PCFG Parsing for Restricted Classical Chinese Texts

Liang HUANG
Department of Computer Science,
Shanghai Jiaotong University
No. 1954 Huashan Road, Shanghai
P.R. China 200030
lhuang@sjtu.edu.cn

Yinan PENG
Department of Computer Science,
Shanghai Jiaotong University
No. 1954 Huashan Road, Shanghai
P.R. China 200030
ynpeng@sjtu.edu.cn

Huan WANG
Department of Chinese Literature and Linguistics,
East China Normal University
No. 3663 North Zhongshan Road, Shanghai,
P.R. China 200062

Zhenyu WU
Department of Computer Science,
Shanghai Jiaotong University
No. 1954 Huashan Road, Shanghai
P.R. China 200030
neochinese@sjtu.edu.cn

Abstract

The Probabilistic Context-Free Grammar (PCFG) model is widely used for parsing natural languages, including Modern Chinese. But for Classical Chinese, the computer processing is just commencing. Our previous study on the part-of-speech (POS) tagging of Classical Chinese is a pioneering work in this area. Now in this paper, we move on to the PCFG parsing of Classical Chinese texts. We continue to use the same tagset and corpus as our previous study, and apply the bigram-based forward-backward algorithm to obtain the context-dependent probabilities. Then for the PCFG model, we restrict the rewriting rules to be binary/unary rules, which will simplify our programming. A small-sized rule-set was developed that could account for the grammatical phenomena occurred in the corpus. The restriction of texts lies in the limitation on the amount of proper nouns and difficult characters. In our preliminary experiments, the parser gives a promising accuracy of 82.3%.

Introduction

Classical Chinese is an essentially different language from Modern Chinese, especially in syntax and morphology. While there has been a number of works on Modern Chinese Processing over the past decade (Yao and Lua, 1998a), Classical Chinese is largely neglected, mainly because of its obsolete and difficult grammar patterns. In our previous work (2002), however, we have stated that in terms of computer processing, Classical Chinese is even easier as there is no need of word segmentation, an inevitable obstacle in the processing of Modern Chinese texts. Now in this paper, we move on to the parsing of Classical Chinese by PCFG model. In this section, we will first briefly review related works, then provide the background of Classical Chinese processing, and finally give the outline of the rest of the paper.

A number of parsing methods have been developed in the past few decades. They can be roughly classified into two categories: rule-based approaches and statistical approaches. Typical rule-based approaches as described in James (1995) are driven by grammar rules. Statistical approaches such as Yao and Lua (1998a), Klein and Manning (2001) and Johnson, M. (2001), on the other hand, learn the parameters the distributional regularities from a usually large-sized corpus. In recent years, the statistical approaches have been more successful both in part-of-speech tagging and parsing. In this paper, we apply the PCFG parsing with context-dependent probabilities.

A special difficulty lies in the word segmentation for Modern Chinese processing. Unlike Indo-European languages, Modern Chinese words are written without white spaces indicating the gaps between two adjacent words. And different possible segmentations may cause consistently different meanings. In this sense, Modern Chinese is much more ambiguous than those Indo-European Languages and thus more difficult to process automatically (Huang et al., 2002).

For Classical Chinese processing, such segmentation is largely unnecessary, since most Classical Chinese words are single-syllable and single-character formed. To this end, it is easier than Modern Chinese but actually Classical Chinese is even more ambiguous because more than half of the words have two or more possible lexical categories and dynamic shifts of lexical categories are the most common grammatical phenomena in Classical Chinese. Despite of these difficulties, our work (2002) on part-of-speech tagging has shown an encouraging result.

The rest of the paper is organized as follows. In Section 1, a tagset designed specially for Classical Chinese is introduced and the forward-backward algorithm for obtaining the context-dependent probabilities briefly discussed. We will briefly present the traditional two-level PCFG model, the syntactic tagset and CFG rule-set for Classical Chinese in Section 2. Features of the Classical Chinese grammar will also be covered in this section. In Section 3 we will present our experimental results. A summary of the paper is given in the conclusion section.

