文章编号: 1671-7872(2008)04-0426-04

一种基于规则优先级的词性标注方法

王广正, 王喜凤
(安徽工业大学 计算机学院, 安徽 马鞍山 243002)

摘要: 词性标注作为汉语自动分词以至中文信息处理领域比较关键的问题之一, 是该领域的研究难点也是研究重点, 对兼类词词性标注的正确率严重影响着词性标注的质量。在基于规则的词性标注的基础上, 提出了一种基于规则优先级的词性标注方法, 即对每条词性标注规则加上优先级, 并在标注算法中通过对优先级进行控制来完成兼类词的词性标注。并用大规模语料对该方法做了试验, 结果表明其词性标注正确率可达到96.4%。

关键词: 汉语自动分词; 词性标注; 兼类词; 规则优先级
中图分类号: TP391 文献标识码: A

A Method of POS Tagging Based on Priority of Rules

WANG Guang-zheng, WANG Xi-feng
(School of Computer Science, Anhui University of Technology Ma′anshan 243002, China)

Abstract: As one of the important problem in the field of Chinese automatic word segmentation and Chinese information process, part of speech tagging is the research difficulty and emphases in the filed. The precision of POS tagging to syntactic category has influenced badly the quality of POS tagging. On the basis of POS tagging based on rules, proposed a method of POS tagging based on PRI of rules, that is adding PRI to each rule. Through controlling the PRI in the tagging algorithm, POS tagging to syntactic categories can be completed. Lots of examples are used to test the method. The result shows that the precision of POS tagging is 96.4%.

Key words: Chinese automatic word segmentation; POS tagging; syntactic category; priority of rules

在中文信息处理领域, 词性的自动标注是一项很重要的研究内容。词性标注就是把已经切分好的输入文本对每个词标记上该词在该上下文环境中合适的词性。但汉语中很多词的词性不只一个, 也即存在词的兼类现象, 给汉语的词性标注增加了很大的困难。兼类词现象在英语、汉语中都是大量存在的。据统计, 虽然兼类词的绝对数目在整个词汇的总数中占的比例不是很大(约4.28%), 但是它们出现的频率极高(约占总词次数的36.8%)[1], 这说明在词性的自动标注时, 超过1/3的词次数要考虑到词的兼类问题。

目前, 国外对英语词性的自动标注研究取得很大的进展, 并且有的系统(CLAWS)标注正确率达到了96%的水平。国内对词性的自动标注研究较晚, 近些年, 清华大学、山西大学、北大计算语言所等研究单位都对该领域做了大量的研究, 并且也取得了很好的效果。清华大学、山西大学对基于统计的词性自动标注做了一些研究, 并提出了用于汉语词性标注的词类标记集, 标注正确率可达95%左右; 北大计算语言所利用了规则和统计相结合的词性标注方法, 充分发挥二者的优势, 通过对约40万字的语料库的处理结果显示, 该方法的标注正确率达到了96%左右[1]。目前对兼类词进行词性标注的方法主要有3种: 基于规则的方法、基于统计的方法[2,3]、规则和统计相结合的方法[1,4]。基于规则的方法是一种传统方法, 能充分利用现有语言学成果, 能通过现有的语言学规律中总结出许多有用的规则, 并能对大多数的语言组合作细致的描述, 但是如果规则描述过细, 词性标注的正确率可提高, 但是规则的覆盖面就会大大减小; 如果使覆盖面增大, 必然要以降低正确率为代价。基于统计的方法是目前应用较多的方法, 其是通过对大规模语料库进行训练而得到, 因此覆盖面很广,

收稿日期: 2008-01-22
作者简介: 王广正(1980-), 男, 山东郓城人, 硕士, 助教, 研究方向为自然语言处理, 图形图像处理。

而这种方法只是取大概率事件,并没有考虑小概率的特殊事件,必然会降低标注的正确率。鉴于以上两种方法的缺陷,把它们结合起来,利用规则标注方法的同时,通过大规模语料库的训练对其中规则的等级进行区分,引入了优先级的概念,在此基础上,提出了基于规则优先级的词性标注算法,并实现了一个词性自动标注系统,通过对近 5 万字的语料进行了标注,结果验证了该算法的有效性。

