中 文 信 息 学 报
第 18 卷 第 2 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol.18 No.2

文章编号:1003-0077(2004)02-0030-06

汉语语料词性标注自动校对方法的研究*

钱揖丽,郑家恒

(山西大学 计算机科学系,山西 太原 030006)

摘要:兼类词的词类排歧是汉语语料词性标注中的难点问题,它严重影响语料的词性标注质量。针对这一难点问题,本文提出了一种兼类词词性标注的自动校对方法。它利用数据挖掘的方法从正确标注的训练语料中挖掘获取有效信息,自动生成兼类词词性校对规则,并应用获取的规则实现对机器初始标注语料的自动校对,从而提高语料中兼类词的词性标注质量。分别对 50 万汉语语料做封闭测试和开放测试,结果显示,校对后语料的兼类词词性标注正确率分别可提高 11.32%和 5.97%。

关键词:计算机应用;中文信息处理;兼类词;汉语词性标注;自动校对;粗糙集

中图分类号:TP391 **文献标识码**:A

Research on the Method of Automatic Correction of Chinese Part-of-Speech Tagging

QIAN Yi-li, ZHENG Jia-heng

(The Department of Computer Science, Shanxi University, Taiyuan, Shanxi 030006, China)

Abstract: The disambiguation of multi-category words is one of the difficulties in part-of-speech tagging of Chinese text, which affects the processing quality of corpora greatly. Aiming at this question, the paper describes an approach to correcting the part-of-speech tagging of multi-category words automatically. It acquires correction rules for the part-of-speech tagging of multi-category words from right-tagged corpora based on the rough sets and data mining, and then corrects the corpora based on these rules automatically. According to the results of close-test and open-test on the corpus of 500,000 Chinese characters, the accuracy of multi-category words' part-of-speech tagging can be increased by 11.32% and 5.97% respectively.

Key words: computer application; Chinese information processing; multi-category word; Chinese part-of-speech tagging; automatic correction; rough sets

1 引言

多年来,众多研究者们使用各种各样的方法,不懈地致力于改进词性标注算法。根据报道,各类词性标注软件的标注正确率均达到了 90%以上,但是其中兼类词的词性标注正确率却不到 90%。我们对 1998 年 1 月人民日报的 200 万语料进行了统计,结果显示其兼类词词性标注正确率在 84%左右,距实用还有一定的差距,仍有待于进一步的提高。所以,对机器标注结果进行校对是重要的、必不可少的一个环节。

* 收稿日期:2003-08-06

基金项目:国家 863 高技术研究发展计划资助(2001AA114031)

作者简介:钱揖丽(1977—),女,硕士,助教,主要研究领域为自然语言处理.

在对语料的机器标注结果进行人工校对的过程中，两点体会：①机器标注结果中的错误有“一错到底”的特性，即对于同样的情况，如果一个地方有错则通篇有错，且出错情况完全相同。所以校对者需要投入大量的时间去校对改正同样的错误，造成校对过程中所需的重复性劳动较多；②由于人为等因素，会造成在相同语境下，同一词的词性标注结果前后不一致的现象。

近年来，波兰科学家 Z. Pawlak 等提出了一个分析数据的数学理论——Rough Sets 理论，它为对不完整数据进行分析、推理、发现数据间的关系、提取有用属性、简化信息处理等提供了有力的工具。十几年来，在机器学习、知识发现、决策支持与分析、数据预测、专家系统、模式识别、语音识别、字符识别等众多人工智能领域得到了成功的应用，并越来越受到国际上的广泛关注[1]。

基于兼类词的词性与它所处的局部上下文环境有关，本文提出了一种兼类词词性校对方法。它利用数据挖掘的方法从训练语料中自动获取校对规则，并利用获取的规则对机器标注语料进行自动校对，从而提高兼类词的词性标注质量。分别对 50 万语料做封闭和开放测试，结果显示，校对后语料的兼类词词性标注正确率分别可提高 11.32% 和 5.97%。

2 词性校对决策表

2.1 构建词性校对决策表

定义 1：$S=(U,A)$为一知识表达系统，且 $C,D\subset A$ 是两个属性子集，分别称为条件属性和决策属性，具有条件属性和决策属性的知识表达系统可表示为决策表，记作 $T=(U,A,C,D)$或简称 CD 决策表[1]。

