基于语料库的中文最长名词短语的自动抽取

李文捷 周明[1] 潘海华[2] 林耀燊 黄锦辉
（香港中文大学系统工程和工程管理学系）

摘 要：从大规模真实文本中抽取名词短语（以下简称 NP ）具有广泛的应用领域。现在国外已有研究者尝试应用统计方法获取英文 NP ，但有关中文 NP 的研究尚无资料显示。本文提出了一种应用简单的统计原理、并以词性标注为基础的汉语最长名词短语的自动抽取方法。实验结果表明，单纯依靠词性信息的统计方法对于汉语 NP 的正确分析是不足够的。通过对错误原因的深入分析，给出了改进建议。

Corpus-based Maximal-length Chinese Noun phrase Extraction

Wenjie Li, Ming Zhou, Haihua Pan, K.F. Wong and Vincent Lum
Department of Systems Engineering and Engineering Management
The Chinese University of Hong Kong

Abstract : Acquiring noun phrases(NP) from running texts is very useful for 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. 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 approach.

1 引 言

名词短语是构成语言的重要组成部分，是传递信息不可缺少的基本单位。在自然语言处理领域，名词短语的正确分析对于机器翻译、文章索引、文献检索以及句法分析等更是具有重要

[1] 周明，中国北京清华大学计算机科学系 .
[2] 潘海华，香港九龙香港城市大学中文翻译与语言学系.

意义。按照传统的方法，对各类短语的获取和分析将意谓著对整个句子进行句法分析甚至语义检验，然后根据某种特定规则从整个句法树中摘取所需的部分子树，其复杂程度和困难程度相当大。因而，目前人们逐渐倾向于词语导向的部分句法分析方法，并已应用于英语名词短语分析以及介词短语搭配等领域。

Church(1988)设计的词性标注和简单非嵌套(non-recursive)名词短语标注系统从已标注词性的英文文本中获取 NP 起始和 NP 终止两种概率矩阵，并根据其概率信息决定是否标注以及如何标注。Church 系统的检索率高达 98%。不足之处在于实验文本太小，而且分析对象是极其简单的 NP 。

Rausch ， Norrback 和 Svensson(1992)设计的“最小名词短语(nuclear NP)”标注系统以瑞典语为研究对象，概率信息仍以词性为基础，但检索率和正确率却只有 85.9%和 84.3%。同年，Bourgault 的 *LECTER* 对法文文本进行了测试，它成功地实别了 95%的“最长名词短语(Maximal-length NP)”，但文献中却无数据显示有多少标记错误的 NP 。

1993 年， Voutilainen 设计的 *NPtool* 尝试利用辞典知识和规则库方法获取英文的“最长名词短语”。其系统性能出色，在不同的文字领域中，可分别获得 98.5%-100%的检索率和 95%-98%的正确率。但据陈光华等统计，附录中所提供的样本中却只有 85%的检索率。

之后，陈光华等(1994)综合了以上两种不同的方法，集统计方法和规则方法于一体，利用英文 SUSANNA 语料库进行了实验。其突出特点在于该系统首先利用统计信息和动态规划方法将句子划分成若干关联块，然后，根据语言特征确定出每块的语法中心词和语义中心词。最后，利用有限状态转移机制抽取合并各类名词短语。其检索率达到 96%。

迄今为止，所有关于名词短语提取和短语分析的研究都是针对英文或者非亚洲语言进行的。从它们的实验结果中可以得出这样一个结论，统计方法对于英文简单 NP 的分析是行之有效的。那末该方法对于具有完全不同语言特征的中文是否适用呢？在本文中，我们将利用统计方法进行汉语名词短语自动抽取的实验，并对结果进行统计和分析。

2 系统设计原理

2.1 语料库的加工

近年来，基于语料库的统计方法在国内外得到越来越广泛的应用。在使用该方法时，语料库的覆盖面以及加工的深度、精度都将直接影响著系统的性能。该实验所用的语料库包括 30 篇新闻报导，共计 750 个复杂句， 16660 个汉语单词。所有文章已经分词和词性标注的预处理。其中词性标注系统的标注集包含 24 大类、 110 种词性类别。此外，用于训练的文本中所有最长名词短语亦预先由人手工标注，方法是在 NP 开始处加入起始符“【”，在 NP 结束处加入终止符“】”。事实上，符号“【”和“】”分别代表 NP 左边界和 NP 右边界。例如：

