中 文 信 息 学 报
第 13 卷 第 2 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 13 No. 2

基于转换的汉语基本名词短语识别模型*

赵 军 黄昌宁

智能技术与系统国家实验室 清华大学计算机科学与技术系 北京 100084

摘要 基本名词短语的识别在自然语言信息处理领域具有重要作用。本文首先从语言学的角度提出了汉语基本名词短语的概念,然后从语言信息处理的角度将用于基本名词短语识别的知识分为两部分,即表示基本名词短语句法组成的基本结构模板(静态知识)与表示基本名词短语出现的上下文环境特征的转换规则(动态知识)。在此基础上设计了一种基于转换的基本名词短语识别模型,该模型可同时结合这两类知识识别基本名词短语。实验结果显示了较高的识别正确率。

关键词 自然语言处理 知识获取 语料库 名词短语

一、引言

在自然语言信息处理领域,名词短语的正确识别和分析对机器翻译、信息检索、文本分类以及句法分析都具有重要作用。本文讨论的是一种特殊的名词短语——基本名词短语(baseNP)的识别问题。关于baseNP,在英语中被定义为"非嵌套的名词短语"[1],有关汉语baseNP的研究还未见报道。本文首先从语言学的角度提出了汉语baseNP的概念,然后从语言信息处理的角度分析了识别汉语baseNP所需的知识,包括表示baseNP句法组成的基本结构模板(静态知识)与表示baseNP出现的上下文环境特征的转换规则(动态知识)。进而,设计了一种基于转换的baseNP识别模型,该模型可同时结合这两类知识识别baseNP。实验结果显示了较高的识别正确率和召回率。

二、汉语基本名词短语的定义及识别知识

2.1 汉语基本名词短语的定义

Church将英语基本名词短语定义为"非嵌套的名词短语",即它的内部不再包含更小的名词短语[1]。这一定义不能满足汉语语言信息处理的要求,例如:"自然语言处理"、"亚洲金融危机"和"政治体制改革进程"等虽然不是非嵌套的名词短语,但对于信息检索和机器翻译等都是很有意义的。在汉语语言学中,名词短语的定语分为三种类型,即限定性定语、描写性定语和区别性定语[6]。本文从限定性定语出发给出汉语基本名词短语的形式化描述。

【定义1】 基本名词短语(简称baseNP):

baseNP →baseNP + baseNP

* 本文于1998年3月3日收到

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

baseNP →baseNP + 名词 | 名动词

baseNP →限定性定语 + baseNP

baseNP →限定性定语 + 名词 | 名动词

限定性定语 →形容词 | 区别词 | 动词 | 名词 | 处所词 | 西文字串 | (数词 + 量词)

由此,名词短语可以分为 baseNP 和 ～ baseNP (非基本名词短语),以下举例说明:

表 1 baseNP 和 ～ baseNP 示例

baseNP	～ baseNP
甲级联赛 产品结构 下岗女工	复杂的特征 这台计算机 对于形势的估计
促销手段 太空旅行 自然语言处理	明朝的古董 11 万职工 高速发展的经济
企业承包合同 第四次中东战争	很大成就 研究与发展 老师写的评语

2.2 基本名词短语的识别知识

从语言信息处理的角度来看,识别文本中的 baseNP 就是对于给定的经过分词和词性标注的文本 $T = w_1/t_1 \quad w_2/t_2 \cdots w_n/t_n$,识别出 T 中所有的 baseNP。一般情况下,组成 baseNP 的词语序列所对应的词性序列呈现一定的规律性,例如:名词和名词同现时可以构成 baseNP,而副词和动词同现时不会构成 baseNP。本文将这类表示 baseNP 内部句法组成的词性序列称为 baseNP 的结构模板,这些结构模板可以作为识别 baseNP 的静态知识。在识别 baseNP 时,可以首先抽取文本中所有符合这些结构模板的词语序列组成 baseNP 候选集。然而,单纯依靠结构模板还不足以正确识别文本中的 baseNP,必须利用 baseNP 在上下文环境中的分布特征来从候选集中筛选出正确的 baseNP。本文将 baseNP 在上下文环境中的分布特征称为识别 baseNP 的动态知识。

综上所述,利用计算机识别文本中的 baseNP 时,应该将 baseNP 识别的静态知识和动态知识结合起来,以获得较高的识别正确率。本文采用的是一种可结合静态知识和动态知识的转换模型。

