Corpus-oriented Acquisition of Chinese Grammar

Yan Zhang
ATR Spoken language Communication Research Laboratories
2-2-2 Keihanna Science City, Kyoto, 619-0288
yan.zhang@atr.jp

Hideki Kashioka
ATR Spoken language Communication Research Laboratories
2-2-2 Keihanna Science City, Kyoto, 619-0288
Hideki.kashioka@atr.jp

Abstract

The acquisition of grammar from a corpus is a challenging task in the preparation of a knowledge bank. In this paper, we discuss the extraction of Chinese grammar oriented to a restricted corpus. First, probabilistic context-free grammars (PCFG) are extracted automatically from the Penn Chinese Treebank and are regarded as the baseline rules. Then a corpus-oriented grammar is developed by adding specific information including head information from the restricted corpus. Then, we describe the peculiarities and ambiguities, particularly between the phrases "PP" and "VP" in the extracted grammar. Finally, the parsing results of the utterances are used to evaluate the extracted grammar.

1 Introduction

Research and development work on spoken language systems for special domains has been gaining more attention in recent years. Many approaches to spoken language processing require a grammar system for parsing the input utterances in order to obtain the structures, especially for rule-based approaches.

Manually developing grammars based on linguistics theories is a very difficult task. Language phenomena are usually described as being symbolic systems such as lexical, syntactic, semantic and common sense. Grammar development has to depend on linguistic knowledge and the characteristics of the corpus to explicate a system of linguistic entities. However, it is expensive and time-consuming to maintain a robust grammar system by manual writing.

Recently some researchers (H. Meng et al., 2002; S. Dipper, 2004 and Y. Ding, 2004) have presented a methodology to semi-automatically capture different grammar inductions from annotated corpora within restricted domains. A corpus-oriented approach (Y. Miyao, 2004) provides a way to extract grammars automatically from an annotated corpus. The specific language knowledge and knowledge relations need to be constructed and oriented to different corpora and tasks (K. Chen, 2004).

The Chinese treebank is a useful resource for acquiring grammar rules and the context relations. Currently there are several Chinese treebanks on a scale of size. In the Penn Chinese Treebank (F. Xia, 2000), each structural tree is annotated with words, parts-of-speech and syntactic structure brackets. In the Sinica Treebank (CKIP), thematic roles are also labeled in the CKIP to provide deeper information.

In this paper, we discuss the extraction of Chinese grammar oriented to a restricted corpus. First, probabilistic context-free grammars (PCFG) are extracted automatically from the Penn Chinese Treebank and are regarded as the baseline rules. Then a corpus-oriented grammar is developed by adding specific information including head information from the restricted corpus. We then describe the peculiarities and ambiguities, especially between the phrases "PP" and "VP" in the extracted grammar. Fi-

nally, the parsing results of the utterances are used to evaluate the extracted grammar.

The outline of this paper is as follows: Section 2 gives the process of acquiring the baseline Chinese grammar and the extension of the current grammar oriented to the corpus. Section 3 explains the grammar properties in our corpus and our approach to disambiguating some special phrase rules, such as "PP" and "VP" and the word "在(ZAI)" in different categories. Section 4 describes the evaluation results of the extracted Chinese grammar. Finally section 5 offers some concluding remarks and outlines our future work.

2 Grammar Acquisition

There are two parts to acquiring grammar in our system. The baseline grammar is extracted automatically from the Penn Chinese Treebank. We define a suitable parts-of-speech and phrase categories and extend them by introducing specific information from our corpus.

2.1 Grammar Extraction from Penn Chinese Treebank

The University of Pennsylvania (Upenn) has released a scale of Chinese treebanks as a kind of resource since 2000 (Xia Fei et al., 2000). Each structural tree includes parts-of-speech and syntactic structure brackets, which provides a good way to extract Chinese probabilistic context-free grammars (PCFG). There are a total of 325 files collected from the Xinhua newswire in this treebank. The majority of these documents focus on economic development and are organized in written formats as opposed to spoken utterances, so the grammars extracted from it are seen as the baseline bank.

