ARE STATISTICS-BASED APPROACHES GOOD ENOUGH FOR NLP?

A CASE STUDY OF MAXIMAL-LENGTH NP EXTRACTION IN MANDARIN CHINESE

Wenjie Li, Haihua Pan*, Ming Zhou†, Kam-Fai Wong and Vincent Lum
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong
Shatin, N.T., Hong Kong
E-mail: {wjli,hpan}@se.cuhk.hk

Abstract

Statistics-based approaches became very popular in recent NLP researches, because of their apparent advantages over linguistics or rule-based approaches. Some even claimed that it would not be necessary to employ the latter approach at all. Thus, it seemed necessary to evaluate such claim and the applicability of the former to NLP in general.

Because of the usefulness of noun phrases (NPs) in many applications, in this paper, we present a simple statistics-based partial parser to detect the boundaries of maximal-length NPs in part-of-speech tagged Chinese texts. On the basis of our experimental results, we will show that statistics-based approaches with purely part-of-speech tags are not adequate for NP extraction in Chinese; they fail to handle cases with structural ambiguity. Our experiments suggest that syntactic and semantic checking is necessary to correctly mark the boundary of maximal-length NPs in Chinese. We conclude with possible solutions to the problematic cases for statistics-based approaches.

1 Introduction

Noun phrases are the basic building blocks of sentences in natural language. They are the basic means for representing concepts in human cognition. They are also

*Department of Chinese, Translation and Linguistics, City University of Hong Kong, 83 Tat Chee Avenue, Kowloon, Hong Kong

†Department of Computer Science, Tsinghua University, Beijing, PRC.

the more appropriate translation units in language translation than words or part-of-speech classes, as argued in Van der Eijk [7]. Furthermore, noun phrases abound in our daily documents and conversation. Thus, extracting NPs from running texts is very useful for many applications such as verb frame characterization, document indexing, information retrieval, sentence parsing, machine translation, etc.

Traditionally, to obtain a noun phrase in a text means to parse the whole sentence first, and then extract the partial tree with NP labels. However, this whole-partial method is quite difficult and involves a great deal of complexity, since various ambiguities cannot be resolved by syntactic or even semantic information. So recently, phrase-oriented partial parser or phrase extractor is gradually explored in noun phrase extraction and preposition phrase attachment (Church [4], Rausch, Norrback and Svensson [11], Bourigault [2], Voutilainen [15], Chen and Chen [3], etc.). The majority of the literature on NP extraction prefer statistics-based approaches over rule-based approaches to avoid detailed and tedious linguistic engineering. Although there are several studies on extracting NPs in English and non-Asian languages using stochastic methods, studies on extracting Chinese NPs have not been reported thus far.

In this paper, a probabilistic partial parser is proposed to extract maximal-length noun phrases in Chinese, which will be used in an information retrieval system. Our research aims at examining the applicability of stochastic methods in parsing Chinese. On the basis of our experimental results, we argue that merely statistics-based approaches with part-of-speech tags are not adequate for maximal-length noun phrase extraction in Chinese, and it is necessary to employ syntactic and semantic information and some kind of rule-based techniques in detecting the boundary of NPs.

2 Previous Work

Church [4] proposes a part-of-speech tagger and a simple non-recursive noun phrase extractor. His noun phrase extractor brackets the "minimal-length noun phrase" (non-recursive) in part-of-speech tagged texts according to two probability matrices: starting NP matrix and ending NP matrix; the same methodology has been used by Garsde and Leech [9] in their probability parser. By calculating the probabilities of inserting an open or close bracket between all pairs of parts of speech, Church achieves a recall rate of 98%, i.e., only 5 out of 248 noun phrases are missed. Although the recall rate is pretty high, the test corpus is too small, and only minimal-length non-recursive NPs are tested.

Rausch, Norrback and Svensson [11] design a *nuclear noun phrase* extractor which takes part-of-speech tagged Swedish texts as input and inserts brackets around noun phrases, i.e., sequences of determiners, premodifiers and nominal phrase. Their system can identify 85.9% of all nuclear noun phrases in a 6,000 word long text collection with a precision of 84.3%.

Bourigault [2] reports a tool, *LECTER*, for extracting terminologies from French texts, and it can extract *maximal-length noun phrases.* His system can recognize 95% of all noun phrases, that is, 43500 out of 46000, from the test corpus. However, no figures are given on how much 'garbage' the system suggests as noun phrases.

