ISCA Archive
http://www.isca-speech.org/archive

International Symposium on Chinese Spoken
Language Processing (ISCSLP 2002)
Taipei, Taiwan
August 23-24, 2002

AN APPROACH TO AUTOMATIC IDENTIFICATION OF CHINESE BASE NOUN PHRASES

Yan ZHANG
National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing, 100080, China
yzhang@nlpr.ia.ac.cn

Chengqing ZONG
National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing, 100080, China
cqzong@nlpr.ia.ac.cn

Bo XU
National Laboratory of Pattern Recognition, Institute of Automation
Chinese Academy of Sciences, Beijing, 100080, China
xubo@nlpr.ia.ac.cn

ABSTRACT

This paper presents an approach to identify Chinese base noun phrases. This method is based on GLR algorithm and extends GLR parsing algorithm further. It is a mixed approach that combines rule-based method and statistical method by using PCFG system. From the experiment results, this method is not only simple but also feasible and efficient to base noun phrases identification.

1. INTRODUCTION

The identification of noun phrases plays an important role in many natural language processing applications, such as partial parsing, information retrieval and machine translation. From the cognitive point of view, human being must recognize, learn and understand the entities or concepts in the texts for natural language processing. These entities or concepts are usually described by noun phrases.

Usually, to identify the noun phrases in the texts means to parse the text and to resolve the relations among the constitutes. The identification of noun phrases belongs to shallow or partial parsing. A base noun phrase (BNP) is a simple noun phrase and a non-recursive noun phrase that does not contain other noun phrases recursively, for example, the element bracketed in the sentence " (**地方**/NN **经济**/NN **实力**/NN) **明显**/AD **增强**/VV" is a BNP, where NN, NR, AD, VV etc are part-of-speech tags defined in Chinese tree bank[1].

Many significant work has been done to deal with the problem of BNP identification in English and many different methods have been applied. Church [2] proposes a simple noun phrase extractor. It brackets the noun phrases of input tagged texts according to two probability matrices: one is starting noun phrase matrix; the other is ending matrix. Bourigault [3] presents a tool which triggers two processing: analysis and parsing. It extracts the maximal length noun phrase. Voutilainen 's NPTool [4] applies a lexicon combined with a constraint grammar and two finite state mechanisms. Abney [5] uses two level grammar rules to implement the parser through pure LR parsing algorithm. The first level grammar rule takes care of the chunk process. The second level grammar rule deal with the attachment problems among chunks.

However, above identification methods all belong to singular methods. Some are statistical methods; the others are rule-based methods. In order to utilize linguistics information and statistical distribution in the identification method, our method uses probabilistic context free grammar as the grammar system. And the algorithm is based on GLR parsing algorithm and is extended further.

There are four sections in this paper. Section 2 briefly describes the grammar system of our parsing system. Section 3 gives the extended GLR parsing algorithm for BNP identification in detail. Section 4 presents the experiment results. The conclusion is drawn in section 5.

2. CHINESE GRAMMAR SYSTEM

2.1 The Structure of rules

Firstly this paper introduces the structures of grammar rules learned from the Penn Chinese tree-bank [1]. The definitions of symbol codes in our system are similar to the classifications of the Penn Chinese tree-bank, but the number of tags is less than that of tree-bank. Our classifications of phrase structures are mainly oriented to noun phrases. There are totally 33 part-of-speech tags, 16 structure tags in the tag set. Each rule is attached with the probability and frequency distributions. The POS tags belong to terminal symbols, while others belong to non-terminal symbols. We don't give the detailed definitions and classifications of the symbols.

In the final rule set there are 400 rules learned from Chinese tree-bank corpus for Chinese base noun phrases identification. The structure of rules is displayed in table 1.

num	rule	probability	frequency
1	NP → NR + NN	0.902135	507
2	NP → NR + NN + NN	0.094306	53
3	NP →NN + PN	0.003559	2

Table 1: the format of grammar rules

In order to denote each rule clearly, the mark '+' is used as the junction mark. Above three rules have the same left side 'NP', so the sum of probabilities of these rules equals to 1. Symbols NR, NN and PN stand for *proper noun*, *common noun* and *pronoun* respectively. In fact, we also look on QP (*quantifier phrase*) as one kind of noun phrases.