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

当生语料被加工成为熟语料之后，下一个步骤便是从中提取所需的知识，并应用学习到的

知识由系统自动标注 NP 。通过对人工标注和系统标注的比较，检验系统的可行性及不足。

2.2 知识获取

所谓知识获取，是指通过对系统进行累积训练，从人工标注的熟语料中提取系统所需的知识。针对目前我们所要解决的问题，可以利用两个概率关系矩阵来描述和存贮出现 NP 左、右边界的概率。具体方法如下。

假设 W_i 和 W_{i+1} 为两个连续的汉语单词，t_i 和 t_{i+1} 分别为它们的词性，若用符号 NP_B 表示左边界，NP_E 表示右边界，则：

$$P(NP_B|t_i,t_{i+1})=\frac{freq(t_i,NP_B,t_{i+1})}{freq(t_i,t_{i+1})} \qquad P(NP_E|t_i,t_{i+1})=\frac{freq(t_i,NP_E,t_{i+1})}{freq(t_i,t_{i+1})}$$

其中 *freq* 可用该模式在语料库中出现的次数近似表示。表 1 和表 2 给出了语料库中出现频率最高的四个词性的概率关系矩阵。'a'代表形容词，'ng'代表一般名词，'p'代表介词，'vgn'代表带体词性宾语的动词。显然 p 和 vgn 之后出现左边界的概率较大，而 ng 之后出现右边界的概率较大。这里值得注意的一点是，'p''p'、'p''vgn'以及'vgn''p'都有可能出现左边界，这和英文是完全不同的。如前所述，中文 NP 的所有修饰成分均在中心名词之前，因而会出现由“PP N”或者以“vgn”开始的“RC N”型 NP 。

表 1 ．**NP 起始概率矩阵**

	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	0.084	0.723	0.333	0.438

表 2 ．**NP 终止概率矩阵**

	a	ng	p	vgn
a	0	0	0	0
ng	0.570	0.028	0.744	0.837
p	0	0	0	0
vgn	0	0	0	0

2.3 ＮＰ自动抽取

系统在抽取ＮＰ时，首先对句中任意一对连续单词，根据其词性类别，分别从上述两矩阵中查找可能出现左边界和右边界的概率。当概率值大于某一阈值时，边界标记“【”和/或“】”做为预选边界加到相应的位置。例如：单词序列「在#vgn 」「学校#ng 」，当阈值为 0.4 时，因为 $P(NP_B|p,ng)=0.728$ ，$P(NP_E|p,ng)=0$ ，则输出为「在#vgn 」【「学校#ng 」。

从下一节的实验结果中不难看出，仅仅根据概率信息所得到的预选边界虽然有较好的覆盖率，但有一半以上是多余的。该现象对于左边界尤其突出。显然，为了获得正确的最长名词短语，需要对预选边界做进一步的筛选和配对。该系统尝试了两种不同的配对原理。

原理一． 最大长度配对原理（**ML**）

最大长度配对是指当出现多个预选左、右边界时，选取距离最大的两个配对，并作为最终边界标注出短语界限。

原理二． 最大概率配对原理（MP）

最大概率配对是指当出现多个预选左、右边界时，选取概率最大的两个配对，并作为最终边界标注出短语界限。

此外，该实验还从不同的方向即从左向右和从右向左对以上两种配对方式进行了比较。

3 实验结果及分析

3.1 实验结果

预选边界的测试结果见表3（阈值=0）

表3．预选边界的统计数据

	正确 边界		错误 边界		NP	检索率 (%)		正确率 (%)	
	左	右	左	右.	总数	左	右	左	右
封闭测试	2717	2722	4040	2882	2724	99.7	99.9	40.2	48.6
开放测试	494	523	770	510	555	89.1	94.2	39.0	50.6

表4和表5给出了配对之后的系统性能评价。性能评价指标包括检索率和正确率。检索率指系统标注正确的NP在人工标注的NP中所占的比例，而正确率则是指系统标注正确的NP在所有系统标注出的NP中所占的比例。通过调整阈值发现当其值为0.1时效果最佳。

表4．配对之后的检索率（阈值=0.1）

组合 方式		从左至右 (%)		从右至左 (%)	
左边界	右边界	封闭测试	开放测试	封闭测试	开放测试
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

表5．配对之后的正确率

组合 方式		从左至右 (%)		从右至左 (%)	
左边界	右边界	封闭测试	开放测试	封闭测试	开放测试
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