Voutilainen [15] also announces an *NPtool* to acquire maximal-length noun phrases. It uses a lexicon with part-of-speech tags and head information, and two rule bases (one is NP-hostile; the other is NP-friendly) for the task. The two mechanisms produce two NP sets and the intersection set of them will be labeled as the final NP. The recall is 98.5-100% and the precision is 95-98% in different domains, which is validated manually by some 20000 words. But as pointed in Chen and Chen ([3]), the recall is only about 85% according to the sample text listed in his appendix.

Chen and Chen [3] design a new and more sophisticated mechanism by combining the statistical method and rule-based method for extracting simple English NPs based on the SUSANNE corpus [13]. They use a probabilistic partial parser with dynamic programming to find out the best liner chunk sequence for the tagged input sentences, and then assign a syntactic head and a semantic head to each chunk with the help of linguistic knowledge. Then the plausible maximal noun phrases are extracted and connected according to the information of syntactic head, semantic head and a finite state mechanism with only 9 states. The average precision is 95%. Due to the difficulty of distinguishing different NP types such as *maximal-length* NPs, *minimal-length* NPs, etc., the average recall is hard to measure, and Chen and Chen only give a suggestive recall of 96% for simple NPs which contain no prepositional phrases (except for the *of*-phrase) or relative clauses. That is, the recall will be much lower if all types of NPs are considered.

Since all the researches discussed so far deal with English and non-Asian languages and seem to suggest that statistics-based approaches are adequate for extracting NPs (except for Voutilainen [15], which employs rule-based methods), it is necessary to examine the applicability of statistics-based approaches to languages like Chinese. In the following, after briefly discussing the complexity of noun phrases, we

will present our experiments on extracting maximal-length NPs in Chinese using a statistics-based partial parser, and discuss the results and their implications to the viability of statistics-based approaches to natural language processing.

3 The Complexity of NPs

Noun phrases in English are usually composed of a determiner, an adjective, and a noun, though the first two elements are optional. They can also be modified by prepositional phrases (PPs) and relative clauses. When they are modified by PPs, a PP-attachment ambiguity may arise, as exemplified in (1), the PP in which can modify either the NP or the verb.

(1) John [$_V$ saw] [$_{NP}$ the girl] [$_{PP}$ with a telescope].

The PP-attachment problem is a very complex issue, and requires utilizing lexical, syntactic, semantic, and pragmatic information, so statistics-based approaches do not necessarily have advantages over rule-based approaches in dealing with such problem. NPs with relative clauses also increase the complexity of noun phrases. Because of the word order in English, it is not easy for a statistics-based parser to mark the boundary of relative clauses and the maximal-length NPs. There is also the structural ambiguity induced by the so-called garden-path sentences, as exemplified below:

(2) John told [$_{NP}$ the boy] [$_{RC}$ the dog bit] [$_S$ Sue liked him].

Simply using statistical information cannot rule out the possibility that *Sue* is analyzed as the object of the verb *bit*, which is the reason for a human parser to backtrack when the verb *liked* is encountered.

Since previous studies on English NP extraction employing statistical methods did not cover NPs with PP or relative clause modification (though Chen and Chen include *of*-phrases in the extracted NPs), they cannot provide solid evidence for the claim that statistical methods are superior to rule-based methods.

In Chinese all the modifiers of NPs precede the head noun. The PP-attachment problem and garden-path sentences induced by relative clauses are avoided in the language. Thus, Chinese presents itself as a testing case for us to examine whether the statistics-based approach can simplify the parsing problem and avoid the complexity of the whole-part method mentioned in the introduction section.

4 Extracting Maximal-length Chinese NPs

4.1 The Corpus

In our current experiments, we use a news report corpus of 30 files which contain 16660 words, 3278 NPs, and 750 sentences. On the average, there are 22 words in a sentence (not including punctuations). All the files have been tagged by TAGGER, a part-of-speech tagging system, developed by Tsinghua University, Beijing, China [1]. The tag set, designed by Beijing University, China [16], contains 24 general categories and 110 part-of-speech tags. The following shows a snapshot of the tagged corpus with marked NP boundaries, where symbol '$' marks the beginning of a sentence, and the English characters after symbol '#' indicate the part of speech of the Chinese word before '#'.[1]

...

