中 文 信 息 学 报
第 18 卷 第 5 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol.18 No.5

文章编号:1003 - 0077(2004)05 - 0011 - 06

语料库词性标注一致性检查方法研究

张 虎,郑家恒,刘 江

(山西大学 计算机科学系,山西 太原 030006)

摘要:在对大规模语料库进行深加工时,保证词性标注的一致性已成为建设高质量语料库的首要问题。本文提出了基于聚类和分类的语料库词性标注一致性检查的新方法,该方法避开了以前一贯采用的规则或统计的方法,利用聚类和分类的思想,对范例进行聚类并求出阈值,对测试数据分类来确定其标注的正误,进而得出每篇文章的词性标注一致性情况,进一步保证大规模语料库标注的正确性。

关键词:计算机应用;中文信息处理;词性标注一致性;兼类词;聚类

中图分类号:TP391 **文献标识码**:A

The Inspecting Method Study on Consistence of Part of Speech Tagging of Corpus

ZHANG Hu, ZHENG Jia-heng, LIU Jiang

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

Abstract: In the deep processing of large-scale corpus, it has been a chief problem to assure the consistence of part of speech tagging to build the high-quantity corpus. A new inspecting method on consistence of part of speech tagging based on clustering and classifying is put forward, firstly we cluster the sequences of part of speech of the example and get the threshold value, then classify the test sequences to judge its' correctness, furthermore, we can know the condition of the consistence of part of speech on every text, and assure the correctness of the part of speech tagging on large-scale corpus further.

Key words: computer application; Chinese information processing; the consistence of part of speech tagging; conversion of parts of speech; clustering

1 引言

随着语料库语言学研究的兴起,建设高质量的大规模语料库已成为首要任务。语料库作为研究资源其价值是通过对语料的标注来体现的。目前对汉语语料库的标注包括词语切分、词性标注、句法标注等。对语料库标注的越准确,语料库的价值就越高。

近年来国内外对汉语的词性自动标注的研究有很多,已取得很大进展。它们大多是采用基于规则和基于统计的方法,标注正确率分别达到 89% 和 96%,对错误标注结果进行分析,可以看出,无论哪种标注算法都有其固有缺陷,概率标注方法总会抑制小概率事件的发生,而规则方法本质上说是一种确定性的演绎推理方法,因此他们很难对词性标注的准确率进行进一步的提高。显然,这样的准确率仍然严重影响语料库的加工质量。

在对语料库深加工的过程中,词性标注的一致性是一个难点。所谓的词性标注的一致性

收稿日期:2004 - 03 - 04

基金项目:国家"863"高技术研究发展计划资助项目(2001AA4031)

作者简介:张虎(1979—),男,硕士研究生,主要研究领域为中文信息处理.

是指在相同的语境下对同一个词标注相同的词性。要保证词性标注的一致性就要尽量避免词性标注上的不一致。词性标注不一致产生的原因主要有下面三个:①各种词性自动标注算法都有这样那样的不足,不可能有百分之百正确的标注结果;②自动标注好的语料要经过人工校对,疏忽和错误在所难免,且不同校对者认识上有差异,导致了不一致现象的出现;③《信息处理用现代汉语词类及标记集》对某些语言现象规定不够明确,同样会造成词性标注前后不一致。因此如何解决这种不一致现象已成为建设高质量语料库的首要问题。

目前国内外对汉语语料库词性标注结果的校对,还停留在人工校对上,对词性标注结果不一致现象还没有进行系统的研究。本文提出了基于聚类和分类的词性标注一致性检查的新方法。该方法首先随机抽出一些含有兼类词的句子,经人工校对后,将含有相同兼类词的词性序列进行聚类并求出阈值,然后对每个含有兼类词的词性序列进行逐一分类,通过计算相似度找出标有该类的词性但相似度不在该类的阈值范围之内的序列,找出的这些序列就被认为是词性标注不一致的。我们通过对含"高"字的1042个句子和50万语料分别进行实验,它们的有效性约为76%,它们的准确率约为83%。

