中 文 信 息 学 报
第 17 卷 第 5 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 17 No. 5

文章编号:1003 - 0077(2003)05 - 0001 - 05

词性标注中生词处理算法研究①

张孝飞[1,2],陈肇雄[2],黄河燕[2],蔡智[1]

(1. 中国科技大学 计算机系,合肥 230000; 2. 中国科学院 计算机语言信息工程研究中心,北京 100083)

摘要:词性兼类是自然语言理解必须解决的一类非常重要的歧义现象,尤其是对生词的词性歧义处理有很大的难度。文章基于隐马尔科夫模型(HMM),通过将生词的词性标注问题转化为求词汇发射概率,在词性标注中提出了一种生词处理的新方法。该方法除了用到一个标注好的单语语料库外,没使用任何其他资源(比如语法词典、语法规则等),封闭测试正确率达 97 %左右,开放测试正确率也达 95 %左右,基本上达到了实用的程度。同时还给出了与其他同样基于 HMM 的词性标注方法的测试比较结果,结果表明本文方法的标注正确率有较大的提高。

关键词:计算机应用;中文信息处理;自然语言理解;词性兼类;隐马尔科夫模型;语料库

中图分类号:TP391 **文献标识码**:A

An Approach of Processing New Words Based on HMM in Tagging of Speech of Part

ZHANG Xiao-fei[1], CHEN Zhao-xiong[2], HUANG He-yan[2], CAI Zhi[1]

(1. Dept. of Computer Science, USTC, Hefei 230000, China;
2. Research Center of Computer & Language Information Engineering, CAS, Beijing 100083, China)

Abstract: Ambiguity of part of speech (POS) which urgent needs to be resolved is a very important ambiguous phenomenon in natural language processing. Furthermore, it is very difficult to disambiguate the ambiguity of part of speech of the new words. In this paper, through converting the problem of tagging of POS to the problem of calculation of word's emission probability; a new approach based on HMM is proposed to solve this problem. This approach uses nothing more than a tagged corpus (e.g. no grammar dictionaries, no grammar rules), and the result shows that the correct rata arrive at 97 % in close test and 92 % in open test.

Key words: computer application; Chinese information processing; natural language processing (NLP); ambiguity of POS; HMM; corpus

一、引言

在自然语言处理过程中,会出现许多歧义问题。其中非常重要的一类歧义就是词性兼类。词性或者称为词类,是词汇最重要的属性,也是连接词汇到句法的主要桥梁。因此词性标注必须为后续自然语言处理过程提供高准确率的中间结果。

现在使用的词性标注方法概括起来主要有两种:基于规则的方法和基于统计的方法。统

① **收稿日期**:2003 - 03 - 11
基金项目:国家自然科学基金资助项目(60272088)
作者简介:张孝飞(1970 —),男,博士生,主要研究方向为自然语言处理、机器翻译.

计语言模型已成功的应用到了许多领域:比如语音识别[1]、信息检索[2]和口语理解[3]等。近年来,基于隐马尔科夫模型(Hide Markov Model,简称 HMM)[4]的统计方法在词性标注中也得到了广泛应用[5~7]。统计的方法可以避免规则方法的许多缺陷,例如,它利用的知识主要是统计数据,可以从语料库中利用有指导或无指导的学习方法得到,从而避免了人工获取规则的繁琐过程。同时,获取的知识具有客观性好、一致性强等特点,处理生词和不规范句子的能力比基于规则的方法有较大的提高。但如何处理生词仍然是统计方法所面临的最大难题之一[8]。

本文基于隐马尔科夫语言模型,利用 PFR 人民日报标注语料库作为实验材料,提出了一种针对词性标注中处理生词的新方法。

二、利用 HMM 进行词性标注的问题描述

假设有一个词序列 $w_1, w_2, \cdots, w_L$,现要找一个词性序列 $C_1, C_2, \cdots, C_L$,由于词性歧义现象普遍存在,所以一个词序列可以对应不同的词性序列,而我们要找的词性序列就是使概率 $P(C_1, C_2, \cdots, C_L | w_1, w_2, \cdots, w_L)$ 最大的词性序列。