$ [他 #rn] 对 #p 着 #utz [报话机 #ng] 拚命 #d 地 #usdi 喊 #vgo 着 #utz ， [大本营 #ng] 一时 #d 寂静 #a ， [整个 #b 绒布 #s 河谷 #ng] 回荡 #vgn 着 #utz [罗则 #npf 颤抖 #vg 的 #usde 声音 #ng] 。

$ [这 #rn 位 #qni 5 2 #mx 岁 #ng 矮 #a 壮 #a 汉子 #ng 的 #usde 眼 #ng] 里 #f 闪 #vgn 着 #utz [泪花 #ng] 。

...

4.2 Method

Our experiment consists of two parts: training and testing. Of the 30 files, we use 25 of them for training and close test, and the rest 5 for open test. First, we manually marked all the maximal-length NPs in the 25 files using "[" for left boundary and "]" for right boundary. We found conjoined NPs and many NPs with PP and/or relative clause modification in our corpus.

Second, we trained our NP extraction program (NPext) using the 25 files with all the maximal-length NPs marked to acquire statistical information about the probability of any two categories for marking left and right boundaries. Thereafter, NPext marked the maximal-length NPs in the 25 files without the boundary markers. Since NPext marked the left and right boundaries independently, we need to pair them, and several pairing methods were examined. Finally, we conducted the open test on the rest 5 files.

[1]The description of the part-of-speech tags is given in the appendix.

Table 1: Probability of Starting an NP

	a	ng	p	vgn
a	0	0.017	0	0
ng	0.031	0.021	0	0
p	0.650	0.728	0.833	0.139
vgn	0.804	0.723	0.333	0.438

4.3 Training

Following Church [4], we acquired two matrices which contain statistical information about the probability of having a left or right NP boundary between any two part-of-speech tags. Suppose that w_i and w_{i+1} are two adjacent words in the sentence, t_i and t_{i+1} are their part of speech, respectively, and NP_B and NP_E are the left and right boundaries. Then the probabilities are defined as below:

$$\begin{aligned} P(NP_B|t_i, t_{i+1}) &= \textit{probability of a left boundary} \\ &= \frac{freq(t_i, NP_B, t_{i+1})}{freq(t_i, t_{i+1})} \\ P(NP_E|t_i, t_{i+1}) &= \textit{probability of a right boundary} \\ &= \frac{freq(t_i, NP_E, t_{i+1})}{freq(t_i, t_{i+1})} \end{aligned}$$

A sample is shown in Tables 1 and 2 for the four common part-of-speech categories in the corpus: a (adjective), ng (general noun), p (preposition) and vgn (verb with an NP object). The first row is the t_{i+1}; the first column is the t_i; and the other entries are probabilities.

From Tables 1 and 2, we can see that "p" and"vgn" are most likely to start an NP, and "ng" to end an NP. Note that in Table 1 there are values larger than zero in the pairs "p" and "p", "vgn" and "p", and "vgn" and "vgn" in Chinese; this is different from English, as shown in Church [4]. The reason is that, unlike English, all the modifiers precede the head noun in Chinese. As a result, Chinese has NPs with the word order "PP N" or "Relative-Cl N", where "Relative-Cl" can start an NP with a verb of the category "vgn".

4.4 Testing

Using the knowledge acquired from the training phase, we conducted close tests on the 25 files used for training, and open tests on the 5 remaining files. In both tests, NPext

Table 2: Probability of Ending an NP

	a	ng	p	vgn
a	0	0	0	0
ng	0.57	0.028	0.744	0.837
p	0	0	0	0
vgn	0	0	0	0

Table 3: Results for Candidate Boundaries

		Close Test	Open Test
Correct No.	left	2717	494
	right	2722	523
Wrong No.	left	4040	770
	right	2882	510
NP No.		2723	555
Recall %	left	99.7	89.1
	right	99.9	94.2
Precision %	left	40.2	39.0
	right	48.6	50.6

first found the candidate boundaries of all NPs by marking left and right boundaries independently, and subsequently, it obtained the final NPs through pairing the left and right boundaries.

4.4.1 Finding Candidate Boundaries

When the probability is larger than a threshold, an appropriate boundary marker is inserted. For instance, for the word pair 在 #p 'at' and 学校 #ng 'school', if the threshold is set to 0.4, "[" will be inserted between 在 and 学校, but not "]", since $P(NP_B|p,ng)$ is 0.728 which is larger than the threshold 0.4, and $P(NP_E|p,ng)$ is zero, which is less than the threshold. Table 3 shows the results for candidate boundaries when the threshold is set to zero.