1 基于规则优先级的词性标注

1.1 规则优先级的引入

在标注的过程中,碰到这样的现象:一个兼类词,它既可以与前驱词搭配,也可以与后续词搭配,此时怎么处理?假设分词系统输入词的序列为:$W=W_1W_2W_3\cdots W_j\cdots W_{m-1}W_m$,其中 W_j 输入的第 j 个词,如果 $W_{j-1}W_j$ $1<j\leqslant m$) 满足规则库中的某条规则 R1,而 W_jW_{j+1}($1\leqslant j<m$) 则满足另一条规则 R2,而 R1 与 R2 有可能相同,也可能不同;另一种情况:如果 $W_{j-1}W_j$ $1<j<m$) 既满足 R1,又满足 R2,R1 与 R2 有可能相同,也可能不同。鉴于上述两种情况的考虑,引进了规则优先级的概念:对规则库中的各条规则加上优先级,0 表示优先级最高,1 表示优先级次之,依次类推。在设置优先级时:(1) 利用统计的方法找出搭配频率出现较高的规则且赋予高优先级;(2) 避免不经常出现的搭配规则一直不被采用。对于(2) 中不常出现的规则,也适当赋予较高的优先级,如:形容词 + 数词(ADJ+NUM),赋予该规则的优先级为 4。例如:“公司的人员在近两年内增加了”中的“近两年”的切分结果为“近 / 两 / 年”,根据“两”和“年”很容易标注为数词和名词,但是“近”的词性怎么标注呢?因为在词性表中,“近”有两个词性:形容词和副词。根据上下文,可以知道此处应为形容词。并且在词性表中,“两”有 3 个词性:形容词,量词,数词。虽然形容词 + 形容词这条规则在大规模语料库中出现的频率要高于形容词 + 数词,但是,形容词后面如果跟像“两”之类的既可为形容词也可为数词的兼类词,文中的规则是把它标记为数词。通过这样的标准来构造规则库,既考虑到了搭配频率较高的规则,又照顾了不常出现的搭配规则。

1.2 词性搭配规则

词的搭配规则主要是在词性的层面上考虑。比如常见的词性搭配有:形容词 + 名词(adj+n)、动词 + 名词(v+n)、量词 + 名词(m+n)等,文中采用的词性标记集是基于知网[6],不同的词性标记集的采用可能会使规则库中的规则以及规则的优先级有所不同,对于规则库的构建,是在对一个大规模语料库进行训练标注的基础上建立的,任何规则库都不可能尽善尽美,所以允许对规则进行人工添加和修改。如在词性标注过程中,若规则库中的规则不能正确地对该输入文本进行正确标注,就要对规则库进行添加或修改,直至只利用规则库中的规则便可以对该文本进行正确地标注为止。目前所用规则库中的规则条目数有100多条,并且这些规则大都是在进行标注的过程中手动添加完成。在已有规则库中添加的规则如"NUM" + "CLAS" , "ADJ" + "NUM", "N" + "V", "N" + "NUM" , "PRON" + "V", "PRON"+"ADV", "PRON"+"ADJ", "PRON"+"N", "AUX""PREP", "V" + "PREP" , "CONJ" + "V" , "ADJ" + "STRU" , "V" + "V", "N"+"N", "AUX"+"V", "N"+"ADJ", "COOR"+"V", "PRON"+"STRU", "COOR"+"PRON", "STRU"+"N" , "ADV"+"PREP", "PREP" + "PRON" , "V" + "STRU" , "N" + "AUX", "V"+"PRON" , "V" + "ADJ" , "PREP"+ "V", "N"+"PREP", "N" + "COOR" , "PUNC" + "PREP" , "V" + "COOR" , "CONJ" + "ADJ" , "N" + "STRU" , "PRON"+"PREP", "PRON"+"CLAS", "V"+"V"+"CLAS", "N"+"ADV", "ADJ"+"V", "PREP"+"ADJ", "CLAS" + "V", "NUM"+"ADJ", "N"+"SUFFIX", "EXPR"+"PUNC"。其它规则见文献[5]。