The probabilistic context-free grammars have proven to be very effective for parsing natural language. The produced rules are learnt by matching the bracketed structures automatically from the trees, and the rule probabilities are calculated based on the maximum likelihood estimation (MLE), presented in the following formula (Charniak, 1996):

$$P(N^i \to \zeta^j) = \frac{C(N^i \to \zeta^j)}{\sum_k C(N^i \to \zeta^k)} \quad (1)$$

The baseline grammar includes about 400 PCFG rules after cleaning and merging some rules with low probabilities (Imamula et al., 2003).

2.2 Extension of the Extracted Grammar

Different corpora produce different grammars that have some specific information. In baseline grammars, many grammars are not suitable for spoken corpora. Therefore, we need to build an appropriate grammar by using specific information in our corpus to improve the parsing results and machine translation systems that operates in a restricted field. The data we used in this system is from the ATR Basic Travel Expression Corpus (BTEC) in which the format of utterances is different from the sentences in Upenn.

Consequently, an appropriate phrase category is required to be constructed by analyzing the knowledge characteristics in BTEC. We define it by comparing three Chinese structure category systems: Sinica, University of Pennsylvania, and HIT (Harbin Institute of Technology). A phrase category should be not too complicated as but cover language phenomenon in the corpus. Our phrase category is defined and explained in table 1.

Categories	Explanation
NNP	Noun Phrase
TNP	Temporal Noun Phrase
LP	Localizer Phrase
NSP	Location Phrase
VP	Verb Phrase
AP	Adjective Phrase
DP	Adverbial Phrase
QP	Quantifier Phrase
PP	Preposition Phrase
VBAP	Phrase with "把 (BA)"
DENP	Nominal Phrase Ending by "的 (DE)"
DEP	Attributive Phrase formed by "的 (DE)"

Table 1 Phrase Categories

In BTEC, Chinese utterances are segmented and labeled as parts-of-speech. We not only construct corpus-oriented grammar rules differently from the baseline grammars but also add head information for each rule.

In the above Table 1, the phrase category "VBAP" is a phrase name including the preposition "把(BA)" and its following noun or verb phrase. The phrase "DENP" is a special nominal phrase which has no word after the auxiliary

word “的 (DE)”, and it is usually put at the end of the utterance. Following are some examples of our grammars.

1. PP → p(sem"和") (head)n
2. DENP → (head)a y(sem"的")
3. PP → p(sem"用") (head)r
4. DEP → (head)DP de

In above rules, the mark “sem” means its following word is a terminal node.

3 Grammar Annotation and Disambiguation

Above constructed Chinese grammars sometimes bring out conflicts when parsing utterances because of the ambiguity phenomenon. Grammar annotation is done to make the grammatical relations of an utterance more explicit. Thus, some ideas are proposed to deal with these ambiguities that are tightly related to Chinese language.

3.1 Annotation and Analysis of Grammar

Plenty of prepositions are rooted in verbs in Chinese language, and most of them still keep the function of verbs. This phenomenon produces ambiguous problems not only between categories preposition “p” and verb “v” but between the phrases “VP” and “PP” in the structures of the utterances. PP-attachment ambiguity is a big problem related to the construction of grammar (S. Zhao. 2004).

Firstly, we extract a lexicon of Chinese prepositions, which have other categories at the same time, such as the category ‘v’, adjective ‘a’, and so on. The following table shows the collocations of these words and their frequencies.

Word	Category	Frequency
按	p	226
	v	85
到	p	2423
	vt	4857
对	p	579
	a	1058
给	p	6422
	v	4309
用	p	1270
	v	1226
在	p	11115
	v	2381
	d	39

Table 2 Some Examples in the Preposition Lexicon

We construct some particular grammar rules for these preposition words showed in Table 2 in order to deal with the conflicts among these words. For example, following rules are related to the word “给”.

PP → p(sem"给") (head)n
VP → p(sem"给") (head)V
VP → v(sem"给") NNP (head)VP

In order to represent the function of the extracted grammar, we compare the coverage of the grammar in different layers between a terminal node and a phrase layer. The different structural trees of the same utterance in Figure 1 are listed as follows.