4.4.2 Pairing Left/Right Boundaries

Since the statistical method only depends on statistical information, the marked left/right boundary can be incorrect. Furthermore, there may be more left bound-

Table 4: Recall After Pairing

Combinations		Forward %		Backward %	
left	right	close	open	close	open
ML	ML	79.7	67.7	79.1	67.8
* MP	MP	81.9	69.4	81.8	69.1
ML	MP	79.6	67.6	79.8	67.8
MP	ML	80.7	69.1	80.6	68.7

aries marked than right boundaries, or the other way round. In order to get correct maximal-length NPs, two methods, *maximal length* (ML) and *maximal probability* (MP), were employed to pair the candidate left/right boundaries. The maximal probability method chooses the candidate boundary with the highest probability, while the maximal length method selects the left and right pair with the maximal length. For example, suppose that we have three left boundaries and two right boundaries marked for a candidate NP, then we will choose the outmost boundaries as the left and right boundaries, if we apply the maximal length method to both left and right boundaries. But, if we use the maximal probability, then the boundaries with the highest probability will be chosen as the left and right boundaries, respectively.

By varying the direction of pairing: forward and backward, and using different combinations of the two methods: ML and MP, we had eight ways of pairing the left and right boundaries. Tables 4 and 5, respectively, show the recall and precision of the final maximal-length NPs, where the candidate boundary set was acquired with a threshold of 0.1.[2]

The comparison of the eight pairing strategies leads to the following conclusions:

- There does not exist much difference between the two directions of pairing: forward and backward, which means that, for Chinese, the characteristic of starting an NP and ending an NP is almost the same.

[2]Note that our precision and recall were calculated based on the definitions given in Chen and Chen, [3] repeated below, where "a" represents the number of NPs marked by both NPext and the human evaluator, "b" the number of NPs marked by NPext only, and "c" the number of NPs marked by the human evaluator only.

$$Precision = a/(a+b) * 100\% \quad (1)$$

$$Recall = a/(a+c) * 100\% \quad (2)$$

Table 5: Precision After Pairing

Combinations		Forward %		Backward %	
left	right	close	open	close	open
ML	ML	77.1	68.9	77.3	70.3
MP	MP	78.0	67.3	77.3	69.7
ML	MP	76.9	68.8	77.2	70.3
* MP	ML	78.1	70.6	78.7	71.3

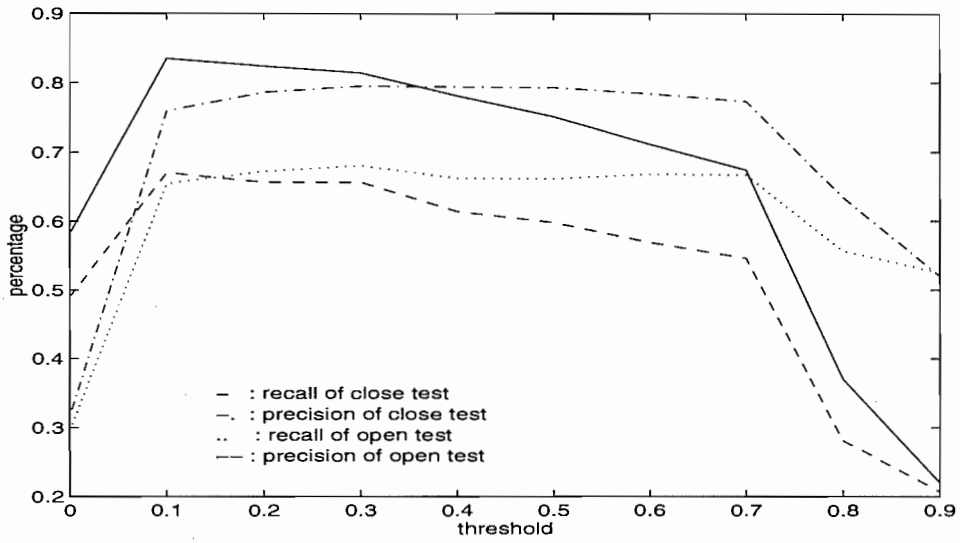

Figure 1: Recall and Precision With Different Thresholds

- Table 4 suggests that the combination of maximal probability for both left and right boundaries with the *forward* direction leads to the best recall, while Table 5 indicates that the combination of maximal probability for left boundary and maximal length for right boundary with the *backward* direction gives us the best precision; both are marked by '*' in the tables.