2.2 The Grammar System

Probabilistic context-free grammars (PCFG) has some merits to natural language processing. It generally considers linguistics property and the statistical distributions of grammar rules. PCFG is the simplest and most natural probabilistic model for tree structures and provide a general computational tool to simulate various probabilistic conditioning. We keep probability and frequency distributions of grammar rules simultaneously. The probability number is the relative value since it is the percentage value in the rule group that have the same left sides. While the frequency number is the absolute value because it is the total numbers occurred in whole corpus. In a word, the probability attribute is the key statistical distribution. And the probability attribute is superior to frequency attribute.

A sample is presented to show how to use the rule.
Suppose there are three rules showed in table 2 and the relations is displayed in figure 1.

Rule	F(r)	P(r)
A+C → X	f1	p1=f1/(f1+f2)
A+C → Y	f2 >f1	p2=f2/(f1+f2)
A+B+C → X	f3 <f1	p3 =1>p1

Table 2: the examples of rule

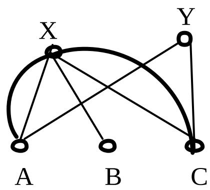

Figure 1: structure of rules

Suppose the current input symbols contain A, B and C. When rule 1 and rule 2 can be matched simultaneously, rule 2 is executed and the item symbol 'Y' is pushed to the stack because p2 is bigger than p1. Probability always has the priority to frequency.

3. ALGORITHM OF BNP IDENTIFICATION

Our parsing algorithm is based on Tomita's GLR parsing algorithm and simplifies GLR algorithm further. We will give the detailed explanation in the following part.

3.1 the Principle of Extended GLR Algorithm

In parsing processing, GLR algorithm [6,7] has four actions: shift, reduce, accept and error in GLR algorithm that are similar to the LR parsing. But GLR parsing algorithm admits multiple entries in the parsing table [8,9]. There are mainly two types of conflicts: shift-reduce conflict and reduce-reduce conflict. Our parsing algorithm also permits that several shift and reduce actions exist in one branch in the parsing table simultaneously. Since we construct the parsing table with PCFG, the search processing is more efficient and faster by statistical distribution than GLR parsing, In the parsing process, when shift action and reduce action conflict, the principle of our parsing method is that reduce action is firstly executed. When different reduce actions conflict, they are executed according to the probabilities of the reduce rules. The parser performs a best-first search through the tree of legal computations. The conflicts among the reduce actions are resolved by the probabilities of the reduced rules. More bigger the probability of a rule, more earlier the rule is executed. The best-first search guarantees that the first solution found is a best solution.

There is another problem with deciding where a noun phrase is stopped. Since the end of a noun phrase is generally not the end of the original sentence, and the ends of noun phrases are certainly not marked. Therefore a noun phrase needs a end marker. We make following work: every word is added an end symbol '#' besides the original part-of-speech tagger. That is to say, every input word bas two markers: part-of-speech tagger and end markers '#' simultaneously. The end of the original input sentence has the same two symbol '#'. If a noun phrase has been recognized, the parser will continue from two different orientations according to the POS tagger and end marker. If the current word has the two markers '#', it means that this word is the end of the input sentence, the parser stops here and output the recognized result. And if one of the markers of the current word is not symbol '#', the parser will continue to perform shift action and search a new noun phrase again.

Although it is not possible a best and simplest method to add an end marker to every word, it is a straightforward way to noun phrases identification. This can avoid showing ending error when the end of the noun phrase is not the end of the input sentence. After the parser performs the reduction operation it will continue to perform shift action or end operation. If it is not possible to get a parser for the entire sentence, the original input sentence is returned without any change. Since sentences often contain structures that are not described by grammar, and since more information is possible to be used, this operation is very important.