从训练语料中抽取兼类词的所有可能词性的真实范例生成范例库，并采集该兼类词被确定为某一可能词性的真实上下文语境信息，然后基于范例库建立词性校对决策表 $T=(U,A,C,D)$如表 1 所示。其中 $U=\{x_1,x_2,\cdots,x_n\}$为所有范例的集合，$x_i(i=1,2,\cdots,n)$表示每个范例；$A=\{a_{-5},a_{-4},\cdots a_{-1},a_0,a_1,a_2,\cdots a_5\}$为所有属性的集合，$a_j(j\in[-5,5])$分别表示某语境中的兼类词词性及其左右各 5 个词性；$C,D\subset A$ 是两个属性子集，$C=\{a_{-5},a_{-4},\cdots,a_{-1},a_1,a_2,\cdots,a_5\}$，即某真实语境中兼类词左右的各 5 个词性构成的条件属性子集；$D=\{a_0\}$，即由该兼类词在某真实语境中的标注词性构成的决策属性子集。

表 1 词性校对决策表

U/A	条件属性										决策属性
	a_{-5}	a_{-4}	a_{-3}	a_{-2}	a_{-1}	a_1	a_2	a_3	a_4	a_5	a_0
x_1	$a_{1,-5}$	$a_{1,-4}$	$a_{1,-3}$	$a_{1,-2}$	$a_{1,-1}$	$a_{1,1}$	$a_{1,2}$	$a_{1,3}$	$a_{1,4}$	$a_{1,5}$	$a_{1,0}$
…	…	…	…	…	…	…	…	…	…	…	…
x_n	$a_{n,-5}$	$a_{n,-4}$	$a_{n,-3}$	$a_{n,-2}$	$a_{n,-1}$	$a_{n,1}$	$a_{n,2}$	$a_{n,3}$	$a_{n,4}$	$a_{n,5}$	$a_{n,0}$

2.2 词性校对决策表的属性约简

如果直接利用词性校对决策表，采取直接匹配的方法校对，由于条件属性约束较多，待处理语料与之完全匹配的可能性较小，匹配度很低(50 万语料开放测试仅为 18.69%)，效果很不理想。所以我们利用粗糙集理论中决策表约简的方法，对条件属性部分进行约简，在不造成冲突错误的前提下，尽量减少条件属性的个数，可有效地提高匹配度，提高校对系统的性能。

定义 2：知识表达系统(决策表)T 的分明矩阵 $M(T)=[c_{ij}]_{n\times n}$是一个 $n\times n$ 矩阵，其中的矩阵项定义为：$c_{ij}=\{a\in A:a(x_i)\neq a(x_j),i,j=1,2,\cdots,n\}$[1]。

因此 c_{ij} 是个体 x_i 与 x_j 有区别的所有属性的集合，利用分明矩阵可以方便地求解属性集合 A 的约简。

我们也很容易看出相对于集合包含关系运算而言，若属性集合 $B\subseteq A$ 是满足条件：$B\cap c_{ij}\neq\Phi$，对于 $M(T)$ 中的任一非空项 $c_{ij}\neq\Phi$ 的一个最小属性子集，则称属性集合 $B\subseteq A$ 是 A 的一个约简。A 的所有约简的集合记为 $RED(A)$[1]。

由于 $M(T)$ 是对称的，且对于每个 $i=1,2,\cdots,n$，$C_{ij}=\Phi$，所以可以用 $M(T)$ 的下三角来表示 $M(T)$，$1\leqslant j<i\leqslant n$。

定义 3：对于每一个分明矩阵 $M(T)$ 对应惟一的分明函数 $f_{M(T)}$，它是一个有 m 元变量 $a_1^*,\cdots,a_m^*$（对应属性 $a_1,\cdots,a_m$）的布尔函数，$f_{M(T)}(a_1^*,\cdots,a_m^*)=\wedge\{\vee c_{ij}^*\mid 1\leqslant j<i\leqslant n, c_{ij}\neq\Phi\}$，$c_{ij}^*=\{a^*\mid a\in c_{ij}\}$[1]。

根据分明函数与约简的对应关系，信息系统 T 的约简 $RED(T)$ 的计算方法如下：

(1)计算信息系统 T 的分明矩阵 $M(T)$；

(2)计算与分明矩阵 $M(T)$ 对应的分明函数 $f_{M(T)}$；

(3)计算分明函数 $f_{M(T)}$ 的最小析取范式，其中每个析取分量对应一个约简。

属性约简算法：

设 $\Omega=a_{-5}\vee a_{-4}\vee\cdots a_{-1}\vee a_1\vee a_2\vee\cdots\vee a_5$，对于 T 做：

(1)for $i=1$ to n do

for $j=1$ to $i-1$ do

if $a_0(x_i)\neq a_0(x_j)$ then

{$c_{ij}=\Phi$;

for $k=-5$ to 5 do

if (k〈〉0) and ($a_k(x_i)\neq a_k(x_j)$) then $c_{ij}=c_{ij}\vee a_k$;

}

else $c_{ij}=\Omega$;

(2)for $i=1$ to n do

{$a_i=\Omega$;

for $j=1$ to $i-1$ do $a_i=a_i\wedge c_{ij}$;