We also calculated the precision and recall of the close and open tests with thresholds varying from 0 to 1 for obtaining the candidate left and right boundaries. Figure 1 shows us the experimental results after pairing. We can see from Figure 1 that the threshold of 0.1 gives us the best precision and recall for both close and open tests.

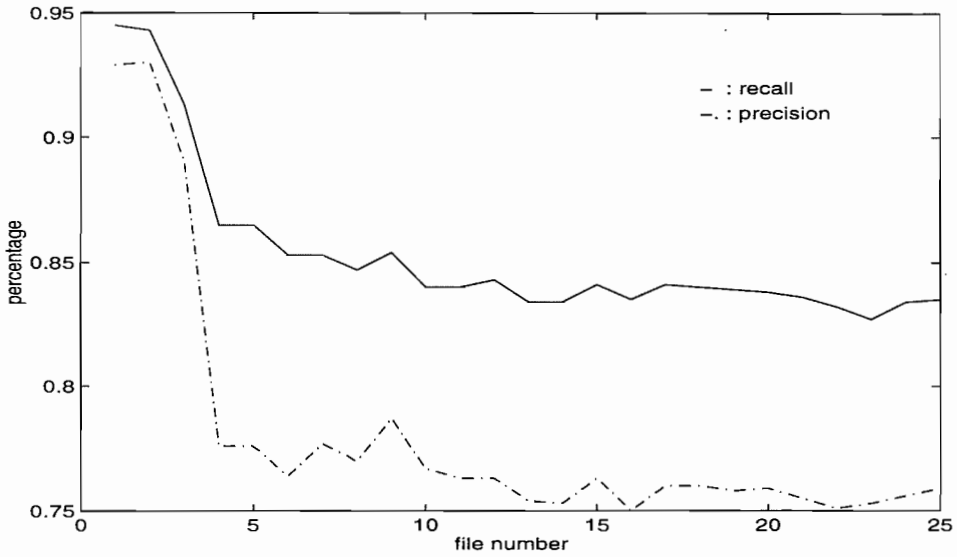

Figure 2: Precision and Recall for Close Test

4.5 Experiment Evaluation

This is the first attempt to find maximal-length Chinese noun phrases using statistical methods based on boundary probabilities. During the research, we found that the best recall and precision for close test are 81.9% and 78.7%, and the best ones for open test, 69.4% and 71.3%, as shown in Tables 4 and 5. Although the corpus size is relatively small, our results are reliable. This is justified by further experiments with varying training set. Results of these experiments show that the recall and precision for both close and open tests stablized within the 25 training files. The recall and precision for close tests emerged when the number of training files reached around 12, as shown in Figure 2. Similarly, those for open test stablized at around 22 training files, as shown in Figure 3.

Table 6 lists the distribution of the errors made by our NPext program. The error types are explained below with examples except for the "others" category; the errors in this type were mainly caused by wrong tags marked in the corpus and wrongly marked boundaries for training.[3]

A: The correct analysis should be two consecutive NPs, i.e., NP1 and NP2, but NPext combined them into one. Typical cases are the double subject and object

[3]We did not give the English translation for our examples, since it is not necessary to understand the content of the sentences for making our point; simply checking the tags and subscripts of the brackets is enough to verify our claims.