In general, one of the actions is error action in the parsing table. This is due to not having enough grammar structures in grammar system. Since the parser only need to give the phrase structures, the BNP identification belongs to shallow parsing. It will give the identified noun phrases in any conditions. When the error action occurs, the shift action is executed continuously. That is to say, we continue to parse the remainder sentence after the word happening error till the entire sentence is completed parsed. If the BNP is identified, the noun phrases are output and bracketed. If the sentence has no NP, then output original sentence without any change.

3.2 Process of BNP Identification

If only grammar rules are used to describe the context relations, they may easily produce more conflicts when several rules satisfy the conditions. So the statistical distributions are associated with the rules in the parsing process and decide which step to continue. We give a flow chart to explain the process of noun phrases identification in detail showed in figure 2.

3.3 example of the algorithm

We give an example to explain the process of our identification algorithm. Each rule has the form shown in section 2 in rule set:

1. NP → NN 0.397626 402
2. NP → NR + NN 0.501483 507
3. NP → NP + NN 0.052423 53
4. NP → JJ + NN 0.049456 50
5. S → NP + VV + NP 1.000000 9

'S' is the start symbol of the sentence. The Chinese input string is "*西藏 金融 工作 取得 显著 成绩*' . Each Chinese word has a determinate part-of-speech in the sentence. And the relevant POS string is "*NR NN NN VV JJ NN*".

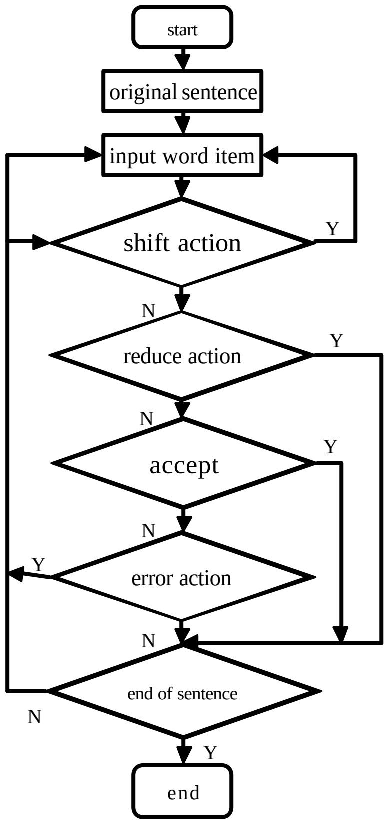

figure 2: process of BNP identification

We use an array to record the identification process.

0 [*NR]	5 [*VV]
1 [*NN]	6 [*JJ]
2 [NP (0,1)]	7 [*NN]
3 [*NN]	8 [NP (6,7)]
4 [NP (2,3)]	9 [S (4,5,8)]

In above array, symbol '*' stands for the POS of every input word. '0 [*NR]' is element 0 of the array and means that the action is shift action and the current symbol of the top of the stack is symbol 'NR'. Element '2 [NP (0,1)]' means that the action is reduce action by element 0 (NR) and element 1 (NN) and symbol 'NP' is return. We give the final structure chart to explain the parsing process in figure 3.

According to above array, we can identify the corresponding noun phrases. Although this sentence is completely identified, we only give the base noun phrase structures. In sentence *(西藏/NR 金融/NN 工作/NN) 取得/VV (显著/JJ 成绩/NN)*, the noun phrases are bracketed.

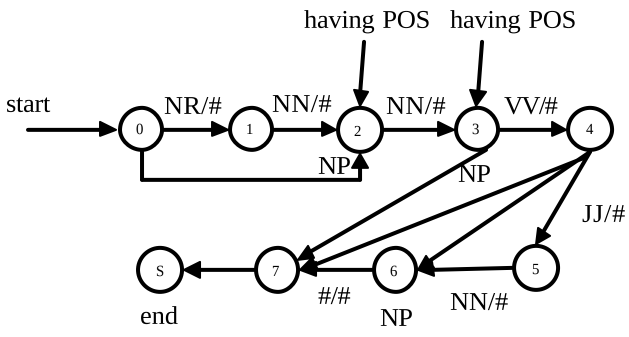

figure 3: the parsing chart

4. EXPERIMENT RESULTS

In our system, base noun phrases are defined to include not only pure noun phrase (NP) but also quantifier phrase (QP), such as QP (**一亿多**/CD **元**/M) .