};

(3)将 $a_1\wedge a_2\wedge\cdots a_n$ 转换成分明析取范式，其中每一析取分量即为 T 的一个约简；

(4)选择一种约简方法，对 T 进行条件属性的约简，形成新的词性校对决策表。

3 词性校对规则集

3.1 规则一致化

决策表属性约简过程中，把以条件属性和决策属性形式描述的 T 公式化为 $\theta\rightarrow\Psi$ 形式的决策规则集（即校对转换规则集），θ 和 Ψ 分别称为决策规则的前驱和后继，对应条件属性和决策属性，类似于前面我们所说明的用条件属性和决策属性描述研究对象，它们表达一种因果关系。

对于任何决策规则 $\theta_1\rightarrow\Psi_1$，如果 $\theta=\theta_1$ 蕴涵 $\Psi=\Psi_1$，我们说该决策规则是 T 中一致的；否则，就说该决策规则是 T 中不一致的[1]。显然，对于 T 中不一致的规则，相同的前驱蕴涵不相同的后继，这种规则间的冲突将导致错误，所以我们要剔除 T 中不一致的规则，将规则集 T 一致化。

3.2 规则集优化

经过条件属性约简和一致化后的形成的校对规则，条件约束大大减少，语料与规则的匹配度也显著提高，开放测试可达 57.07%，比约简前提高近 40 个百分点，但仍然有大量的语境无法找到与其完全匹配的校对规则。为了进一步提高校对正确率，本文对规则集进行了进一步优化，优化的依据就是规则间的相似度。

规则是由许多属性组成，规则间的相似度就是根据属性之间的相似度定义的[1]。规则的相似度也常常是通过距离来定义的。

定义 4：基于统计思想和欧氏距离，定义规则间的距离：

$$d(x_i,x_j)=\sqrt{\frac{1}{n}\sum_{k=1}^{n}(f_k(x_i)-f_k(x_j))^2}.$$

其中，$n=10$ 表示属性的总数；$f_k(x_i)$ 表示第 i 条规则 x_i 的第 k 个属性值（即词性）出现的频率，即某个词性标记出现在第 k 个属性位置的次数与出现总次数的比率。

定义 5：基于规则间的距离，定义规则间的相似度：

$$\mathrm{SIM}(x_i,x_j)=1-d(x_i,x_j),(d(x_i,x_j)\in[0,1])^{[1]}$$

选取合适的阈值对规则集中满足条件的规则进行归并，减少规则的数目，从而对规则集进行简化和优化。通过实验，选定阈值为 0.87。

使用规则集中的规则对语料库进行词性自动校对的同时，需要将规则的使用情况反馈到规则集，对规则的使用情况做出评价，从而为规则集的不断动态更新提供依据。为了记录、反应和评价规则的使用情况，本文做出以下定义：

定义 6：定义规则的可信度：$f_e=f_r/f_w$。其中，f_r 表示规则使用正确的次数，f_w 表示规则被使用的总次数。

随着校对语料规模的不断扩大，语料数量及校对次数的不断增加，每一规则的可信度信息越来越丰富和客观，我们就可以根据规则集中每一规则的可信度信息，有选择的对规则集中可信度过低的规则进行删除，进一步优化规则集。

通过规则集的优化过程，我们不仅简化了规则的表示形式，同时表示形式的简化以及相似度计算的引入，也使得校对时规则与语料的匹配度大大提高，从而提高了系统的性能。

4 词性自动校对

利用自动获取生成的词性校对规则，按照以下步骤对机器自动标注语料进行处理：

1. 从待处理的初标语料中抽取所有的兼类词及其上下文语境信息，建立待处理语料的兼类词表及范例表；

2. 对于词表中有相应校对规则的兼类词，逐一进行以下处理：

(1)从语料中抽取包含该词的一个句子，获取相关的上下文信息，即条件属性部分；

(2)搜索规则集，查找匹配规则完成校对，并同时对规则使用情况做出评价：

a. 若有匹配规则，则比较语料中实际标注的词性与依据规则推出的词性（即决策属性 a_0）是否相同，若相同则认为规则使用正确；若不同，则修改；

b. 若没有匹配的规则，则计算当前语境与每一规则间的相似度，依据 MAX(SIM(x_i, x_j))对应的规则完成校对；

3. 对于新出现的频次比小于 10:1 的兼类词，将其实例另外保存，经过人工校对后，按照获取规则的方法，获取该词的校对规则加入到规则集，对规则集进行动态更新。

5 实验结果及分析

从训练语料中选取50万作为封闭测试语料集,另外选取50万语料作为开放测试语料集。

5.1 上下文窗口长度的选择

迄今为止,在中文信息处理领域的许多研究中,考虑上下文信息时,大都是在长度为2～5的窗口上进行的。我们对2～5的上下文窗口长度作了测试,结果如表2和表3所示。

表2 不同窗口长度封闭测试结果

	原正确数	原正确率	校对后正确数	校对后正确率	提高率
长度为2	168	84.85%	168	84.85%	0.00%
长度为3	168	84.85%	173	87.37%	2.52%
长度为4	168	84.85%	179	90.40%	5.55%
长度为5	168	84.85%	185	93.43%	8.58%