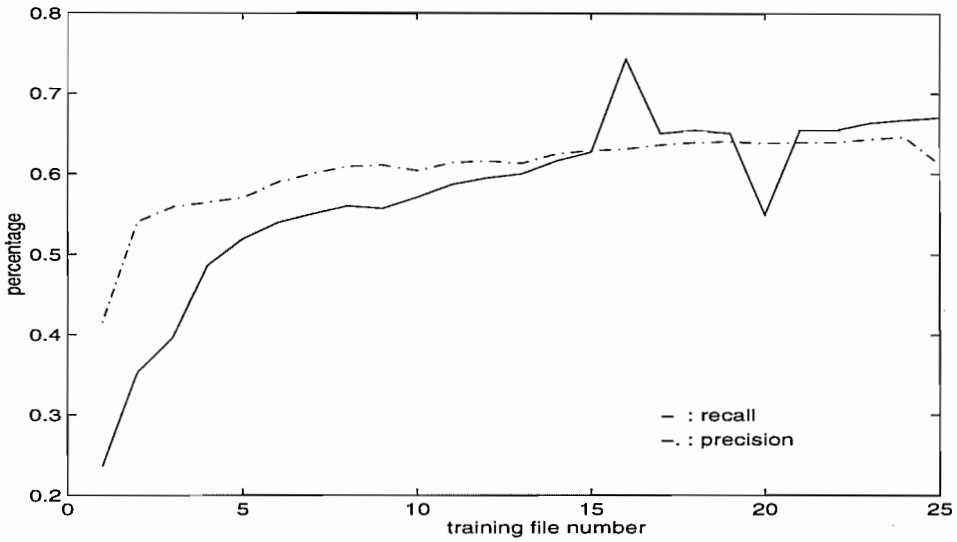

Figure 3: Precision and Recall for Open Test

Table 6: Error Distribution

Error types	NP missed		NP marked wrong	
	No.	%	No.	%
A	60	13.3	30	4.16
B	91	20.1	186	25.76
C	87	19.3	205	28.4
D	10	2.2	10	1.39
E	9	2.0	9	1.38
Others	195	43.1	282	39.1
Total	358	100	722	100

constructions in Chinese. Below are two examples with the correct analysis followed by the wrongly marked one.

Correct: [$_{NP1}$ 原材料 #j 价格 #ng] [$_{NP2}$ 上涨 #vg 幅度 #ng] 大 #a.
Wrong: [$_{NP}$ 原材料 #j 价格 #ng 上涨 #vg 幅度 #ng] 大 #a.
Correct: ... 给 #vgn 了 #utl [$_{NP1}$ 我们 #rn] [$_{NP2}$ 做人 #vg 的 #usde 权力 #ng]
Wrong: ... 给 #vgn 了 #utl [$_{NP}$ 我们 #rn 做人 #vg 的 #usde 权力 #ng]

The two NPs: NP1 and NP2 in the correct sentences above were wrongly merged into one NP, as indicated by the bracketed NPs in the wrong sentences.

B: This type is the opposite of type A, i.e., the correct analysis is one NP, but it was marked as two NPs: NP1 and NP2.

Correct: ... 是 #vy [$_{NP}$ 治理 #vg 整顿 #vg 的 #usde 关键 #ng 一 #mx 年 #ng].
Wrong: ... 是 #vy [$_{NP1}$ 治理 #vg 整顿 #vg 的 #usde 关键 #ng] [$_{NP2}$ 一 #mx 年 #ng].

Our program NPext incorrectly split the one NP in the correct sentence above into two NPs: NP1 and NP2.

C: The correct analysis is an NP containing a relative clause (RC), but NPext split it into a verb and an NP, so the result is a verb phrase (VP), not an NP.

Correct: ... 在 #pzai [$_{NP}$[$_{RC}$ 制造业 #ng 中 #f 居 #vgn 主导 #ng 地位 #ng] 的 #usde 汽车 #ng 工业 #ng] 尤 #d 为 #vi 明显 #a 。
Wrong: ... 在 #pzai 制造业 #ng 中 #f [$_{VP}$[$_{V}$ 居 #vgn] [$_{NP}$ 主导 #ng 地位 #ng 的 #usde 汽车 #ng 工业 #ng]] 尤 #d 为 #vi 明显 #a 。

D: This type involves compounds and sequences like "vgn" "vgn" "ng". The correct boundary should be between "vgn" and "vgn" for the sentence below.

Correct: ... 实施 #vgn [贴 #vgn 花 #ng 分割 #vg 等 #x 合理化 #ng 建议 #nvg]
Wrong: ... 实施 #vgn 贴 #vgn [花 #ng 分割 #vg 等 #x 合理化 #ng 建议 #nvg]

E: This type involves sequences like "p" "vgn" "ng", and the correct boundary should be between "p" and "vgn", but NPext wrongly marked it between "vgn" and "ng". That is, the result should be P + NP, but it was marked as P + VP.

Correct: ... 从 #p [$_{NP}$[$_{RC}$ 稳定 #vgn 社会 #ng] 的 #usde 大局 #ng] 出发 #vgo.
Wrong: ... 从 #p [$_{VP}$[$_{V}$ 稳定 #vgn] [$_{NP}$ 社会 #ng 的 #usde 大局 #ng]] 出发 #vgo.