根据 Bayes 公式得:

$$P(C_1, C_2, \cdots, C_L | w_1, w_2, \cdots, w_L) = \frac{P(w_1, w_2, \cdots, w_L | c_1, c_2 \cdots, c_L) \times P(c_1, c_2, \cdots, c_L)}{P(w_1, w_2, \cdots, w_L)} \tag{1}$$

对于同一词串,上式分母的值都相同,所以上面公式可以简化为求下式的最大值:

$$P(w_1, w_2, \cdots, w_L | c_1, c_2, \cdots, c_L) \times P(c_1, c_2, \cdots, c_L) \tag{2}$$

对于一阶 HMM,即假设(1)当前词的词性只与它的前一个词的词性有关,(2)输出观察值(某个词)概率只与当前词性有关,则整个问题就可以进一步具体的利用下面公式来表示:

$$T^* = \arg\max_{c_1, c_2, \cdots, c_L} P(c_1)\, P(w_1 | c_1) \prod_{i=2}^{L} P(c_i | c_{i-1})\, P(w_i | c_i) \tag{3}$$

其中,$P(C_i | C_{i-1})$ 为 HMM 中的状态转移概率,$P(W_i | C_i)$ 为词汇发射概率。

三、生词处理

3.1 生词处理的问题描述

前面提到,生词处理仍然是困扰基于 HMM 统计词性标注方法的最主要问题之一。系统进行开放测试时,肯定会遇到许多生词,因此对生词的处理就成为标注能否成功的关键之一。

设句子中有生词 x_j,假设其在句子中的词性为 c_j,则(3)式可表示为:

$$T^* = \arg\max_{c_1, c_2, \cdots, c_L} P(c_1)\, P(w_1 | c_1) \cdots P(c_{j-1} | c_{j-2})\, P(w_{j-1} | c_{j-1})\, P(c_j | c_{j-1})\, P(x_j | c_j) \times \prod_{i=j+1}^{L} P(c_i | c_{i-1})\, P(w_i | c_i) \tag{4}$$

(4)式中由于 x_j 是个生词,我们不知道其词汇发射概率 $P(x_j | c_j)$,导致计算无法进行下去。因此确定 x_j 的词性问题,我们可以把它转化为确定 x_j 词汇发射概率 $P(x_j | c_j)$ 的问题。

3.2 生词词汇发射概率的估算模型

文献[5]尝试把生词的词汇发射概率赋值为 1,即令 $P(x_j | c_j) = 1$。这种方法的优点是易于实现、处理效率高,但毕竟没有利用任何统计先验知识和词法分析知识,准确率难以提高。还有的采用词形分析、规则和统计相结合等其他方法来确定生词的词汇发射概率[9,10]。这种

方法的特点是,如果规则设计得非常完善,则对生词处理的准确率能达到比较高的程度。但考虑到统计方法处理过程的统一性以及规则知识的难以获取,我们没有采用词形分析和规则方法,从而也免去了建立形态还原系统词典和消岐规则的困难。

假设有输入句子 $S = w_1, \cdots, w_{j-1} x_j w_{j+1} \cdots, w_N$,其中 S 表示整个句子,$w_i (1 \leq i \leq N)$ 表示单个的词,x_j 为生词。

现在面临的问题是如何来估算生词 x_j 的词汇发射概率?我们可以设想这样的情景:假设把 S 加入训练集中,由于加入的只有一个句子,因此对其他词的发射概率和整个模型的词性转移概率的影响可以忽略不计。遵循 HMM 模型的假设,x_j 的词性 c_j 由 w_{j-1} 的词性决定。即

$$P(c_j \mid x_j) \approx \sum_{m=1}^{M} P(c_m \mid w_{j-1})\ P(c_j \mid c_m) \tag{5}$$

其中 M 表示词性种类的总数。而根据贝叶斯公式,词汇发射概率

$$P(x_j \mid c_j) = \frac{P(x_j)}{P(c_j)}\ P(c_j \mid x_j) \tag{6}$$

将(5)式代入(6)得

$$P(x_j \mid c_j) = \frac{P(x_j)}{P(c_j)} \sum_{m=1}^{M} P(c_m \mid w_{j-1})\ P(c_j \mid c_m) \tag{7}$$