1.Sentence (这里/r 的/de 鞋/n 哪/r 双/q 是/v1 特价/n 商品/n ？ /w)

```
|__ NNP__NNP(head)__DEP__r(head) 这里
|   |          |          | ___ de 的
|   |          | _________NNP __ n(head) 鞋
|   |
|   | ___ QP __ r(sem"哪")
|         | ____q(head) 双
|
| __ v1(sem"是")
|
| __ NNP(head) __ NNP _____n(head) 特价
|      | _________NNP(head)__n(head) 商品
| __ w ？
```

2. Sentence (这里/r 的/de 鞋/n 哪/r 双/q 是/v1 特价/n 商品/n ？ /w)

```
|__NNP__NNP(head)*__ r 这里
|   |          | ___ de 的
|   |          | ____  n(head) 鞋
|   |
|   | ___ QP __ r(sem"哪")
|         | __ q(head) 双
|
| __ v1(sem"是")
|
| __ NNP(head)** __ n 特价
|      | __________ n(head) 商品
| __ w ？
```

Figure 1 Annotation of Different trees in the same sentence

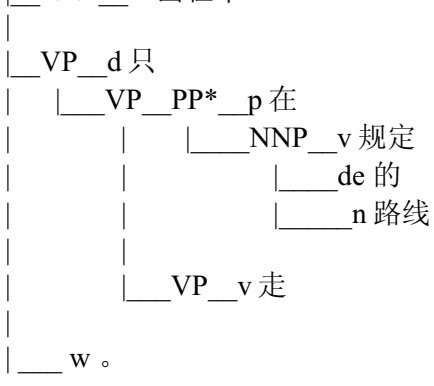
The same utterance obtains different structural trees from different levels of grammar rules by parsing results, although these two trees are cor-

rect and acceptable. The grammar plays an important role in the machine translation system when we build the mapping relations with the goal languages by transform rules. This problem is also called Granularity (K. Chen, 2004). Symbol '**' in Figure 1 denotes that the phrase "NNP" is produced by the rule "NNP → n (head)n" rather than "NNP → NNP (head)NNP".

3.2 Grammar Disambiguation

A grammar inevitably includes ambiguities among its rules. To some extent, certain kinds of ambiguities are produced by the same ambiguous problems found among part-of-speech tags. As with the expression in Section 2, the ambiguity between the phrases "PP" and "VP" is partly produced by the multiple categories 'p' and 'v' of the words. This is a common case where the phrases "PP" and "VP" are nested against each other. For example, the rule "PP → p (head)v" and "VP → PP (head)VP". This situation is described in the following two utterances in Figure 2.

1. Sentence (出租车/n 只/d 在/p 规定/v 的/de 路线/n 走/v 。)



2. sentence (请/vw 给/p 我/r 送/v 杯/q 咖啡/n。)

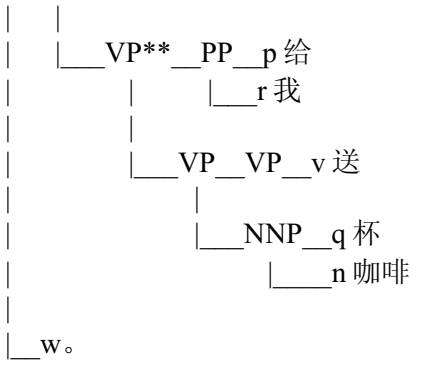


Figure 2 The Correct Trees of Utterances Including phrases "PP" and "VP"

In sentence 1 of figure 2, the phrase "PP" (在/p 规定/v 的/de 路线/n) contains the verb "规定", and is produced by the rule "DEP → v de" and "NNP → DEP n" firstly. Likewise, in sentence 2, the phrase "VP(送杯咖啡)" is produced firstly rather than phrase "给我送" is got by rule "VP → PP v". That is to say, the phrase "VP" has higher priority to be produced than "PP".