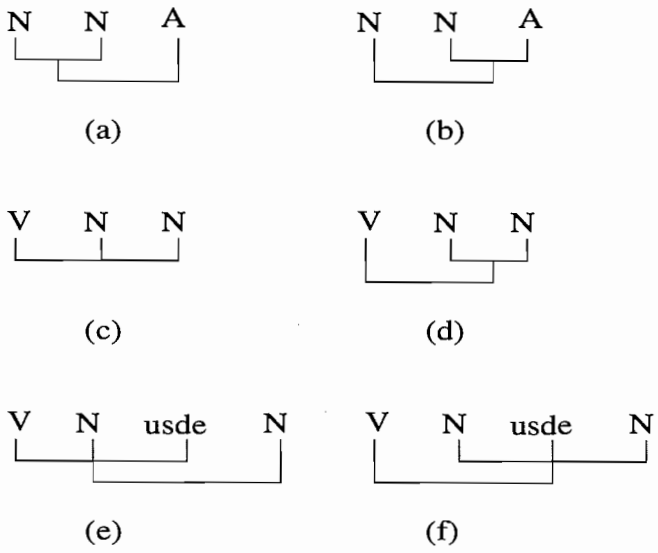

Figure 4: Structural Ambiguity

4.6 Discussion

By carefully examining the errors made by NPext, we see that most of them are structurally ambiguous. Figure 4 shows the possible structures for the sequences in (3) below, which are the possible sources for the error types mentioned earlier; note that (3a) and (3b) cover the error types A and B, and (3c) corresponds to error type C.

(3) a. N N A b. V N N c. V N usde N

Pattern (3a) has two possible structures (a) and (b) in Figure 4, (3b), (c) and (d), and (3c), (e) and (f). Note that (c) in Figure 4 is the double object configuration, but (d) is the single object configuration. The tag *usde* in (e) and (f) is a relative clause marker, or in general a modifier marker in Chinese. Structure (e) gives us an NP with a relative clause, but (f), a VP. Thus, we see that the sequences in (3) are ambiguous, but statistics-based approaches cannot differentiate them.

One of the reasons is that both structures for an ambiguous sequence are equally plausible, so using statistical information the system or program will have an equal chance in making wrong or correct prediction. For the sequence in (3c), the situation is even worse. The statistical information prefers structure (f) in Figure 4, since the probability for having a left boundary between a verb and an N is 0.723, as shown in Table 1, but the probability for a right boundary between an N and a *usde* is 0.005. Consequently, it is very unlikely for our statistics-based program to

favor structure (e) in Figure 4. That is, for sequences like (3c), an analysis of an NP with a relative clause is very unlikely. This is the reason why the C type error is very high, as shown in Table 6. Even though one could collect statistical information for more detailed tag classification, which may reflect semantic differences of some syntactic categories, we see no easy and clear ways to collect statistical information which could differentiate the structures of (e) and (f) in Figure 4 for sequence (3c).

Furthermore, although it is very difficult for a statistics-based parser to analyze the "V N" sequence in (3c) as a relative clause, it is relatively easy for a rule-based system to obtain that, since we can detect the patterns of relative clauses once a 'usde' is encountered. For instance, we can write a simple procedure to determine whether the sequence in (3c) contains a relative clause by appealing to relative clause rules or patterns.[4] The same conclusion can be applied to the sequence in (3b).

Hence statistics-based approaches are not adequate to make the necessary distinction, and some kind of rule-based approaches is necessary in extracting maximal-length NPs in Chinese. Therefore, statistics-based approaches and rule-based approaches are complimentary, and should both be employed in parsing natural languages.

5 Conclusion

In this paper, we have proposed a simple statistics-based maximal-length NP extractor for Chinese. Our experiments showed that statistics-based approaches were not adequate for maximal-length NP extraction in Chinese, since the best recall is 69.4% and the best precision, 71.3% for open tests. Therefore, it is not enough to have just the part-of-speech information and the probabilities of beginning an NP and ending an NP for NP extraction. Rule-based patterns, syntactic and semantic information should also be utilized in resolving structural ambiguities for the sequences of tags such as those, as listed in (3), which are the most problematic cases in NP extraction for Chinese, and thus a combination of statistics and rules and patterns should fare better than approaches which only employ one of them.

[4]Here, we ignore the possibility that rule-based approaches need to check semantic factors to decide whether a relative clause analysis is feasible. But it suffices to say that statistics-based approaches do not have any advantage over rule-based approaches on this matter.

Acknowledgement

The work reported in this paper is partially supported by the Hong Kong Research Grant Council under the 1994/95 Earmarked Grant for Research Initiative (RGC Ref no. CUHK 258/94E).