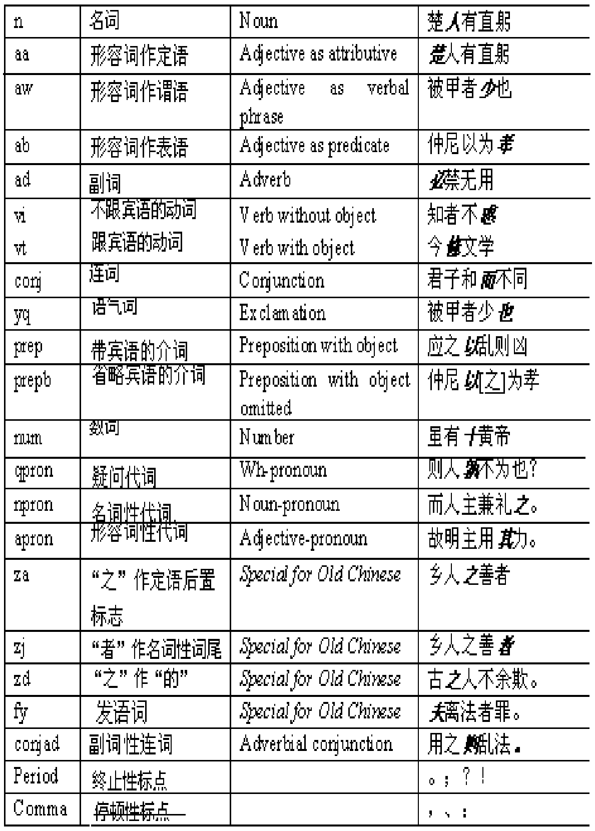
1 Tagset and Context-Dependent Probabilities

Generally speaking, the design of tagset is very crucial to the accuracy and efficiency of tagging and parsing, and this was commonly neglected in the literature where many researchers use those famous corpora and their tagset as the standard test-beds. Still there should be a tradeoff between accuracy and efficiency. In our previous work (2002), a small-sized tagset for Classical Chinese is presented that is shown to be accurate in their POS tagging experiments. We will continue to use their tagset in this paper. We will also use a forward-backward algorithm to obtain the context-dependent probabilities.

1.1 Tagset

The tagset was designed with special interest not only to the lexical categories, but also the categories of components, namely *subcategories* a word may belong. For example, it discriminates adjectives into 4 subcategories like *Adjective as attributive*, etc. (See table 1). And several grammatical features should be reflected in the tagset. These discriminations and features turn out to be an important contributing factor of the accuracy in our parsing experiments.

Table 1. The tagset for Classical Chinese



n	名词	Noun	楚**人**有直躬
aa	形容词作定语	Adjective as attributive	**楚**人有直躬
aw	形容词作谓语	Adjective as verbal phrase	被甲者**少**也
ab	形容词作表语	Adjective as predicate	仲尼以为**孝**
ad	副词	Adverb	**必**禁无用
vi vt	不跟宾语的动词 跟宾语的动词	Verb without object Verb with object	知者不**惑** 今**修**文学
conj	连词	Conjunction	君子和**而**不同
yq	语气词	Exclamation	被甲者少**也**
prep	带宾语的介词	Preposition with object	应之**以**乱则凶
prepb	省略宾语的介词	Preposition with object omitted	仲尼**以**[之]为孝
num	数词	Number	星有**十**黄帝
qpron	疑问代词	Wh-pronoun	则人**孰**不为也?
rpron	名词性代词	Noun-pronoun	而人主兼礼**之**。
apron	形容词性代词	Adjective-pronoun	故明主用**其**力。
za	“之”作定语后置标志	*Special for Old Chinese*	乡人**之**善者
zj	“者”作名词性词尾	*Special for Old Chinese*	乡人之善**者**
zd	“之”作“的”	*Special for Old Chinese*	古**之**人不余欺。
fy	发语词	*Special for Old Chinese*	**夫**离法者罪。
conjad	副词性连词	Adverbial conjunction	用之**则**乱法.
Period	终止性标点		。；？！
Comma	~~停顿性标点~~		，、：

1.2 Tagging Algorithms

We apply the Hidden Markov Model (HMM) (Viterbi, 1967) and the forward-backward algorithm (James, 1995) to obtain the context-dependent probabilities.

Generally there are 2 types of HMM taggers for parsers, the trigram model and the bigram forward-backward model. Charniak (1996) suggested that the former is better for parsers.

But the former only result in a *deterministic* sequence of most probable POS, in other words, it assigns only *one* POS tag for each word. Although the accuracy of trigram by our previous work (2002) is as high as 97.6%, for a sentence of 10 words long, the possibility of all-correctness is as low as low as $(97.6\%)^{10} = 78.4\%$, and the single-tag scheme does not allow parsers to *re-call* the correct tags, as is often done if we apply the forward-backward model. So in this paper we still apply the traditional bigram forward-backward algorithm. We suggest that a combination of trigram and forward-backward model would be the best choice, although no such attempt exists in the literature.