三、基于转换的 baseNP 识别模型

基于转换的 baseNP 识别模型的特点是:通过一个基于转换规则的文本转换机制,将识别 baseNP 的静态知识和动态知识结合起来,从而充分利用 baseNP 的内部组成结构模板和在上下文环境中的分布特征进行识别。以下,首先介绍表示 baseNP 在上下文环境中的分布特征的上下文有关规则的获取,然后介绍基于转换规则的 baseNP 识别模型。

3.1 识别 baseNP 的上下文有关规则的获取

将 baseNP 在上下文环境中的分布特征用一组上下文有关规则表示。本研究利用错误驱动的学习方法[3]获取识别 baseNP 的上下文有关规则。规则的获取流程为:首先根据基本结构模板对文本进行 baseNP 的初始标注,然后比较初始标注结果和正确答案,以发现标注中的错误。同时定义上下文有关的转换规则空间,遍历规则空间中的每一条规则,将其运用于对错误识别结果的校正中,然后用评价函数对各候选规则打分,挑选出得分最高的一条规则,并应用该规则对当前标注结果进行转换。循环利用上述方法从错误中逐渐学习到新的上下文有关的转换规则,组成转换规则序列。这种规则获取方法的关键在于以下两点: ①转换规则空间的

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

定义；②评价函数的定义。

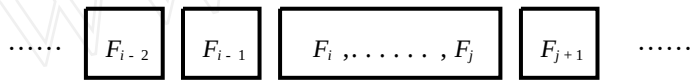
1. 转换规则空间的定义

在基于转换的学习算法中，转换规则空间定义为转换规则模板的集合。每一条转换规则模板包括转换动作和触发环境两个要素。其中触发条件限定了上下文环境的若干特征，转换动作根据这些特征更新标注结果。

本文将转换动作定义为 baseNP 候选标记、baseNP 确认标记和 baseNP 否定标记之间的转化。当某个 baseNP 候选词串的上下文环境满足某个触发条件时，就执行相应的转换动作。转换动作的形式如下：

- 转换动作 1：将词串 W 上的候选标记转化为确认标记；
- 转换动作 2：将词串 W 上的候选标记转化为否定标记；
- 转换动作 3：将词串 W 上的确认标记转化为否定标记；
- 转换动作 4：将词串 W 上的否定标记转化为确认标记。

根据语言学知识，本文将 baseNP 候选词串的触发环境限定为其上下文中的前两个词和后一个词以及该词串中的第一个词和最后一个词，而对 baseNP 产生影响的是其中每个词的词性、义类和音节数等属性。由此，本研究如下定义触发环境。令 $w_0, \cdots, w_i, \cdots, w_j, \cdots, w_n$ 为一个汉语句子，其中 $w_i, \cdots, w_j$ 为一个 baseNP 候选词串，若用 F_i 表示词 w_i 的相关属性集，则 $F_i = \{POS_i, SENSE_i, SYL_i\}$，其中 POS(i)、SENSE(i)和 SYL(i)分别表示词 w_i 的词性、义类和音节数信息，其中词性标记集见附录，义类代码取自《同义词词林》[6]的大类、中类和小类。触发环境可描述如下：



baseNP 候选串

图 1　baseNP 触发环境示意图

进而，本研究从上述的触发环境中定义了 20 条触发条件。

2. 评价函数

该学习过程由训练集的错误标注作为驱动，以标注结果错误率的降低程度作为评价函数。如果应用某一条转换规则于当前语料库，正确标注的词串数越多，则该规则的得分越高；错误标注的词串数越多，则该规则的得分越低。所以，对规则的评价函数定义如下：

$$F(r) = C(r) - E(r)$$

其中，F 为转换规则的评价函数，r 为规则，C(r)为应用规则 r 后错误标记改为正确标记的数目，E(r)为应用规则 r 后正确标记改为错误标记的数目。

3.2　基于转换的 baseNP 识别模型

该模型的输入为经过分词和词性标注的汉语文本，输出为标注了 baseNP 的文本。基本原理为：首先利用 baseNP 的基本结构模板(识别 baseNP 的静态知识)，对文本进行 baseNP 的初始标注，抽取文本中所有符合这些结构模板的词语序列组成 baseNP 候选集。然后依次利用习得的转换规则(识别 baseNP 的动态知识)对初始标注结果进行修改，直至转换规则集中的所有规则都已用过。

算法 1：基于转换的 baseNP 识别算法

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

设 C 是未标注 baseNP 的语料库,SM 是 baseNP 的基本结构模板,TS 是转换式集合;

(1) 应用 SM 对 C 进行初始标注,得到当前标注 $C^{[0]}$;