对式(7)中的各概率值采用极大似然估计,得

$$P(x_j \mid c_j) = \frac{C(x_j)}{C(c_j)} \sum_{m=1}^{M} P(c_m \mid w_{j-1})\ P(c_j \mid c_m) = \frac{1}{C(c_j)} \sum_{m=1}^{M} \left(\frac{C(w_{j-1} c_m)}{C(w_{j-1})} \times \frac{C(c_m c_j)}{C(c_m)} \right) \tag{8}$$

其中 $C(c_j)$ 表示标注符号 c_j 在训练语料中的出现次数,$C(c_m c_j)$ 表示标注符号串 $c_m c_j$ 的共现次数。

(8)式即为我们将要测试的生词词汇发射概率估算模型。

四、实验设计

4.1 语料

我们这次试验使用的是 PFR 人民日报标注语料库。该语料库使用了 26 个大类标注符号和 42 个细类标注符号。试验是这样设计的:PFR 语料库共计 110 多万词,在进行封闭测试时,首先分别以 10 万、20 万、30 万 ……110 万词的语料进行训练,从而建立相应的语料词典和模型参数。然后采取 Viterbi 算法对用来训练的所有语料重新进行词性标注,求出每一个句子的最佳词性标注序列,即完成了封闭测试。

在进行开放测试时,事先从 110 多万词的语料库中抽出大约 10 %的句子。这些句子不参与训练,而用于后面的测试,我们称之为测试集。同样的也是分别以 10 万、20 万、30 万 …… 100 万词语料进行训练,建立相应的语料词典和模型参数,然后采取 Viterbi 算法对测试集进行词性标注,从而完成了开放测试。

4.2 封闭测试

封闭测试结果如表 1 和表 2 所示,与其他文献报道的结果差不多。图 1 绘出了标注正确率与训练语料规模的关系曲线。

表 1 封闭测试结果(26 大类)

语料规模(万词)	10	20	30	40	50	60	70	80	90	100	110
正确率(%)	97.86	97.69	97.51	97.35	97.42	97.48	97.48	97.49	96.95	96.97	97.04

表 2　封闭测试结果(42 细类)

语料规模(万词)	10	20	30	40	50	60	70	80	90	100	110
正确率(%)	96. 39	96. 21	96. 11	95. 85	95. 79	95. 76	95. 74	95. 71	95. 68	95. 65	95. 64

4. 3　开放测试

开放测试时,肯定会遇到许多生词。对生词的处理算法是影响开放测试准确率的关键因素。对生词处理地比较好,则开放测试结果的准确率就会比较高。

开放测试结果如表 3 和表 4 所示,细类标注准确率达 92. 53 %,大类标注准确率达 94. 8 %,基本上达到了实用的程度。其中细类标注相对于大类标注其准确率不高,这主要是由于数据稀疏的原因(许多细类标注符在训练语料库出现的次数非常少,有几个细类标注符甚至只出现了一次)。因此,我们在做实用机器翻译系统时,对于词性标注集的确定,标注集的规模应该适当,分类不要过细,够用就行。比如我们在 IMT/ EC 系统[12]中只使用了 20 来个标注符。这样就能得到比较高的词性标注正确率,为机器翻译的后续处理过程提供有力保障。图 2 绘出了标注正确率与训练语料规模的关系曲线。

表 3　开放测试结果(26 大类)

语料规模(万词)	10	20	30	40	50	60	70	80	90	100
正确率(%)	88. 40	90. 98	92. 41	93. 31	93. 69	93. 92	94. 24	94. 47	94. 65	94. 86

表 4　开放测试结果(42 细类)

语料规模(万词)	10	20	30	40	50	60	70	80	90	100
正确率(%)	85. 66	88. 35	89. 82	90. 78	91. 18	91. 46	91. 76	92. 03	92. 29	92. 53