2 词性一致性现象分析

语料库中已标注词性的词语有两种标注结果:单标记词语和多标记词语。所谓单标记词语,即在语料库中只有一种词性标记的词语;多标记词语则为在语料库中标有两种或两种以上词性的词语。我们所做的不一致性检查是针对多标记词语而言的。经分析,这些多标记词语包括两种情况:

Ⅰ.词表中是单标记词语,但在语料中标了不同的词性,出现了不一致。

词条:财税

语料中的多标记词性:j、n

错例:并/c 结合/v 整个/b 财税/n 体制/n 改革/v

校正词性:j

分析:"财税"这一词比较有争议,在这我们认为它是非兼类词,只有缩略语j这唯一词性。

Ⅱ.词表中是兼类词,即有不同标记的词语,这部分词语可能存在词性标注错误,即:在相同的语境中出现了不同的词性。

词条:高

语料中的多标记词性:a、ad、an、d、j、Ng、nr、v

错例:比/p 以往/t 高/a 出/v 许多/m 。/w

校正词性:v

分析:"高"有八种词性,按照我们的标注规范,它在例句中的词性应为动词v。

对于第一种情况的不一致,如果词表中的词比较全的话,这种不一致比较容易解决,只要把标注不一致的词条与词表比较后,改正即可。

我们对北京大学网上公布的200万汉字语料进行了统计,在200万汉字的语料库中兼类词占到11%,但兼类词的词次却占到了47%。面对大规模语料库,如何判断在相同的语言环境下兼类词出现的错标现象就成为一个重要的问题。如:"高"

句①:比/p 获得/v 亚军/n 的/u 德国/ns 选手/n 高/a 出/v 18/m 分/q 多/m 。/w

句②:比/p 其他/r 农户/n 高/v 出/v 1000/m 多/m 元/q 。/w

显然上边两句话语言环境相同,但在语料中句①标了 a,句②标为 v。

3 词性一致性检查模型

兼类词词性标注是否一致是按照其语境来判断的,所以我们以每个兼类词及其上下文语境所形成词性标记序列作为研究对象。首先对范例进行聚类并求出阈值,然后根据聚类结果对测试语料进行词性标注的一致性检查。

3.1 直接模型

为了描述兼类词的语境,我们建立含有兼类词的词性标记序列表:

表 1 词性标记序列表

词	前三词	前两词	前一词	兼类词	后一词	后两词	后三词
词性标注	词性 1	词性 2	词性 3	词性 4	词性 5	词性 6	词性 7

注:其中"前(后)几词"指从所要考查的兼类词数起前(后)边的第几个词。

定义 1:位置属性

兼类词词性标记序列的前、后词的词性由于离兼类词的距离不同,所以对包含兼类词的词性序列影响程度也不同,称之为位置属性。用向量 $X=\{(1/22),(1/11),(2/11),(4/11),(2/11),(1/11),(1/22)\}$ 表示。

1/22: 前 (后) 第三个词的位置属性值。 1/11: 前 (后)第二个词的位置属性值。

2/11: 前 (后) 第一个词的位置属性值。 4/11: 兼类词的位置属性值。

位置属性值表明该位置的词性对该词性标记序列影响程度的数值化表示,所有位置属性值之和为 1。如:1/22 表明该位置对词性标记序列的影响程度是 1/22。

直接模型的算法只考虑词性标记序列的位置属性。如果两个序列对应位置的词性标记相同则 Y 取 1,否则取 0。即:$Y=\begin{cases}1, if(CXni=CXnj)\\0, if(CXni<>CXnj)\end{cases}$