The Chinese word "在" is a special individual word in our corpus. Its correlative disambiguation rules are constructed by the knowledge relations listed in the following table:

Category of " 在 (ZAI)"	The order of rules	Ambiguity parts bracketed in utterance
P (preposition)	1. VP → V(sem" 在 ") (head)r 2. VP → PP (head)VP	我/r 一直/d [[在/p 这里/r] [等/v 电话/n]]。
V (verb)	1. VP → V(sem" 在 ") (head)r 2. VP D (head)VP	我/r [一直/d 在/v 学校/n]
D (adverb)	1. VP → D(sem" 在 ") (head)VP 2. DP → D (head)d	我/r [一直/d 在/d 想/v 着/u 你/n]

Table 3 The characteristics of word "在"

The following steps are used to identify the ambiguities between the phrases "PP" and "VP":

1. The first step is to look up the preposition lexicon based on the category of the word and find the relative rules from the extracted grammar.
2. When the "PP" rules conflict with the "VP" rules, we firstly consider the verb and then select an appropriate rule by comparing the relationship to neighboring preposition words.
3. Long distance rules have priority. For instance, rule "PP → p v nd" is preferred to rule "PP → p v".
4. It is clear that the fine-grained rules have less representational ambiguity than the coarse-

grained grammar rules in relation to the tree presentations.
5. The head information in the rules is viewed as being types of reference knowledge because of their own ambiguities.

4 Evaluation for Grammar

We use the extracted grammar described in section 3 to parse Chinese utterances in BTEC and to evaluate the roles of the grammar.

4.1 Parsing with Grammar

The parser adopts a bottom-up parsing algorithm in order to obtain the phrase structures of utterances. There are 200 Chinese utterances selected in our experiment. The number of rules totals 682 that are constructed manually except baseline rules from Upenn Chinese treebank. Table 4 lists the number of PCFG rules which have different left-side phrases.

Left-side phrase	frequency	Proportion as head information
AP	42	15
DENP	20	2
DEP	15	2
DP	9	5
LP	10	3
NNP	240	114
NSP	18	2
PP	39	1
QP	50	17
TNP	28	15
VBAP	7	0
VP	162	106
sentence	40	--

Table 4 The number of rules with different left-side phrases

In our current experiment, the evaluation is limited to obtaining several special phrase structures including "NNP", "VP", "PP", and "DENP" by using the extracted grammar. Therefore, the parsing results are calculated using the precision of these phrases in the following formula (2) and are listed in Table 5. We give the evaluation results of the word "在" separately in Table 6.

$$\Pr ec(phrase) = \frac{N_c}{N_t} \times 100\% \quad (2)$$

where N_c denotes the number of correct phrases in the parsing results, and N_t is the total number of the phrases in the utterances.

Phrase	Precision without disambiguation	Precision with disambiguation
Prec(NNP)	70.03	70.43
Prec(PP)	81.51	84.17
Prec(VP)	69.01	70.13
Prec(DENP)	82.61	82.81

Table 5 The evaluation results

Phrase with "在"	Precision without disambiguation	Precision with disambiguation
Prec(PP)	79.12	83.67
Prec(VP)	89.34	91.72
Prec(DP)	87.71	88.02

Table 6 The evaluation results of "在"

From the evaluation results, we found that the precisions of the phrases "NNP" and "VP" were not high due to the diversity and complexity. We only processed the ambiguity between "VP" and "PP" and improve the precision of phrase "PP". From the condition of the word "在", it is very useful for the grammar extraction to construct information on high-frequency words and word-to-word collocation relations.

4.2 Discussion

The Chinese language is one of the most difficult languages to process. There is still no uniform standard for acquiring Chinese grammar that covers all domains. Hence, a grammar should be constructed from the view of point of real research requirements in real corpora. It is the most important to maintain consistency and satisfy the actual requirements of a real corpus.
One of the main purposes in constructing a Chinese grammar is to improve its validity and robustness to machine translation in a restricted corpus. The development of a robust grammar based on linguistics is difficult because of the complexity of deep linguistic analysis. For example, how many annotated grammars are suitable for the parsing system and a real machine translation? What is the balance between the granularity of grammar structures and grammar

coverage including the ambiguities? In general, the coarse-grained grammar rules have a higher coverage rate compared with fine-grained rules, which contain more terminal nodes. There is also the major problem of determining which Treebank size is required to acquire the grammar rules.