(2) 重复以下几步,直到不存在转换规则 r,r ∈TS 并且 r 未被本识别过程应用过。

循环第 i 步,

1 从 TS 中顺序取第一条未被本过程应用过的转换规则 $r^{[i]}$;

2 利用 $r^{[i]}$对文本 $C^{[i-1]}$进行转换,得文本 $C^{[i]}$;

(3) 在文本 $C^{[i]}$中,如果仍然存在歧义标记,即某个词串对应两个或两个以上的确定标记,则保留其中最长的词串所对应的标记,删除其余的标记,得到输出文本 CT。

(4) 输出文本 CT 为 baseNP 的识别结果。

四、实验及评测

实验分四部分:①baseNP 基本结构模板的获取;②利用错误驱动的学习方法从训练集中获取上下文有关的转换规则;③将基本结构模板与上下文有关的转换规则结合起来识别文本中的 baseNP;④词汇语义属性对于 baseNP 识别的作用。

4.1 baseNP 基本结构模板的获取

从 baseNP 的定义可以看出,文本中的名词 N(或名动词 VN)与前面的若干个限定性定语组合成 baseNP。而限定性定语可以是形容词 A、区别词 B、动词 V、名词 N、处所词 S、西文字串 XCH 以及序数词 MX 和量词 Q 的组合,因此,可以将 baseNP 的初始结构模板定义为:

baseNP ::≡< A| B| V| N| S| X| MX + Q > { < A| B| V| N| S| X| MX + Q > } + N | VN

该模板是一个递归定义的模板。在真实语料中,其中的许多模板是不出现或很少出现的,因此应该通过学习标注了 baseNP 的语料库,从初始模板中剔除这些模板,而仅保留那些与识别 baseNP 密切相关的结构模板。

本研究用来标注 baseNP 的语料规模为 10 万字,包含 65568 词次、6712 词形,分布在科技、科普、法律、军事和新闻五个领域。在对它进行自动分词和词性标注的基础上,根据 baseNP 的定义,进行 baseNP 的人工标注。从人工标注 baseNP 的语料库中统计得到 407 个 baseNP 的结构模板,其中出现5次以上的结构模板有64个,覆盖了语料库中98.6 %的 baseNP,本文称为基本结构模板。表 2 列出了一些二词组合的基本结构模板。

表 2 二词组合的基本结构模板

模板	示例	模板	示例	模板	示例
A + NG	旧模式	VN + NG	检索方法	VGO + NG	下岗女工
B + NG	国产电冰箱	NG + VN	太空旅行	S + NG	海底光缆
N + N	市场经济	VGN + NG	主管部门	XCH + NG	IBM 公司

4.2 转换规则的获取

用错误驱动的转换式学习方法从上述 10 万字的训练集中学习 baseNP 的上下文环境特征,在评价函数阈值 T = 0 时,得到了 380 条转换规则,以下列出 10 条常用的规则。

1. change 候选标记 to 确定标记

when POS(p_{-1}) = QN AND POS(p_1) = 。

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

例: 该/R 公司/NG 与/CM 外商/NN 签定/VN 两/A 项/QN [承包/VNN 合同/NG] 。/。

2. change 候选标记 to 确定标记

when POS(p_{-1}) = CM AND POS(p_{-2}) = BN

例: …将/P 各/R 种/QN [反/H 坦克/NG 火力/NG] 和/CM [防/H 坦克/NG 障碍物/NG] 密切/A 结合/VGN …

3. change 候选标记 to 确定标记

when POS(p_{-1}) =“ AND POS(p_1) = ”

例:这/R 种/QN 语法/NG 已经/D 成为/VGN 许多/MG 立足/VGO 于/P“/“ [复杂/A 特征/NG] ”/”的/USDE“/“ [合一/NG 运算/VNN] ”/”的/USDE [形式化/VNO 方法/NN] 的/USDE 基础/NG 。/。

