规则和边界统计相结合的英语基本名词短语识别*

梁颖红[1] 赵铁军[1] 翟舒[2]

([1]哈尔滨工业大学计算机学院 哈尔滨 150001) ([2]东北林业大学外语学院 哈尔滨 150001)

E-mail:{[1]liangyh; tjzhao}@mtlab.hit.edu.cn [2]zhaishu_315@yahoo.com.cn

摘 要：基本名词短语识别在自然语言处理领域具有重要作用。本文以英语基本名词短语识别为目标，采用规则和边界统计相结合的策略识别英语基本名词短语，把基本名词短语识别分成依规则标注和用边界概率校正两个过程，通过对规则标注结果边界的修正，在一定程度上弥补了上下文无关规则不能解决边界歧义的缺点。与基于规则的方法相比，本方法可以在召回率没有明显下降的情况下大幅度提高基本名词短语识别的精确率。

关键词：基本名词短语；词性组合；规则；概率矩阵

English Base Noun Phrase Identification Based on the Combination of Rule Template and Boundary Statistic

Liang YingHong[1] Zhao TieJun[1] Zhai Shu[2]

([1]Computer Science and Engineering Department ,Harbin Institute of Technology ,Harbin 150001)

([2] Foreign Language Department , North-East Forest University ,Harbin 150001)

E-mail:{liangyh; tjzhao E-mail:{[1]liangyh; tjzhao}@mtlab.hit.edu.cn [2]zhaishu_315 @yahoo.com.cn

ABSTRACT: Finding base noun phrase is very important in the field of natural language processing. This paper aims at the identification of English base noun phrase, In particular, we present a strategy of the combination of the rule and the boundary statistic. We divide the procedure of the identification of English base noun phrase into two steps: One is tagging the corpus by rules , the other is verifying the boundary which tagged by rule through the boundary probability metric. In this way, we remedy the shortcoming that context independent rule can't deal with the ambiguity of boundary. Compare to the rule method, our method may rapidly improve the precision in the precondition of lower descendent of recall.

Key words: base noun phrase; POS combination; rule template; probability metric

*本文研究受到国家 863 计划资助(项目号 2002AA117010-09)。

1 引言

识别基本名词短语是自然语言处理领域的非常重要的子任务，名词短语的识别可以应用到机器翻译、信息检索、主题内容分析和文本处理，对名词短语的识别直接关系到文本分析和文本处理的正确性。

英语基本名词短语是非递归的名词短语，国内外有很多研究人员进行了英语名词短语识别的研究工作。文献[3]利用统计方法，得到开始和结束位置的概率矩阵，并用此概率值进行了 NP 的边界识别；文献[5]将表示基本名词短语句法组成的基本结构模板与表示基本名词短语出现的上下文环境特征的转换规则相结合用于识别英语基本名词短语；文献[8]在提出汉语基本名词短语的概念后，也用和文献[5]相同的方法对汉语基本名词短语进行识别，并得出基本结构模板是识别基本名词短语的必要条件，而不是充要条件；文献[6]采用基于实例的学习方法，把待识别语料和事先存储的实例集相比较，把和待标语料距离最近的例子作为标注的标准对待标语料进行标注；文献[7]建立了用于识别英语基本名词短语的统一统计模型，把词性标注和基本名词短语识别集成在一个统一的统计模型中，目的是为了保证基本名词短语的识别尽可能在词性标注正确率较高的情况下进行。

我们在识别过程中把 Church 的边界统计方法和规则方法相结合，通过设定阈值把统计得到的开始和结束概率矩阵值作为对规则标注结果的校正依据，弥补了上下文无关的规则在识别基本名词短语边界方面的缺陷。通过测试，本方法可以在召回率没有明显下降的情况下大大提高精确率。

2 基于规则的英语基本名词短语识别

2．1 规则的获取

规则方法是自然语言处理领域中经常采用的方法，我们在识别过程的前部分采用了规则的方法。规则的获取过程如下：首先从已经标注有基本名词短语信息的 WSJ（15-18）训练语料中得到组成基本名词短语的词性组合，然后去掉小于出现 5 次的词性组合，最后把剩余的词性组合作为规则。我们从训练语料中共得到 1162 个规则。

表 1 英语基本名词短语的规则例表

模板	示例	模板	示例
NN NNS	administration/NN officials/NNS	DT NN	the/DT administration/NN
PRP	he/PRP	PRP$ VBG NNS	its/PRP$ continuing/VBG problems/NNS

2．2 用规则标注语料

用规则标注语料时采用最长匹配原则。例：

[consumer/NN expenditure/NN data/NNS]　正确

[consumer/NN] [expenditure/NN][data/NNS]　错误

统计显示，如果将训练语料中所有符合规则的词语序列全部标为基本名词短语，封闭测试召回率是 88.55%，精确率是 78.04% ；开放测试召回率是 87.57%，精确率是 77.47% 。这说明仅仅依靠上下文无关的规则不能解决基本名词短语识别中边界歧义问题。

