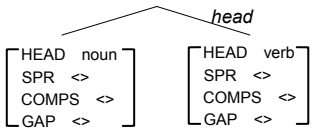

Figure 3.2: The Topic-Sentence schema

3.2.9 P_On_Ov Construction

Some predicates in Chinese take a nominal object and a predicate phrase as arguments, which we call the P_On_Ov construction. The structures of 11a and 11b look similar, but their semantic interpretations are different. The predicate of 11a is a subject-control verb, while the predicate of 11b is an object-control verb. Our grammar distinguishes between these constructions as in the English HPSG, and we obtain different semantic structures for them (11c and 11d).

11a. 约翰 帮 麦克 打 玛丽
John help Mike beat Mary
John beats Mary to help Mike.

11b. 约翰 派 麦克 打 玛丽
John send Mike beat Mary
John sends Mike to beat Mary.

11c. help(John, Mike, beat(John, Mary))

11d. send(John, Mike, beat(Mike, Mary))

3.2.10 Relative Clauses

Relative clauses in Chinese are introduced by the relativizer "的". It is used as the relative pronoun (12a) and as the relative adverb (12b). We therefore define a specialized lexical entry for each usage. The relativizer takes a clause as a specifier and modifies a noun. Note that the relativizer takes a gapped clause in the case of 12a.

12a. 麦克 买 的 书
Mike buy relativizer book
a book that Mike buys

12b. 麦克 玩 球 的 公园
Mike play ball relativizer park
a park where Mike plays ball

4 Conclusion and Future Work

We have presented our framework of Chinese HPSG. Our framework covers Chinese constructions that frequently appear in Chinese real-world text. Currently, we are implementing our grammar and evaluating the coverage of the grammar with sample sentences taken from a Chinese grammar textbook and newswire texts. In future work, we will obtain an HPSG lexicon and a disambiguation model from Penn Chinese Treebank, using the grammatical framework presented in this paper.

References

Charniak, E. and M. Johnson (2005) *Coarse-to-fine n-best parsing and MaxEnt discriminative reranking.* In Proc. ACL 2005.

Zhu, D. (1982) *语法讲义 (Lectures on Grammar*), Beijing: Comercial Press.

Fang, J. and T. H. King (2007) *An LFG Chinese grammar for machine use*. In Proc. GEAF 2007.

Gao, Q. (2000) *Argument Structure, HPSG and Chinese Grammar.* PhD thesis, Ohio State University.

Guo, Y., J. van Genabith and H. Wang (2007) *Treebank-based acquisition of LFG resources for Chinese*. In Proc. LFG 2007.

Levy, R. and C. D. Manning (2003) *Is it harder to parse Chinese, or the Chinese Treebank?* In Proc. ACL 2003.

Li, W. (2001) *The Morpho-Syntactic Interface in a Chinese Phrase Structure Grammar.* PhD thesis, Simon Fraser University.

McDonald, R. and F. Pereira (2006) *Online learning of approximate dependency parsing algorithms*. In Proc. EACL 2006.

Miyao, Y. and J. Tsujii (2005) *Probabilistic disambiguation models for wide-coverage HPSG parsing.* In Proc. ACL 2005.

Sag, I., T. Wasow and E. Bender (2003) *Syntactic Theory: A Formal Introduction.* University of Chicago Press.

Wang, Q., D. Schuurmans, and D. Lin (2005) *Strictly lexical dependency parsing.* In Proc. 9th IWPT.

Wang, X. and M. Miyazaki (2007) *文構造文法に基づく中国語構文解析 (Chinese syntactic analysis using sentence structure grammar).* Journal of Natural Language Processing, vol.14 no.2, pp.69-93.