2 PCFG Model and Classical Chinese Grammar

In this section we will cover the PCFG model and context-sensitive rules designed for Classical Chinese. Features of the rule-set will be also discussed.

2.1 PCFG Model and Rule Restriction

CFG: A context-free grammar (CFG) is a quadruple (V_N, V_T, S, R) where V_T is a set of terminals (POS tags), V_N is a set of non-terminals (syntactic tags), $S \in V_N$ is the start non-terminal, and R is the finite set of rules, which are pairs from $V_N \times V^+$, where V denotes $V_N \cup V_T$. A rule $< A, \alpha >$ is written in the form $A \to \alpha$, A is called the left hand side (LHS) and α the right hand side (RHS).

PCFG: A probabilistic context-free grammar (PCFG) is a quintuple (V_N, V_T, S, R, P), where (V_N, V_T, S, R) is a CFG and $P : R \mapsto (0,1]$ is a probability function such that $\forall N \in V_N : \sum_{\alpha : N \to \alpha \in R} P(N \to \alpha) = 1$

Rule Restriction: We restrict the CFG rules to be binary or unary rules, but ***NOT*** as strict as the Chomsky Normal Form (CNF). Each $R_i \in R$ could be in the following two forms only:

1. $R_i : N_j \to AB$
2. $R_i : N_j \to A$

where $N_j \in V_N$ and $A, B \in V$

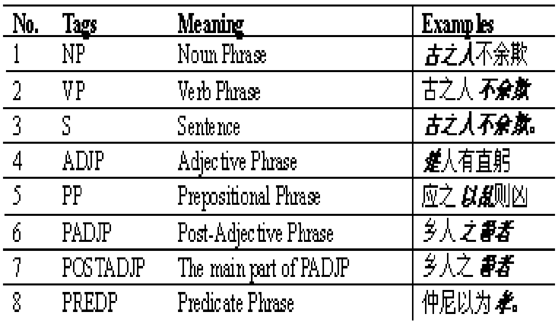
The advantage of binary/unary rules lies in the simplicity of parsing algorithm, and will be discussed in Section 4.

The major difference between our model and CNF is that for unary rules, we do not require the right-hand-side to be terminals. And this enables us easier representation of the Classical Chinese language.

2.2 Rule-Set for Classical Chinese

An important advantage of PCFG is that it needs fewer rules and parameters. According to our corpus, which is representative of Classical Chinese classics, only 100-150 rules would be sufficient. This is mainly because our rule set is linguistically sound. A summary of the set of rules is presented as follows.

Table 2. Our non-terminals (also called syntactic tagset, or constituent set)



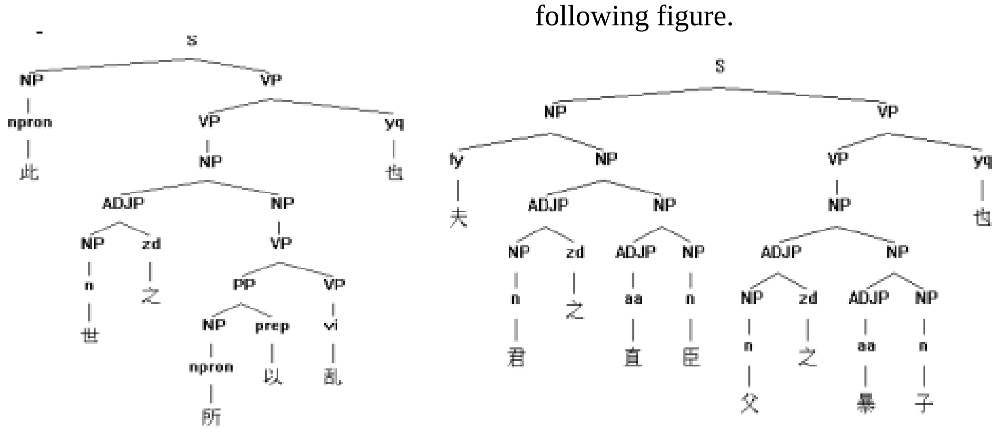
No.	Tags	Meaning	Examples
1	NP	Noun Phrase	**古之人**不余欺
2	VP	Verb Phrase	古之人 **不余欺**
3	S	Sentence	**古之人不余欺。**
4	ADJP	Adjective Phrase	**楚**人有直躬
5	PP	Prepositional Phrase	应之 **以乱**则凶
6	PADJP	Post-Adjective Phrase	乡人**之善者**
7	POSTADJP	The main part of PADJP	乡人之 **善者**
8	PREDP	Predicate Phrase	仲尼以为**孝。**

A subset of most frequently used rules is shown in the following table.