1.3 基于规则优先级的词性标注算法

1.3.1 算法描述　提出的词性标注算法的基本思想如下:对给定的分词结果中各个词进行扫描,若当前词的词性是唯一的,则对该词标记为此唯一词性;若词性不唯一,则考虑此词与直接前驱词和直接后继词的组合,选择出直接前驱词与当前词的搭配符合规则中优先级最高的一个,记为 index1,再选择当前词与直接后继词的搭配符合规则中优先级最高的一个,记为 index2,然后比较规则 index1 和规则 index2 优先级的大小,选优先级大的规则中对应的词性作为当前词的词性;再标注下一个词,直至输入词串中所有的词被标注完为止。

输入:切分出的词序列,各个词之间由“/”隔开;

输出:标注好的词序列,各个词后跟“/”及该词对应的词性;

基于规则优先级的词性标注算法 T 描述如下:

T1: [初始化] 用数组 wordlist 把切分好的词串存起来, 数组中每个元素存储每个词语;

T2: [在词表中查找词性] 对数组中的从第 2 个词到倒数第 2 个词进行循环, 对每一个词语, 从词性表中查出其所有可能的词性, 若当前词的词性唯一, 则直接把当前词标记为该词性, 否则, 查出其 W_{j-1} 和 W_{j+1} 对应的所有可能的词性, 并转到 T3;

T3: [查找规则库] 找出前驱词 W_{j-1} 和当前词 W_j 的词性搭配规则中优先级最高的, 记为 index1, 然后查找出当前词 W_j 和后继词 W_{j+1} 的所有搭配规则中优先级最高的, 记为 index2;

T4: [比较优先级的高低] 若 index1 比 index2 高, 则取规则 index1, 并把当前词标记为 index1 中对应的词性; 若 index2 比 index1 高, 或两者相等, 则把当前词标记为 index2 中对应的词性; 并修改当前词的词性集合为该词性, 避免以后的查找中再查到该词的其它词性;

T5: [标注第一个词和最后一个词] 若第一个词的词性唯一, 则取该词性, 否则, 对前二个词查找规则库, 若同时满足多条规则, 则取优先级最高者; 对最后一个词也同样处理; 不同地方只是与倒数第 2 个词查找搭配规则;

T6: [人工修改规则库] 输出标注结果, 若有些词标记错误, 则人工修改规则库: 或添加规则, 或调整相关规则的优先级。

1.3.2 算法复杂度 假设分词结果中所含的词语个数为 n, 要对每个词查找词性数据库, 设数据库中记录个数为 M, 数据库查找采用的是折半查找, 其复杂度是 $O(\log_2^M)$, 故此算法的时间复杂度为 $O(n*\log_2^M)$。

2 实验结果及分析

2.1 实验环境

系统实验的硬件环境: CPU: P4 1.40 G; 内存: DDR- SDRAM 256 M; 开发工具: Microsoft Visual Studio,NET 2003, Microsoft SQL server 2000; 操作系统: Microsoft Windows 2000 professional SP4。

2.2 实验数据来源

采用《人民日报》1998 年 1 月份全部内容的语料库, 这其实是一个熟语料库, 实验时把它还原成了生语料库, 这样便于计算标注的正确率。此语料库有 400 万字左右, 切分和标注形式如: "19980101- 01- 002- 005/m 今年 /t 是 /v 党 /n 的 /u 十一 /m 届 /q 三中全会 /j 召开 /v 20/m 周年 /q ,/w 是 /v 我们 /r 党 /n 和 /c 国家 /n 实现 /v 伟大 /a 的 /u 历史 /n 转折 /vn,/w 进入 /v 改革 /vn 开放 /vn 历史 /n 新 /a 时期 /n 的 /u 20/m 周年 /q 。/w 在 /p 新 /a 的 /u 一 /m 年 /q 里 /f ,/w 大力 /d 发扬 /v 十一 /m 届 /q 三中全会 /j 以来 /f 我们 /r 党 /n 所 /u 恢复 /v 的 /u 优良 /z 传统 /n 和 /c 在 /p 新 /a 的 /u 历史 /n 条件 /n 下 /f 形成 /v 的 /u 优良 /z 作风 /n 。"。