表3 不同窗口长度开放测试结果

	原正确数	原正确率	校对后正确数	校对后正确率	提高率
长度为2	167	84.34%	158	79.80%	-4.54%
长度为3	167	84.34%	169	85.35%	1.01%
长度为4	167	84.34%	175	88.38%	4.04%
长度为5	167	84.34%	179	90.40%	6.06%

可以看出,随着窗口的增大,实验结果越来越好,所以我们选择上下文窗口长度为5。

5.2 实验方法的选择

以兼类词"要求"为例,采用三种不同的方法分别做封闭和开放测试:

方法1:直接利用自动获取生成的校对决策表(未做约简)校对;

方法2:依据属性约简算法约简校对决策表,利用约简后生成的规则集自动校对;

方法3:选定阈值为0.87,对初始规则集进行优化,并引入相似度计算自动校对。

表4 规则集规模对比

	规则数量(条)
方法1	481
方法2	329
方法3	124

表5 封闭测试结果对比

	原正确数	原正确率	校对后正确数	校对后正确率	提高率
方法1	168	84.85%	188	94.95%	10.10%
方法2	168	84.85%	187	94.44%	9.59%
方法3	168	84.85%	185	93.43%	8.58%

表6 开放测试结果对比

	原正确数	原正确率	校对后正确数	校对后正确率	提高率
方法1	167	84.34%	170	85.86%	1.52%
方法2	167	84.34%	174	87.88%	3.54%
方法3	167	84.34%	179	90.40%	6.06%

从表4中很容易看到,引入属性约简和相似度计算后,规则数量由481条降到124条。由表5知,随着规则数量的大大减少,虽然封闭测试的提高率有所降低,但是降低的幅度不大。观察表6,规则数量的大大减少反而使得开放测试的改对数量大增,且标注正确率大大提高,这个结果是非常振奋人心的。所以我们采用方法3作为最终的校对方法。

5.3 实验结果及分析

50万封闭测试语料中共有兼类词1480个,出现82464词次;50万开放测试语料中共有兼类词1292个,出现83269词次。表7给出测试结果。

表 7　实验测试结果对比

	总词次	原正确词次	原正确率	校对后正确词次	校对后正确率	提高率
封闭测试	82464	69055	83.74%	78390	95.06%	11.32%
开放测试	83269	70212	84.32%	75183	90.29%	5.97%

实验结果显示，经过校对后，封闭测试语料和开放测试语料的兼类词词性标注正确率分别提高 11.32%和 5.97%。影响和制约正确率进一步提高的因素主要有以下几点：

(1)对冲突规则的简单剔除会使得一部分语言环境信息丢失，从而使得一部分潜在的规则丢失。

(2)从训练语料获取规则时，对于各个词性标记的频次比大于 10:1 的兼类词，我们认为其各个词性所占的比重过于悬殊，没有予以考虑，所以对于这部分兼类词，没有相应的校对规则，不做校对处理。

(3)采用粗糙集理论的约简方法获取校对规则时，只考虑了上下文范围内的词性，而未涉及上下文的具体词，因此，词对于兼类词的词性影响的大小是多少，还需要进一步深入研究。

采用机器自动校对的方法提高词性标注的正确率，这是对语料加工质量保证的有益尝试。今后我们将围绕这个问题，做进一步的研究和探讨，改进和完善校对方法，期待能够取得更好的结果。

参 考 文 献：

[1] 史忠植．知识发现[M]．北京：清华大学出版社，2002.

[2] ZDZISLAW PAWLAK. Rough Sets-Theoretical Aspects of Reasoning about Data[M]. Kluwer Academic Publisher, 1991.

[3] Eric Brill. Unsupervised Learning of Disambiguation Rules for Part of Speech Tagging[A]. In: Yarowsky D. Churchk. Proceeding of 3rd Workshop on Very Large Corpus[C]. Cambridge, Massachusetts, USA, 1995, 1－13.

[4] 李晓黎，史忠植．用数据采掘方法获取汉语词性标注规则[J]．计算机研究与发展. 2000, 37(12): 1409－1414.

[5] 朱靖波，张玥杰，姚天顺．一种短语结构规则的自动获取方法[J]．计算机研究与发展. 1999, 36(5): 601－607.