Table 3. A simple subset of PCFG Rules for Classical Chinese

1.	S ->	NP VP	; simple S/V
2.	S ->	VP	; S omitted
3.	S ->	VP NP	; S/V inversion
4.	S ->	ad S	
5.	VP ->	vi	
6.	VP ->	vt NP	; simple V/O
7.	VP ->	NP vt	; V/O inversion
8.	VP ->	ad VP	
9.	VP ->	PP VP	; prepositioned PP
10.	VP ->	VP PP	; postpositioned PP
11.	VP ->	NP	; NP as VP
12.	VP ->	VP yq	
13.	NP ->	n	
14.	NP ->	npron	
15.	NP ->	ADJP NP	
16.	NP ->	POSTADJP	
17.	NP ->	VP	; V/O as NP
18.	NP ->	fy NP	
19.	ADJP ->	aa	
20.	ADJP ->	apron	
21.	ADJP ->	NP zd	
22.	PP ->	prep NP	; P+NP
23.	PP ->	NP prep	; inversion
24.	PP ->	prepb	; object omitted
25.	PP ->	NP	; prep. omitted
26.	POSTADJP->	VP zj	

Examples of parse trees are shown in the following figure.



(a) (b)

Fig. 1. the parse trees of 2 sentences (a) 此世之所以乱也。 (b) 夫君之直臣父之暴子也。

2.3 Features of Classical Chinese Grammar Rules

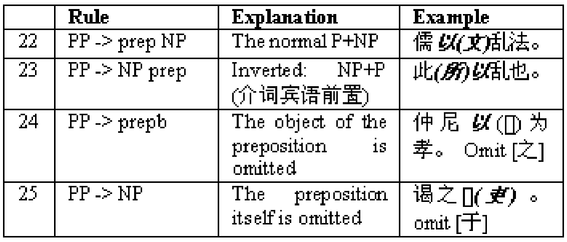
As an aside, it is worthwhile to point out here some peculiarities of the Classical Chinese grammar used in our work. Readers not interested in grammar modeling may simply skip this subsection. As mentioned before, the grammar of Classical Chinese is entirely different from that of English, so a few special features must be studied. Although these features bring many difficulties to the parser, we have developed successful programming techniques to solve them.

From the rule-set, the reader might find that two special grammatical structures is very common in Classical Chinese:

1. Inversion: subject/verb inversion (rule 3), preposition/object inversion (rule 23).
2. Omission: Subject omitted (rule 2), preposition’s object omitted (rule 24), preposition omitted (rule 25).

Maybe the strangest feature is the structure of PP. English PP is always P+NP. But here in Classical Chinese, by inversion and omission, the PP may have up to 4 forms, as shown in rule 22-25.

Table 4. The 4 rules from PP. The object of the preposition is in brackets, and [] indicate an omission.



	Rule	Explanation	Example
22	PP -> prep NP	The normal P+NP	儒 *以(文)*乱法。
23	PP -> NP prep	Inverted: NP+P (介词宾语前置)	此*(所)以*乱也。
24	PP -> prepb	The object of the preposition is omitted	仲尼 *以*(□) 为孝。 Omit [之]
25	PP -> NP	The preposition itself is omitted	谒之 □(*吏*) 。 omit [于]

Another feature that must be pointed out here is the cycle. In our rule-set, there are 2 rules (rule 11 and rule 17) forming a cycle:

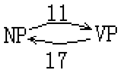


Fig. 2. A cycle in the rule-set. Rule 11: NP-> VP, Rule 17: VP-> NP.

It will ease our parsing because Classical Chinese is lexically and syntactically very ambiguous. An NP can act as a VP (a main verb), while a VP can act as a NP (subject or object). These two features are exemplified in figure 3. There are actually more cycles in the rule-set. Helpful as they are, the cycles bring great difficulty to the memory-based top-down parser. In practice, we develop a closure–based method to solve this problem, as shown in the following pseudo-code:

```
better_results_found=true;
while (better_results_found)
{
        better_results_found=false;
        memory_based_top_down_parse();
        // if better results found, the variable will be set true
}
```

Another point is the use of preferences for ambiguity resolution. While the ambiguities in our rule-set greatly ease our modeling Classical Chinese grammar, it causes the parser to make a lot of ridiculous errors. So we here apply some predefined preferences such as 'an fy must be at the first of an NP' and 'a yq must be at the end of a VP'. This consideration results in a significant increase in the parsing accuracies.