2.3 实验结果

在对近 5 万字的语料进行了标注实验, 利用本方法标注正确率可达到 96.4%, 表 1 为一些分词和标注的例子。

表 1 部分词性标注的例子

序号	分词例句	词性标注结果
1	我们要坚持稳定、发展和平等的方针	我们 /PRON 要 /AUX 坚持 /V 稳定 /ADJ,/PUNC 发展 /V 和 /COOR 平等 /N 的 /STRU 方针 /N
2	从小学到大学	从小 /ADV 学到 /V 大学 /N
3	发展、和平等问题	发展 /V 、/PUNC 和 /COOR 平等 /N 问题 /N
4	安装在桌子上的灯亮了	安装 /V 在 /PREP 桌子 /N 上 /N 的 /STRU 灯 /N 亮 /ADJ 了 /STRU
5	着重要解决下列问题	着重 /V 要 /ADV 解决 /V 下列 /ADJ 问题 /N
6	提高人民生活水平	提高 /V 人民 /N 生活水平 /N
7	病人的妻子伤心得很	病人 /N 的 /STRU 妻子 /N 伤心 /V 得 /STRU 很 /ADV
8	不同情况下有不同解释	不同 /ADJ 情况 /N 下 /N 有 /V 不同 /ADJ 解释 /V
9	程序已经编制好了	程序 /N 已经 /ADV 编制 /N 好 /V 了 /STRU
10	怪不得我最近没见到他, 原来他病了	怪不得 /CONJ 我 /PRON 最近 /ADJ 没 /ADV 见到 /V 他 /PRON ,/PUNC 原来 /ADJ 他 /PRON 病 /V 了 /STRU

例句 1 中的“稳定”、“发展”、“平等”都应为“方针”的定语, 但这里的标注是不一样的, 因为, 分词词典中“稳定”有 2 个词性, 形容词(ADJ)和动词(V), 这里应为形容词, 而“发展”在词典中只有 1 个词性—动词(V), “平等”在词典中也只有 1 个词性—名词(N)。这就是这 3 个词虽然结构相同, 但词性标注不同的原因。例句 3 的情况和例句 1 是一样的。例句 8 中的“解释”被标注为动词, 也是因为在词典中, 此词只有 1 个词性—动词。例句 9 中“编制”和“好”分别被标注为名词和动词, 这是因为在词典中, “编制”只有名词 1 个词性, 而“好”的词性有 2 个—形容词和动词, 此处应为动词, 其对应的解释为“BeWell| 健壮”。

3 结　　语

基于规则的词性标注方法,重点就是构建完善而合理的规则库, 并对总结到的所有搭配规则在一个大规模语料库中进行统计, 求出各条规则的出现频度。在规则库中, 对每条规则都加上了优先级。有了频度值, 就可以按照它来调整优先级, 然后把不常见的搭配规则找出来, 对它们的优先级进行特别处理。方法和规则库都很简单, 并且汉语的兼类词出现次数太多, 搭配的规则很难完全总结, 所以下一步的工作就是进一步完善规则库, 并对它们的优先级进行更加合理的调整, 以期达到更高的标注正确率。

参考文献:

[1] 周强. 规则和统计相结合的汉语词类标注方法[J]. 中文信息学报, 1995, 9(3): 1-10.

[2] 魏欧, 吴健, 孙玉芳. 基于统计的汉语词性标注方法的分析和改进[J]. 软件学报, 2000, 11(4): 473-480.

[3] 王素格, 张永奎. 汉语词性标注排歧方法探讨[J]. 计算机工程与应用, 2001 (7): 70-72.

[4] 黄德根, 张丽静, 张艳丽, 等. 规则和统计相结合的兼类词处理机制[J]. 小型微型计算机系统, 2003, 24(7): 1252-1255.

[5] 温锁林. 中文文本歧义字段切分技术[J]. 语文研究, 2001 (3): 36-40.

[6] 董振东, 董强. 知网[H/OL]. Http://www.keenage.com.