5 Conclusion and Future Work

Corpus-oriented grammar extraction is conducted for the purpose of constructing more explicit grammar knowledge and improving the machine translation system in a restricted corpus. Treebanks provide a useful resource for acquiring grammar rules. However, it is time consuming to construct a much larger size Treebank, which is better for grammar extraction. It would be better if the knowledge extraction process could be carried out iteratively. The parser could use the initial grammar to produce a large amount of structural trees. These new trees would provide more information on the grammar to improve the robustness of the grammar and the power of the parsing system. This whole process can be regarded as an automatic knowledge learning system.

The principal idea in this paper was to acquire Chinese grammar from a restricted corpus for a machine translation system. The extracted grammar was not only from the Penn Chinese treebank but also from new information added to our experimental corpus. The corpus-oriented Chinese grammar was evaluated by parsing the phrase structures that includes “NNP”, “VP”, “PP”, “DENP”, and the phrases relative to the word “在”.

Currently, we only focus on a few limited phrases, and the disambiguation process has been explored with specific rules manually. Therefore, to improve grammar extraction in the future, we will aim at increasing the robustness and coverage of the rules and try to automatically reduce the ambiguity rate by constructing more knowledge relations. The word-to-word collocation relations provided useful information on grammar extraction for the detailed processing.

Acknowledgment

This research was supported by a contract with the National Institute of Information and Communication Technology (NICT) of Japan.

References

Helen M. Meng and Kai-Chung Siu. 2002. Semi-Automatic Acquisition of Domain-Specific Semantic Structures, IEEE Transactions on Knowledge and Data Engineering, vol 14, n 1, January/February, pp. 172-180

Stefanie Dipper. Grammar Modularity and its Impact on Grammar Documentation. In Proceedings of the 20th International Conference on Computational Linguistics (COLING), pp. 1-7, Geneva, Switzerland, 2004

Claire Gardent, Marilisa Amoia and Evelyne Jacquey. Paraphrastic Grammars. ACL Workshop on text meaning, Barcelona, July 2004

Yuan Ding and Martha Palmer. Automatic Learning of Parallel Dependency Treelet Pairs. In Proceedings of the First International Joint Conference on Natural Language Processing (IJCNLP2004). March, Sanya, pp. 30-37, 2004

Shaojun Zhao and Dekang Lin. A nearest-neighbor method for resolving pp-attachment ambiguity. In Proceedings of the First International Joint Conference on Natural Language Processing (IJCNLP2004). March, Sanya, pp. 428-434, 2004

Kenji Imamura, Eiichiro Sumita and Yuji Matsumoto. 2003. Feedback Cleaning of Machine Translation Rules Using Automatic Evaluation. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL 2003), pp. 447-454.

Keh-Jian Chen and Yu-Ming Hsieh. Chinese Treebank and Grammar Extraction. In Proceedings of the First International Joint Conference on Natural Language Processing (IJCNLP2004). March, Sanya, pp. 560-565, 2004

CKIP (Chinese Knowledge Information Processing). 1993. The Categorical Analysis of Chinese. [In Chinese]. CKIP Technical Report 93-05. Nankang: Academic Sinica.

Fei Xia, Martha Palmer, Nianwen Xue, Mary Ellen Okurowski, John Kovarik, Fudong Chiou, Shizhe Huang, Tony Kroch, and Mitch Marcus. 2000. Developing Guidelines and Ensuring Consistency for Chinese Text Annotation. Proceeding of the second International Conference on Language Re-

sources and Evaluation (LREC-2000), Athens, Greece.

Rashmi Prasad, Elini Miltsahaki, Aravind Joshi and Bonnie Webber. Annotation and Data Mining of the Penn Discourse Treebank. In Proceedings of the ACL 2004 Workshop on Discourse Annotation, Barcelona. 2004.

Yusuke Miyao, Takashi Ninomiya and Jun'ichi Tsujii. Corpus-oriented Grammar Development for Acquiring a Head-driven Phrase Structure Grammar from the Penn Treebank. In Proceedings of the First International Joint Conference on Natural Language Processing (IJCNLP2004). March, Sanya, pp. 390-397, 2004

E. Charniak. 1996. Treebank Grammars. Technical Report CS-96-02, Department of Computer Science, Brown University.