2.3 边界概率矩阵方法识别英语基本名词短语

概率矩阵方法是 Church(1998) 使用的一种识别名词短语的方法。这种方法首先从标注好的语料中统计名词短语开始和结束位置的词性符号和这一符号的前一个词性符号，进而得到两个矩阵，一个是名词短语开始位置的概率矩阵，一个是结束位置的概率矩阵。

表 2 名词短语开始的位置矩阵

	AT	NN	NNS	VB	IN
AT	0	0	0	0	0
NN	0.99	0.01	0	0	0
NNS	1.0	0.02	0.11	0	0
VB	1.0	1.0	1.0	0	0
IN	1.0	1.0	1.0	0	0

该方法从输入的句子中得到相邻的两个词性标注，根据概率矩阵取出概率最大的插入开始和结束标志，名词短语的左边界用"["，右边界用"]"表示。Church 在 Brown 语料上实现了该方法，但 WSJ(15-18 训练语料，20 测试语料)是识别基本名词短语常用的测试集[5]，因此我们在 WSJ(15-18 训练语料，20 测试语料)语料上按照 Church 方法把插入的基本名词短语边界进行了括号匹配，得到了识别出的基本名词短语精确率和召回率。结果如下：精确率 92.97%，召回率 86.98%。

3 基于规则和边界统计相结合的方法识别英语基本名词短语

3．1 系统的流程

按照规则对英语基本名词短语识别，只是考虑到词性的组合规律，没有考虑组成基本名词短语边界词性的特点，因此这种方法标注的错误主要发生在开始和结束的边界位置上。而 Church 的边界统计方法却对组成基本名词短语的边界位置的词性和前一词性的组合特点进行了统计，该方法在识别英语基本名词短语边界方面具有一定的优势。鉴于此，我们充分利用边界统计方法在边界识别方面的优势，以弥补规则不能解决边界歧义的弱点。提出了规则

和边界统计相结合的方法，系统流程图见图 1。

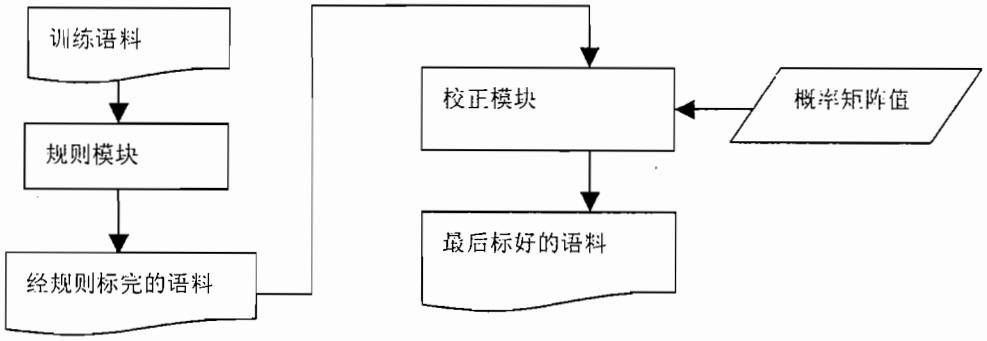

图 1 本系统的流程图

3．2 阈值的设定

在边界统计中，我们通过计算得到了英语基本名词短语开始和结束位置的概率值，把是否插入开始和结束边界的概率分界值作为阈值，当待标语料前后两个词性之间插入开始或结束边界的概率值大于此阈值时，插入开始或结束符号，否则不插入边界符号。下表是当分界值分别为五种不同情况时得到的结果：

表 3　5 次实验结果表

阈值	封闭精确率	开放精确率	封闭召回率	开放召回率
0.1	96.59	96.37	86.47	85.03
0.3	98.23	97.76	85.41	84.06
0.5	98.49	97.96	83.85	82.61
0.7	98.84	98.23	81.89	80.54
0.9	98.87	98.16	76.95	75.71

我们发现当分界值为 0.3 时，系统的总体性能最好，因此，把 0.3 最为最后的阈值。

3．3 边界概率矩阵对依据规则标注结果的校正

我们提出的规则和边界统计相结合的方法充分发挥了规则的词性组合具有规律性的优势，也融合了边界统计在识别基本名词短语边界方面的独到之处。具体实现步骤是：先用训练得到的规则对语料按照最大长度匹配来进行标注，然后对得到的标注结果用边界概率矩阵进行修正：取出经规则插入的开始和结束标志的词性和前一个词性，在概率矩阵中查出在这两个词性之间插入该标志的概率，如果此概率大于阈值,则保留此标记，否则去掉该标记，同时也去除和此标记相对称的标记符号。例：

[Arbitragers/NNS] [were/VBD n't/RB] [the/DT only/JJ big/JJ losers/NNS]　in/IN [the/DT collapse/NN]　of/IN [UAL/NNP Corp./NNP stock/NN]　./.