$n=1\cdots7$ i,j 分别代表各个句子

直接模型词性标记序列的相似度公式:

$$S_{i,j}=\sum_{n=1}^{7}Y(ni,nj)X(n) \tag{1}$$

n 代表词性标记序列中的各个位置;$X(n)$代表各个位置的位置属性值

例如:"高"

a. 缀/v 满/a 彩灯/n 的/u 高/a 塔/n 直/d 插/v 夜空/n ,/w

b. 这/r 是/v 一/m 项/q 高/a 科技/n 的/u 硬仗/n ,/w

对例句可以建立如下的词性标记序列:

a.(a n u a n d v)

b.(v m q a n u n)

a 与 b 中有两个位置的词性标记是相同的,按照公式(1)求出 a 与 b 两个词性标记序列的相似度是 6/11。

3.2 向量模型

直接模型的算法只考虑了词性标记序列的一个属性即位置属性,因此可以进一步把词性的属性考虑进去。对每个含兼类词的词性标记序列进行向量化表示,然后求出任何两个向量之间的相似度。

定义 2:词性属性

兼类词词性标记序列前、后词的词性和词性标记的位置,对确定兼类词的词性影响程度不同,称之为词性属性。用一个 7 行 m 列的二维矩阵来描述。其中:行表示兼类词词性标记序列前、后三个词及兼类词本身;列表示语料库所采用的词性标记集的标记。

例如:"高" 缀/v 满/a 彩灯/n 的/u 高/a 塔/n 直/d 插/v 夜空/n ,/w

词性标记序列是:(a n u a n d v)

设:词性标记集为:{n v a d u p r m q c w l f s t b z e o l j h k g y}

"高"的词性属性矩阵:

$$\boldsymbol{Y}=\begin{pmatrix}0,0,1,0,0\cdots\cdots\\1,0,0,0,0\cdots\cdots\\0,0,0,0,1\cdots\cdots\\0,0,1,0,0\cdots\cdots\\1,0,0,0,0\cdots\cdots\\0,0,0,1,0\cdots\cdots\\0,1,0,0,0\cdots\cdots\end{pmatrix}$$

定义 3:词性标记序列向量

位置属性向量与词性属性矩阵的乘积定义为词性标记序列向量。即:$Vec = X \times \boldsymbol{Y}$

例句"缀/v 满/a 彩灯/n 的/u 高/a 塔/n 直/d 插/v 夜空/n ,/w"的词性标记序列向量如下:

Vec = (1/22, 1/11, 2/11, 4/11, 2/11, 1/11, 1/22) × $\boldsymbol{Y}$

= (3/11, 1/22, 9/22, 1/11, 2/11, 0, 0……)

采用马氏距离计算方法定义向量模型词性标记序列的相似度公式:

$$S_{i,j} = (x_i, y_i)' V^{-1}(x_i, y_i) \tag{2}$$

其中:x_i 和 y_i 是两个任意的词性标记序列向量

$$V = \frac{1}{m-1}\sum_{i=1}^{m}(x_i - \bar{x})(x_i - \bar{x})'$$

例如:"高"

a. 缀/v 满/a 彩灯/n 的/u 高/a 塔/n 直/d 插/v 夜空/n ,/w

b. 这/r 是/v 一/m 项/q 高/a 科技/n 的/u 硬仗/n ,/w

对例句可以生成下边特征向量:

a.(3/11, 1/22, 9/22, 1/11, 2/11, 0, 0……)

b.(5/22, 1/11, 4/11, 0, 1/11, 0, 0, 1/11, 2/11, 0……)

按照公式(2)可以求出上边两个词性标记序列的相似度约是 0.236。

3.3 聚类

聚类是把某些对象按其相似性加以分组的一种数据划分。它是通过较为少数的聚类簇去表现大量的数据,每个聚类簇都有自己的特征。本文采用的是基于重心的聚类方法。

词性标记序列向量集合中任一向量 x_i 与重心向量 x_j 间的距离 d_{ij} 满足:

$$\frac{1}{k-1}\sum d_{ij} \leqslant H \tag{3}$$

称集合对于 H 组成一类。

其中：k 为集合中元素个数，H 为阈值。

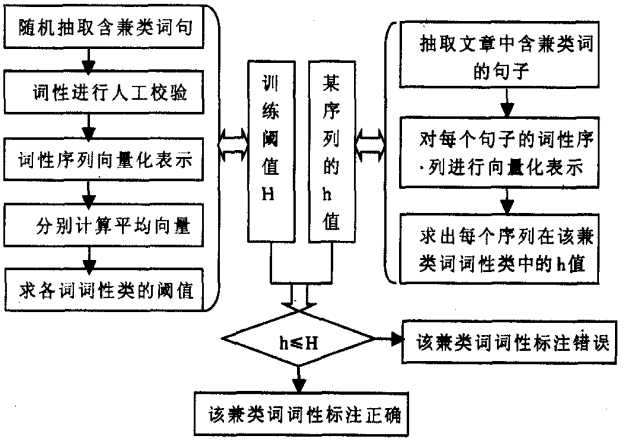
H 值是通过训练范例求得的，具体部骤如下：

step1：随机选取一些含有兼类词的句子，进行人工校对，对每个兼类词的所有词性分别计算含有同一兼类词且其词性标注相同的所有词性标记序列的向量的平均值 VA，根据公式(2)计算所有该类的词性标记序列向量与 VA 的马氏距离，计算该类中所有马氏距离的平均值。

Step2：step1 中求出的每一兼类词词性类的平均值作为该兼类词词性类的 H。

4 测试及分析

4.1 词性一致性检查的流程



注：该图的说明是以某一词的某个词性作为对象。

4.2 阈值的选取

我们从 200 万字语料库中，选取了含“高”字的 1045 个句子，作部分测试语料。对语料中词语的词性只标注大类，词性标记集为：{n v a d u p r m q c w I f s t b z e o l j h k g y}。但是“高”的词性标记 d、j、g，在 200 万语料中只出现了一次，为了验证基于聚类的词性一致性检查算法的有效性，我们不考虑“高”的“d、j、g”三个词性标记。

表 2 “高”字的统计结果和阈值表

词条	高		
兼类数目	6		
总词数	1045		
词性	a	n	v
词性词数	910	112	20
词性所占比例(%)	87.08	10.71	1.91
词性的阈值	0.212	0.103	0.259

统计结果和阈值选取如表 2。

对 50 万测试语料计算阈值方法和上例相同。

4.3 测试

1)两种模型的测试数据

对含“高”字的 1042 句子和 50 万语料分别进行测试，结果见表 3。

设总词次数为 M，存在错标的兼类词次数为 m，查到的不一致词次数为 N，包含的查错数为 n。

模型的有效性 $=(N-n)/m$；模型的无效性 $=n/(M-m)$。

2)结果分析

200 万语料中含“高”字的 1042 个句子和随机抽出的 50 万语料的测试结果显示：

表3 实验结果表

词 条	词性	实验结果					
		总词次数	存在错标的兼类词次数	查到的不一致词次数(查错数)		模型的有效性(无效性)(%)	
				直 接	向 量	直 接	向 量
高	a	910	8	5(1)	7(1)	50.00(0.11)	75.00(0.11)
	n	112	0	0(0)	0(0)	100.00(0.0)	100.00(0.0)
	v	0	5	3(1)	5(1)	40.00(6.67)	80.00(6.67)
总 数		1042	13	8(2)	12(2)	46.15(0.19)	76.92(0.19)
50万语料	所有	64040	562	343(85)	518(91)	45.91(0.13)	75.98(0.14)