4. change 候选标记 to 确定标记

when SENSE(p_{-1}) =Ja02

例: 这/R 种/QN 气候/NG 叫做/VGN/Ja02 [热带/NG 雨林/NG 气候/NG] …

5. change 候选标记 to 确定标记

when POS(p_{-1}) = P AND POS(p_1) = V

例: 在/P [上海/NG 战役/NG]结束/VGO 后/F , ……

6. change 候选标记 to 确定标记

when POS(p_{-1}) = M

例: 许多/MG [亏损/VGO 企业/NG] 将/D 转产/VGO , ……

7. change 确定标记 to 否定标记

when POS(p_{-1}) = M AND POS(p_1) = U

例: 许多/MG 地方/NG 分布/VN 着/UT 茂密/A 的/USDE 热带/NG 雨林/NG , ……

8. change 候选标记 to 否定标记

when POS(p_{-1}) = D .AND. POS(BEGIN(W)) = VGN

例: 两/MJ 国/NG 政府/NG 今天/T 联合/D 发表/VGN 建交/VNO 公报/NG, ……

9. change 候选标记 to 确定标记

when SENSE(p_{-1}) = Hc11 .AND. SENSE(END(W)) = Dk14

例: 两/MJ 国/NG 政府/NG 今天/T 发表/VGN/Hc11 [建交/VGO 公报/NG/Dk14] , ……

10. change 候选标记 to 确定标记

when SENSE(p_{-1}) = Ie02. AND. POS(p_1) = 。

例: …组成/VGN/Ie02 [防/H 步兵/NG 火力/NG 配系/NG] 。/。

从上述示例中可以看出,基于转换的学习方法获取的转换规则是确定性的。如果独立地看某一条规则,它可能不是完全正确的(如:规则示例 6),即转换动作并不完全适用于该触发环境。但是该规则的使用是确定的,少部分错误转换的词语序列将由后续的转换规则(如:规则示例 7)补救。因此,规则的使用是有序的,前面的规则更具归纳性,而后续的规则则更具特殊性。

4.3 baseNP 的识别

实验首先用单纯基于基本结构模板的方法识别 baseNP,并对标注结果进行测试;再用基本结构模板和转换规则相结合的方法重新识别文本中的 baseNP,并对标注结果进行测试。

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

实验分封闭测试和开放测试两部分,封闭测试语料和开放测试语料的规模都为 10 000 字,其中封闭测试语料来自训练集的 5 个分类领域;开放测试语料来源于训练集外。测试指标如下:

$$\text{精确率}: p = \frac{a}{b} \times 100\% \qquad \text{召回率}: r = \frac{a}{c} \times 100\%$$

其中 a 是识别正确的 baseNP 数,b 是识别为 baseNP 的词串数,c 是文本中实际存在的 baseNP 数。测试结果如表 3 所示。

表 3 baseNP 识别的测试结果

测试类型	基于结构模板的方法		结构模板与转换规则相结合的方法	
	精确率	召回率	精确率	召回率
封闭测试	72.9%	77.5%	93.1%	95.2%
开放测试	72.7%	78.3%	89.3%	92.8%

从以上测试结果可以看出,结构模板和上下文转换规则相结合的 baseNP 识别方法优于单纯基于结构模板的 baseNP 识别方法。

4.4 词汇语义属性对 baseNP 识别的作用

在 4.2 节的转换规则实例中,规则 4、9 和 10 是与词汇语义属性相关的转换规则。在实验得到的 380 条转换规则中,与词汇语义属性相关的转换规则有 335 条,可见词汇语义属性对 baseNP 识别有重要作用。为了对这种作用做定量测试,本文进行了如下对比实验。

1. 基于词语句法属性的方法:从转换式空间中删除所有与词汇语义属性相关的规则模板,在该转换规则空间上进行转换规则的学习,获得 116 条基于词语句法属性的转换式规则,将它们用于 baseNP 识别;

2. 基于词语句法属性和词汇语义属性的方法:在未作删除的转换规则空间上进行转换规则的学习,获得 380 条基于词语句法属性和词汇语义属性的转换规则,将它们用于 baseNP 识别,测试结果如表 4 所示。

表 4 词汇语义属性对 baseNP 识别作用的测试数据

测试类型	基于句法属性的方法		基于句法、词汇语义属性的方法	
	精确率	召回率	精确率	召回率
封闭测试	86.5%	95.0%	93.1%	95.2%
开放测试	83.2%	93.5%	89.3%	92.8%

从以上测试结果可以看出,加入词汇语义属性后,汉语 baseNP 识别精确率提高 6%左右,而召回率没有明显的变化。而 L. A. Ramshaw 的实验表明[5],加入词汇语义属性后,英语 baseNP 识别精确率提高 1%。说明词汇语义属性对于缺乏形态标记的汉语 baseNP 识别的作用更大一些。

五、结束语

本文从语言学的角度定义了汉语 baseNP,介绍了用于识别 baseNP 的静态知识(baseNP 结构模板)和动态知识(baseNP 的上下文环境特征)。在此基础上提出了一种将两种知识相结合的 baseNP 识别模型,该模型的实验结果显示了较高的识别正确率。分析实验结果可以看