A The Part of Speech of Chinese Used by the System

nf	姓氏	npf	人名	npu	机构名	npr	其它专名
ng	普通名词	t	时间词	s	处所词	f	方位词
vg	一般动词	vgo	不带宾	vgn	带体宾	vgv	带动宾
vga	带形宾	vgs	带小句宾	vgd	带双宾	vgj	带兼语宾
va	助动词	vc	补语动词	vi	系动词	vy	动词“是”
vh	动词“有”	vv	来、去连谓	a	形容词	z	状态词
b	区别词	mx	系数词	mw	位数词	mg	概数词
mf	分数词	mb	倍数词	mm	数量词	mh	数词“半”
mo	数词“零”	qni	个体量词	qnc	集合量词	qnk	种类量词
qng	名量词“个”	qnm	度量词	qns	不定量词	qnv	容器量词
qnf	成形量词	qnt	临时量词	qnz	准量词	qvp	专用动量词
qvn	名动量词	rn	体词性代词	rp	谓词性代词	rd	副词性代词
p	介词	pba	把 (将)	pbei	被 (让, 叫)	pzai	在
d	副词	cf	连词前段	cpw	并连词	cpc	并连分句
cps	并连句子	cbc	分句词语间	cbs	句子间	usde	“的”
uszh	“之”	ussi	“似的”	usdi	“地”	usdf	“得”
ussu	“所”	ussb	“不”	utl	“了”	utz	“着”
utg	“过”	upb	被	upg	给	y	语气词
o	象声词	e	叹词	hm	数词前缀	hn	名词前缀
k	后缀	i	成语	j	简称语	l	习用语
x	其他	xch	非汉字	xfl	数学公式		

References

[1] Bai, Shuan-Hu. (1992) Studies and implementation of probability-based automatic part-of-speech tagging for Chinese corpora. Master's report, Tsinghua University, Beijing, China (in Chinese).

[2] Bourigault, Didier. (1992) Surface grammatical analysis for the extraction of terminological noun phrases. In *Proceedings of COLING-92*, pages 977–981, Nantes, France.

[3] Chen, Kuang-hua and Hsin-Hsi Chen. (1994) Extracting noun phrases from large-scale texts: A hybrid approach and its automatic evaluation.

[4] Church, K. (1988) A stochastic parts program and noun phrase parser for unrestricted text. In *Second Conference on Applied Natural Language Processing*, pages 136-143. Association of Computational Linguistics, Austin, Texas.

[5] Church, K. and P. Hanks (1989) Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1):22-29.

[6] Church, K. and W. Gale, et al. (1989) Parsing, word associations and typical predicate-argument relations. In *Proceedings of the 1989 DARPA Speech and Natural Language Workshop.*

[7] van der Eijk, P. (1993) Automating the acquisition of bilingual terminology. In *Proceedings of EACL'93*, Utrecht, the Netherlands.

[8] Feng, Zhiwei. (1988) The complex feature in the description of Chinese sentences. *Chinese Information Processing*, 4(3):20-29 (in Chinese).

[9] Garside, Poger and Geoffrey Leech. (1985) A probabilistic parser. In *Proceedings of Second Conference of the European Chapter of the ACL*, pages 166-170.

[10] Magerman, D. and M. Marcus. (1990) Parsing a natural language using mutual information statistics. In *Proceedings of the 28th National Conference on Artificial Intelligence.*

[11] Rausch, Norrback, and Svensson. (1992) Excerpering av nominalfraser ur löpande text. Ms., Stockholms Universiet, Institutionen för linfvistik.

[12] Salton, Gerard and Maria Smith. (1989) On the application of syntactic methodologies in automatic text analysis. *ACM.*

[13] Sampson, Jeofray. (1995) *English for the Computer: The SUSANNE Corpus and Analytic Scheme.* Oxford University Press, Oxford, UK.

[14] Sheridan, Paraic and Alan F. Smeaton. (1992) The application of morpho-syntactic language processing to effective phrase matching. *Information Processing and Management*, 28(3).

[15] Voutilainen, Atro. (1993) NPtool: a detector of English noun phrases. In *Proceedings of Workshop on Very Large Corpora: Academic and Industrial Perspectives*, pages 48–57.

[16] Yu, S. W. (1992) The design of modern Chinese grammatical electronic dictionary. In *Proceedings of the International Conference on Chinese Information Processing* (in Chinese).