①向量模型和直接模型查到的正确的不一致数相差较大,它们的有效性分别约为76%和46%,显然向量模型比直接模型更有效。

②应用向量模型进行词性一致性检查时,有一些错标的未查到,未查到数占存在错标的兼类词次数的比例约为24%。向量模型检查的召回率约为76%。

③利用向量模型进行一致性检查时,在查到的不一致词中存在一些查错的。向量模型检查的准确率约为83%。

④无论是直接模型还是向量模型,它们的无效性都很低,都底于0.20%。

我们提出的基于聚类和分类的语料库词性标注一致性检查的新方法,通过对含"高"字的1042个句子和50万语料分别进行实验,从纵向和横向对200万语料作了代表性测试,它们的有效性约为76%,它们的准确率约为83%,而其无效性很低,不到0.20%。结果表明这种解决词性标注一致性的方法是有效的,它避开了以前一贯采用的规则或统计的方法,利用了聚类和分类的思想,对范例进行聚类并求出阈值,对测试数据进行分类确定其标注的正误,这在一定程度上解决了机器自动标注中容易出现而且较难解决的一致性问题,而且也大大的提高了人工校对的效率和质量。今后,我们将进一步改进词性标注一致性检查的模型,并努力建立词性标注不一致的自动校正模型,为建设高质量语料库提供更好的方法。

参 考 文 献:

[1] 俞士汶,等. 大规模现代汉语标注语料库的加工规范[J].中文信息学报,2000,14(6):58-64.

[2] 孙即祥,等.现代模式识别[M].国防科技大学出版社,2002.

[3] 刘开瑛.中文文本自动分词和标注[M].北京:商务印书馆,2000.

[4] 张仰森,丁冰青.中文文本自动校对技术现状及展望[J].中文信息学报,1998,12(3):50-56.

[5] 钱卫宁,周傲英.Analyzing Popular Clustering Algorithms from Different Viewpoints[J].软件学报,2002,1382-1394.

[6] 冯志伟.计算语言学基础[M].北京:商务印书馆,2001.

[7] 史忠植.知识发现[M].北京:清华大学出版社,2002.

[8] 齐璇,王挺,陈火旺.义类自动标注方法的研究[J].中文信息学报,2001,15(3):9-15.

[9] 游荣彦,等.向量空间模型中特征词的区分度的定量研究[J].中文信息学报,2002,16(3):15-19.

[10] 贺宏朝,等.一种基于上下文的中文信息检索查询扩展[J].中文信息学报,2002,16(6):32-37.

语料库词性标注一致性检查方法研究

作者： 张虎， 郑家恒， 刘江
作者单位： 山西大学,计算机科学系,山西,太原,030006
刊名： 中文信息学报 ISTIC PKU
英文刊名： JOURNAL OF CHINESE INFORMATION PROCESSING
年，卷(期)： 2004,18(5)
被引用次数： 4次

参考文献(10条)

1. 俞士汶 大规模现代汉语标注语料库的加工规范[期刊论文]-中文信息学报 2000(06)

2. 贺宏朝 一种基于上下文的中文信息检索查询扩展[期刊论文]-中文信息学报 2002(06)

3. 游荣彦 向量空间模型中特征词的区分度的定量研究[期刊论文]-中文信息学报 2002(03)

4. 齐璇;王挺;陈火旺 义类自动标注方法的研究[期刊论文]-中文信息学报 2001(03)

5. 史忠植 知识发现 2002

6. 冯志伟 计算语言学基础 2001

7. 钱卫宁;周傲英 Analyzing Popular Clustering Algorithms from Different Viewpoints 2002

8. 张仰森;丁冰青 中文文本自动校对技术现状及展望 1998(03)

9. 刘开瑛 中文文本自动分词和标注 2000

10. 孙即祥 现代模式识别 2002

引证文献(4条)

1. 张虎.郑家恒 基于分类的汉语语料库词性标注一致性检查[期刊论文]-计算机工程 2008(8)

2. 周文.徐国梁 翻译记忆中语句相似度计算方法的研究[期刊论文]-计算机应用 2007(5)

3. 牛洪梅.加米拉·吾守尔.吐尔根·依布拉音 现代维吾尔语的词性标注校对技术研究[期刊论文]-伊犁师范学院学报（自然科学版） 2007(1)

4. 牛洪梅.吐尔根.伊不拉音 维吾尔语的词性标注校对初探[期刊论文]-微型电脑应用 2006(12)