这是依据规则得到的标注结果，用概率矩阵校正时，首先获得第一个边界符号的词性 NNS

和前一个单词的词性“ ”（对于句首的词，其前一个词的词性为空串，句尾的词，其后一个词的词性为空串）。然后在开始概率矩阵里查找空串和 NNS 之间插入"["的概率，如果此概率值大于阈值,保留"["，否则去掉"["，并把和它相对应的"]"也删除。依此类推，检查其他的标记符号。如果开始概率矩阵中 NNS 和 VBD 的概率值小于阈值,其他的标记的概率矩阵值都大于等于阈值，上句的校正结果如下：

[Arbitragers/NNS] were/VBD n't/RB [the/DT only/JJ big/JJ losers/NNS] in/IN [the/DT collapse/NN] of/IN [UAL/NNP Corp./NNP stock/NN] ./.

4 实验结果及分析

4．1 测试指标

性能评估指标为英语基本名词短语识别的精确率、召回率和 $F_{\beta=1}$，公式如下：

精确率： $precision = \frac{a}{b} \times 100\%$；

召回率： $recall = \frac{a}{c} \times 100\%$；

$$F_{\beta=1} = \frac{(\beta^2+1) \times 精确率 \times 召回率}{\beta^2 \times 精确率 + 召回率}$$；

其中，a 是识别出的正确的基本名词短语个数，b 是被识别为基本名词短语的词串总数，c 是测试集中的基本名词短语总数。

4．2 实验结果及分析

我们使用普遍采用的基本名词短语识别的 WSJ（15-18）作为训练集，WSJ(20)作为测试集，分别在三组测试语料（一个为 WSJ15-18（2.18 兆），一个为 WSJ00-18（10.4 兆）和 WSJ00-23（13.8 兆））上训练得到开始和结束的概率矩阵，并分别用它们的概率矩阵对同一经规则标注的结果进行了校正，在阈值为 0.3 时，结果见下表：

表 4 三组概率值下的结果比较

训练规模		WSJ15-18	WSJ00-18	WSJ00-23
精确率	封闭测试	98.23	97.82	97.79
	开放测试	97.76	97.55	97.55
召回率	封闭测试	85.41	85.4	85.41
	开放测试	84.06	84.11	84.15

表 5 本文方法与规则和边界统计方法的比较表

		基于规则模版	基于边界统计	规则和边界统计相结合的方法
精确率	封闭测试	78．04	93．59	98．23
	开放测试	77．47	92．97	97．76
召回率	封闭测试	88．55	88．87	85．41
	开放测试	87．57	86．98	84．06
$F_{\beta=1}$	封闭测试	82．96	91．17	91．37
$F_{\beta=1}$	开放测试	82．21	89．88	90．39

表 5 表明：与基于规则的方法相比，规则和边界统计相结合的方法可以在召回率只下降 3%的前提下，使精确率提高了 20%。这说明边界统计方法在一定程度上可以弥补上下文无关规则不能解决边界歧义的缺陷，因而大大提高精确率。

5 结束语

本文研究用规则和边界统计方法识别英语基本名词短语，将名词短语的识别分为依规则标注和据概率矩阵校正两个层次，后者对前者进行了修正。通过把统计得到的概率矩阵信息融入到上下文无关的规则中，扩大了上下文信息，很好地实现了统计、规则的统一，在一定程度上解决了利用规则无法避免边界歧义的困难。通过测试，证明了该方法对提高英语基本名词短语识别的精确率是有效的。

参考文献

[1] Brill: "Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part of Speech Tagging."Computational Linguistics, Dec. '95.

[2] Cardie,Claire and Pierce,David: "Error-driven pruning of treebank grammars for base noun phrase identification."In Proceedings of COLING-ACL'98. pp. 218-224.

[3] Church,,K: "A stochastic parts program and noun phrase parser for unrestricted text." In Proceedings of the Second Conference on Applied Natural Language Processing,1988., pp136-143.

[4] Eric F. Tjong Kim Sang : "Introduction to the CoNLL-2000 Shared Task: Chunking." In proceedings of CoNLL-2000 and LLL-2000,pages 127-132,Lisbon,Portugal,2000.

[5] Lance A.Ramshaw and Mitchell p.Marcus : "Text chunking using transformation_based learning. In Natural language processing using very large corpora."Kluwer. Originally appeared in WVLC-95,pp.82-94.

[6] Walter Daelemans, Sabine Buchholz, Jorn Veenstra: " Memory-Based Shallow Parsing." the CoNLL-99 workshop

[7] Xun EnDong: "An Unified Statistical Model to Identify English BaseNP."ACL-2000,The 38th Annual Meeting of the Computational Linguistics ,HongKang,3-6 October 2000.

[8] 赵军、黄昌宁： "基于转换的汉语基本名词短语识别模型"，1999《中文信息学报》第 13 卷第 2 期。