To each Chinese sentence, BNP identification always gives the final identification results and the base noun phrases are bracketed. We select 1000 Chinese sentences as the test corpus. Two examples Chinese sentences are displayed with the identification results.

1. (**外商**/NN **投资**/NN **企业**/NN) **不断**/AD **增加**/VV
2. (**中国**/NR **建筑业**/NN) **对**/P **外**/NN **开放**/VV **呈现**/VV (**新**/JJ **格局**/NN)

We give the experiment results shown in table 3.

Num of experiment	Characteristics	Precision
1	Length of Sent. = 15	87.40%
2	Length of Sent. >15	81.11%

Table3: the experiment results

Our system is different from GLR parsing algorithm in following aspects. (1) We use PCFG to construct parsing table and use statistical distribution to resolve conflicts. This extends GLR parsing algorithm. (2) Since we only consider phrase structure rather than the whole relation of the sentence. When error action occurs, we still parsing the sentence to its end. This operation is a simplified work. (3) We also simplify the parsing process that don't trace back.

From the result, the extended GLR parsing algorithm is efficient for Chinese base noun phrases identification. The experiment results prove that it is more useful to add more abundant linguistics knowledge and statistical attribution to the parsing algorithm. We also find that the precision of base noun phrases is closely associated with the length of the input sentences.

5. CONCLUSIONS

This paper presents an approach to identify the base noun phrases. This approach is based on GLR parsing algorithm. Since BNP identification belongs to shallow parsing, we simplify GLR method in the parsing processing. And we utilizes PCFG grammar system that considers statistical attribution as well as the grammar attributions simultaneously. So our parsing system combines Chinese grammar phenomena with statistics distribution. The experiment results prove that this approach is a straightforward and simpler approach to complete Chinese BNP identification. In the future task, we further improve the efficiency and robustness of our parsing algorithm and expand Chinese grammatical rules with both statistical attributions and language information.

6. ACKNOWLEDGEMENTS

The research work described in this paper is supported by the National Nature Science Foundation of China under grant number 9835003 and the National Science Foundation of China under grand number 60175012, and also the National Key Foundational Research Program of China under grand number G1998030504.

7. REFERENCES

[1] Xia F., the Segmentation Guidelines for the Penn Chinese Tree bank (3.0), 2000

[2] Church, K. A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Text, *In Proceedings of Second Conference on Applied Natural Language Processing*, pp.136-143, 1988

[3] Bourigault D., Surface Grammatical Analysis for the Extraction of Terminological Noun Phrases, *In Proceedings of the 30th Annual Meeting of the Association of Computational Linguistics*, pp.977-981, 1992

[4] Voutilainen A., NPTool, a Detector of English Noun Phrases, *In Proceedings of the Workshop on Very Large Corpora*, pp.48-57, 1993

[5] Abney S., Parsing by Chunks, Kluwer Academic Publishers, 1991

[6] Tomita M., an Efficient Augmented-Context -Free Algorithm, Comp utational Linguistics, Volume 13, Numbers 1-2, 1987

[7] Feng Z. W. Automatic syntactic parsing based on Phrase Structure Grammar, Modern Linguistics, Vol.2, 2000, pp.84-98, Beijing

[8] Weng F. L., Wang Y. Y., Introduction to Computational Linguistics, China Social Sciences Press, 1998

[9] Chen K.H. and Chen H.H., Extracting Noun Phrases from Large-Scale Texts: A Hybrid Approach and Its Automatic Evaluation, *In Proceedings of the 32nd Annual Meeting of Association of Computational Linguistics*, New York, pp.234-241, 1994