3 Evaluations

In our preliminary experiments, we constructed a treebank of 1000 manually parsed sentences (quite large for Classical Chinese treebank), in which 100 sentences are selected as the test set using the cross-validation scheme, while the others as the learning set. The majority of these sentences are extracted from classics of pre-Tsin Classical Chinese such as *Hanfeizi* and *Xunzi* because in these texts there are fewer proper nouns and difficult words. That is the restriction we put on the selection of Classical Chinese texts. It must be pointed out here that compared from other languages, Classical Chinese sentences are so short that the average length is only about 4-6 words long.

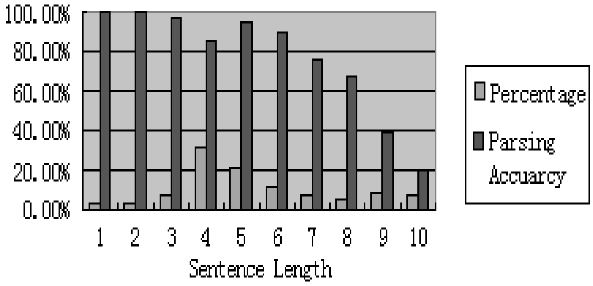


Fig. 3. Sentence Distributions and Parsing Accuracies

Figure 3 shows the distribution of sentences and parsing accuracies for different sentence lengths. For distribution, we can see that those 4-word, 5-word, and 6-word sentences constitute for the majority of the corpus, while those 1-word and 2-word sentences are very few. For accuracy, the parser is more effective for shorter sentences than for longer sentences. And for 1-word and 2-word sentences, there is no error report from the parse results.

Conclusion

Computer processing of Classical Chinese has just been commencing. While Classical Chinese is generally considered too difficult to process, our previous work on part-of-speech tagging has been largely successful because there is almost no need to segment Classical Chinese words. And we continue to use the tagset and corpus into this work. We first apply the forward-backward algorithm to obtain the context-dependent probabilities. The PCFG model is then presented where we restrict the rules into binary/unary rules, which greatly simplifies our parsing programming. According to the model, we developed a CFG rule-set of Classical Chinese. Some special features of the set are also studied. Classical Chinese processing is generally considered too difficult and thus neglected, while our works have shown that by good modelling and proper techniques, we can still get encouraging results. Although Classical Chinese is currently a dead language, our work still has applications in those areas as Classical-Modern Chinese Translation.

For future work of this paper, we expect to incorporate trigram model into the forward-backward algorithm, which will increase the tagging accuracy. And most important of all, it is obvious that the state-of-the-art PCFG model is still two-leveled, we expect to devise a three-level model, just like trigram versus bigram.

Acknowledgements

Our special thanks go to Prof. Lu Ruzhan of Shanghai Jiaotong University for his sincere guidance.

References

Allen, J. (1995) *Natural Language Understanding,* The Benjamin/Cummings Publishing Company, Inc.

Viterbi, A. (1967) *Error bounds for convolution codes and an asymptotically optimal decoding algorithm.* IEEE Trans. on Information Theory 13:260-269.

Yao Y., Lua K. (1998a) *A Probabilistic Context-Free Grammar Parser for Chinese,* Computer Processing of Oriental Languages, Vol. 11, No. 4, pp. 393-407

Huang L., Peng Y., Wang H. (2002) *Statistical Part-of-Speech Tagging for Classical Chinese,* Proceedings of the 5th International Conference on Text, Speech, and Dialog (TSD), Brno, in press

Klein, D., and Manning C. (2001) *Natural Language Grammar Induction using a Constituent-Context Model*, Proceedings of Neural Information Processing Systems, Vancouver.

Yao Y., Lua K. (1998b) *Mutual Information and Trigram Based Merging for Grammar Rule Induction and Sentence Parsing,* Computer Processing of Oriental Languages, Vol. 11, No. 4, pp. 393-407

Johnson, M. (2001) *Joint and conditional estimation of tagging and parsing models,* Proceedings of International computational linguistics conference, Toulouse

Charniak, E. (1996) *Taggers for Parsers*, Artificial Intelligence, Vol. 85, No. 1-2, pp. 45-47.