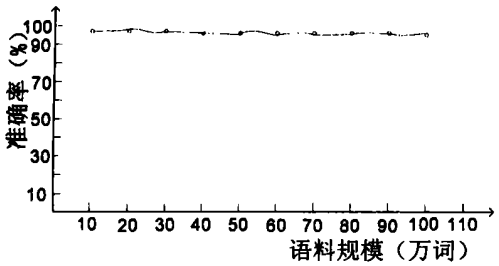

图 1　封闭测试时准确率与语料规模的关系

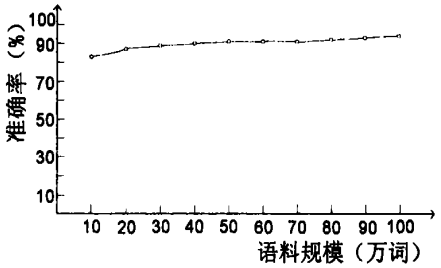

图 2　开放测试时准确率与语料规模的关系

4. 4　开放测试结果与其他方法的比较

为了考察本文提出的生词发射概率估算模型的有效性,开放测试时我们还跟其他同样基于 HMM 的方法作了比较。用来做比较的算法详见文献[5],该算法把生词的词汇发射概率赋值为 1,即令 $P(x_j|c_j)=1$,这里我们称之为 P - 1 算法。比较结果如表 5 和表 6 所示,结果显示本文算法的标注正确率平均高出近 1 %。虽然本文算法还远没有完全解决词性标注的全部问题,但测试结果说明本文提出的生词发射概率估算模型还是比较有效的。

表 5　本文算法与其他算法测试结果的比较(26 大类)

语料规模(万词)	50	60	70	80	90	100
P - 1 算法正确率(%)	92. 08	92. 35	92. 75	93. 29	93. 84	94. 12
本文算法正确率(%)	93. 69	93. 92	94. 24	94. 47	94. 65	94. 86

表 6　本文算法与其他算法测试结果的比较(42 细类)

语料规模(万词)	50	60	70	80	90	100
P - 1 算法正确率(%)	90. 71	90. 88	90. 91	91. 07	91. 51	92. 33
本文算法正确率(%)	91. 18	91. 46	91. 76	92. 03	92. 29	92. 53

五、结论

除了用到一个标注好的单语语料库外,没使用任何其他资源(比如词典库、规则库等),封闭测试正确率达 97 %左右,开放测试正确率也达到 95 %左右。同时与其他同样基于 HMM 的方法的测试结果比较,表明本文方法的标注正确率有较大的提高。

六、进一步的研究

我们现在所做的工作还没有充分利用我们现有的 IMT/ EC 系统的功能,尤其是 IMT/ EC 系统强大的规则分析能力[11]。下一步我们将把本文方法与 IMT/ EC 系统的规则分析方法结合起来,进行优势互补,希望能进一步提高词性标注的正确率。

参 考 文 献:

[1] Jelinek, F.. Self-organized language modeling for speech recognition. Readings in Speech Recognition [C], A. Waibel and K. F. Lee, eds., Morgan-Kaufmann, San Mateo, CA, 1990, 450 - 506.

[2] Miller, D., Leek, T., and Schwartz, R. M.. A hidden Markov model information retrieval system. Proc. 22nd International Conference on Research and Development in Information Retrieval[C], Berkeley, CA, 1999, 214 - 221.

[3] Zue, V. W.. Navigating the information superhighway using spoken language interfaces[R]. IEEE Expert, October, 1995,10 (5):39 - 43.

[4] L. E. Baum. An inequality and associated maximization technique in statistical estimation for probabilistic functions of a Markov process [J], Inequalities, 1972, 3: 1 - 8

[5] 赵铁军,等. 机器翻译原理[M],哈尔滨工业大学出版社,2000,6,141 - 143.

[6] 刘开瑛,等. 语料库词类自动标注算法研究,陈肇雄. 机器翻译研究进展[C]. 北京:电子工业出版社,1992,378 - 386.

[7] 黄昌宁,李娟子. 语料库语言学[M]. 北京:商务印书馆,2002,115 - 120.

[8] Ralph Weischedel, Marie Meteer, Richard Schwartz, Lance Ramshaw, Jeff Palmucci. Coping with Ambiguity and Unknown Words through Probabilistic Models [J]. Computational Linguistic, 1993, 19 (2) : 359 - 382.

[9] 周强,俞士汶. 一种切分和词性标注相融合的汉语语料库多级处理方法. 陈力为. 计算语言学研究与应用[C]. 北京:北京语言学院出版社,1993,126 - 131.

[10] 白栓虎. 基于统计的汉语语料库词性自动标注的研究与实现,黄昌宁,夏莹,语言信息处理专论[C]. 北京:清华大学出版社.

[11] 陈志忠,陈肇雄,高庆狮. 通用的自然语言词法分析机制[J]. 计算机学报,1991,2(2).