实验结果表明：

【结论1】 从不同方向进行配对对名词短语的提取基本上没有影响。

【结论2】 几种不同的配对方式区别不大。比较而言，最大概率配对方法的性能略好些。

3.2 错误分析

通过进一步分析系统标记错误的NP短语，我们发现共有以下五类主要错误：

A：错误地将两个本来应该分开的 NP 合并在一起。该现象主要出现于有双宾语和有主题及主语的句子中；

B：与A相反，有时系统亦会将一个本应合在一起的 NP 断开；

C：当 NP 中包含 RC 时，系统会不可避免地将从句中的小 NP 抽出，有时亦会错误地划出 NP 边界；

D：若采用 ML 方法，当连续出现两个动词时（如 V_1V_2 N），不论 V_1V_2 是连动关系还是由 V_2 引导的从句作 NP，系统都会将左边界划在 V_2 和 N 之间；

E：与D类似的问题亦出现于 P+V 组合，系统会将 P+NP 和 P+VP 混肴。

主要错误类型的分布如表6所示。

表6．错误类型的分布

错误	遗漏的 NP		标错的 NP	
类型	数目	%	数目	%
A	60	13.3	30	4.2
B	91	20.1	186	25.8
C	87	19.3	205	28.4
D	10	2.2	10	1.4
E	9	2.0	9	1.4
其它	195	43.1	282	39.1
总共	358	100	722	100

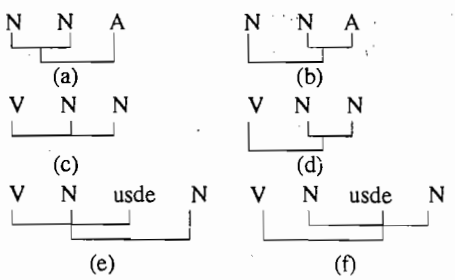
3.3 讨论

在认真地分析了该系统出错原因之后，我们看到大部分问题出在结构歧义上。下图给出了三种单词序列的可能结构。

从下图中不难看出，统计方法很难区分上述不同结构，其原因在于，在文本中两种结构出现的次数几乎相等，此时系统标注正确或标注错误的机率是相等的。对于序列 c 情况则更糟，因为 VN 之间出现左边界的概率为 0.723，而 N 和 usde 却只有 0.005 的概率，因而统计方法将永远不会倾向于(e)。这就是为什末C类错误率较高的原因。

然而规则方法却有可能解决上述问题。因此，如果将统计方法与规则方法结合起来，相信会取得较好效果。

a. N N A　　b. V N N　　c. V N usde N



4 结束语

本文提出了一种基于统计原理自动抽取汉语最长 NP 的方法，实验结果显示该方法的效果不佳，开放测试的最好检索率和正确率只有 69.4%和 71.3%。所以，仅仅依靠词性标注的统计方法是不成功的。语法结构的歧义分析有赖于基于规则的模式匹配、更深层次的语法标注以及语义知识的辅助，才能取得较好的效果。

附录　部分词性代码及含义

npf	人名	ng	普通名词	t	时间词	s	处所词
vg	一般动词	vgn	带体宾动词	vgo	不带宾动词	nvg	动名词
mx	系数词	mw	位数词	rn	体词性代词	p	介词
d	副词	usde	"的"	usdi	"地"	utl	"了"
utz	"着"	j	简称语	x	其它		

参考文献

[1] Bourigault 、Didier, Surface grammatical analysis for the extraction of terminological noun phrases, In *Proceedings of COLING-92*, pages 977-981, 1992

[2] Church, K., A stochastic parts program and noun phrase parser for unrestricted text. In *Second Conference on Applied Natural Language Processing*, pages 136-143, 1988

[3] Feng, Zhiwei, Complex features in description of Chinese language, *Chinese Information Processing*, 4(3):20-29, 1989

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

[5] Rausch, Norrback and Svensson. (1992), Excerpering av nominalfraser ur löpande text, Manuscript, Stockholms universiet, Institutionen för linfvistik, 1992

[6] Salton, Gerard and Maria Smith, On the application of syntactic methodologies in automatic text analysis, ACM, 1989

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

[8] Van der Eijk, P, Automating the acquisition of bilingual terminology, In *Proceedings of EACL'93*, Utrecht, the Netherlands. 1993

[9] Voutilainer, Atro, NPtool: a detector of English noun phrases, In *Proceedings of Workshop on Very Large Corpora: Academic and Industrial Perspectives*, Pages 48-57, 1993