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

出,目前的识别模型不适合考虑远距离依存现象,因此有必要在模型中加入词语搭配信息,这种知识可以来源于搭配词典,也可以依据两个词的关联程度[2]从语料库中学习而来。

附录:词性标记集

1. 名词 N
1.1 普通名词 NG
1.2 人名 NF
1.3 地名 NL
1.4 机关名 NU
2. 时间词 T
3. 处所词 S
4. 方位词 F
5. 动词 V
5.1 助动词 VA
5.2 系动词 VI
5.3 趋向动词 VQ
5.4 是 VY
5.5 有 VH
5.6 名动词 VN
5.6.1 可带宾语的动词 VNN
5.6.1 不可带宾语的动词 VNO
5.7 普通动词 VG
5.7.1 体宾动词 VGN
5.7.2 谓宾动词 VGV
5.7.3 不可带宾语的动词 VGO
6. 形容词 A
7. 状态词 Z
8. 区别词 B
9. 数词 M
9.1 基数词 MJ
9.2 序数词 MX
9.3 其它数词 MG
10. 量词 Q
10.1 名量词 MQ
10.2 动量词 QV
11. 代词 R
12. 介词 P
13. 副词 D
14. 连词 C
14.1 前置连词 CF
14.2 中置连词 CM
14.3 后置连词 CN
15. 结构助词 U
15.1 “的”USDE
15.2 “地”USDI
15.3 “得”USDF
15.4 “似的”USSI
15.5 “所”USSU
15.6 “之”USZH
15.7 时态助词 UT
15.8 其它助词 UX
16. 语气词 Y
17. 象声词 O
18. 叹词 E
19. 前缀 H
20. 后缀 K
21. 成语 I
22. 简略语 J
23. 习用语 L
24. 其它 X
24.1 非汉字字符串 XCH
24.2 标点符号(各自成一类)

参 考 文 献

[1] K. W. Church, A Stochastic Parts Program and Noun Phrase Parser for Unrestricted Test, Proceedings of the Second Conference on Applied Natural Language Processing, p136～143, 1988

[2] K. W. Church, Word Association Norms, Mutual Information, and Lexicography, Proceedings of the 27th Annual Meeting of the Association for Computational Linguistics, 1989

[3] Eric Brill, Transformation-Based Error-Driven Learning and Natural Language Processing: A Case Study in Part-of-Speech Tagging, Computational Linguistics, V21. No.4, 1995

[4] L. A. Ramshaw, M. P. Marcus, Text Chunking Using Transformation-Based Learning, Proceedings of the Fourth Workshop on Very Large Corpus, p82～94, 1995

[5] 张卫国.三种定语、三类意义及三个槽位.中国人民大学学报.1996 年 No.4: p97～100

[6] 梅家驹、竺一鸣等.同义词词林.上海辞书出版社.1983

(下转 39 页)

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

Modification of the Chinese Character Segmentation Method Based On Units Amalgamation

Zhou Pin , Ma Shaoping and Jiang Zhe
State Key Laboratory of Intelligent Technology and Systems
Department of Computer Science and Technology , Tsinghua University , Beijing , 100084

Abstract This paper introduces the modification of the Chinese character segmentation method based on units amalgamation. This modified method alters the advanced amalgamation part which is the core of the original method. Because the modified method looks for the best amalgamating combination from all the units which can be amalgamated , it can avoid some amalgamation errors which will be caused by the advanced amalgamation in the original method. By many tests on actual samples , the modification does not decline the performance of the original method , and it removes some errors and effectively improves the segmentation correct rate.

Keywords units amalgamation segmentation method advanced amalgamation the best amalgamating combination

---

(上接 7 页)

A Transformation - Based Model for Chinese BaseNP Recognition

Zhao Jun Huang Changning
Department of Computer Science & Technology Tsinghua University Beijing 100084

Abstract It is important to recognize the baseNP in the field of natural language processing. At first , the paper defines Chinese baseNP from the linguistic standpoint. Then the knowledge which is essential for baseNP recognition is analyzed from the standpoint of automatic language information processing. The recognition knowledge includes the basic construction templates which specify the syntactic composition of baseNPs (static knowledge) and the context-sensitive transformative rules (dynamic knowledge) which reflect the context features. Based on the above knowledge , a transformation-based model for recognizing Chinese baseNP is put forward , which incorporates the static knowledge and the dynamic knowledge into an organic whole to recognize the baseNPs in Chinese texts. The experiment shows a satisfatory precision and recall.

Keywords Natural Language Processing Knowledge Acquisition Corpus Noun Phrase

© 1994-2009 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net