关联规则在汉语词属性中的研究

王 玮 蔡莲红
清华大学计算机科学与技术系多媒体所 （北京 100084）
E-mail: wangwei.188@263.net

摘 要 在语音识别和合成中文本分析是很重要的，文本分词是文本分析正确与否的基础。目前语音合成系统中的分词一般是基于词典分析建立的，对于多音字易产生错误。文章基于数据挖掘中的关联规则的发现方法对文本分词中词语的关联关系进行研究，通过文本数据的文本特征及语音特征描述的有机结合，获取词语自身属性的关联关系，最后进行了实例测评。

关键词 数据挖掘 文本分析 分词 关联规则

Research on Correlation of China Word Attribution

Wang Wei Cai Lianhong
（Department of Computer Science, Tsinghua University, Beijing 100084）

Abstract: Text Analysis is very important in Speech Recognition and Synthesis. Word Segmentation is on the basis of text analysis. The models in most speech synthesis systems that are now being used are constructed by dictionary, qualitatively and with high errors. This paper applies data mining approach to finding association rules from text data. Through analyzing text attributes to decrease word segmentation errors. The Rules can improve segmentation precision.

Keywords: Data Mining, Text Analysis, Word Segmentation, Association Rules

1 引言

语音合成技术之所以不能广泛应用主要是由于其合成的质量较低，这其中很大程度上是由于文本预处理过程文本分析的分词错误造成的，因此要提高语音合成的自然度就必须提高文本分析的质量，尽可能的减少分词错误。

数据挖掘是发现新颖的、有效的和完全能够被人们理解的数据模式的一种方法。它结合统计和计算技术从大量的数据集中获取有用的模式，进而产生指导性的规则集合，这些规则是对数据库中数据属性、对象集的有效描述，为被挖掘处理的数据集产生摘要，提供给决策支持系统。

由于数据挖掘能够较好地找出数据集中数据项之间的关联关系，文章基于二元文法的统计概率分词模型，对其中容易产生语义及发音错误的多音字进行自身属性的关联关系发现，根据其在词或句中的位置、词性和声调等特征的描述，得到词语自身属性的描述规则。

2 文本分词

首先来描述语音合成系统中的分词算法的算法设计思想：输入待处理的文本，从输入文本的第一个字开始扫描输入的文本，根据已有的词表搜索词表中的项目，得到最长匹配的词，如果没有找到可以匹配的词语，则假设当前的汉字只是一个单个词，跳过当前的汉字，重复扫描过程。如果匹配到最长的词语，若词语中包含有 n 个汉字，记为 $C_1C_2,\cdots,C_n$，文本中对应的匹配的文字片断和后面的部分为 $C_1C_2\cdots C_nD_1D_2$，这里 D_1 以后的字符序列为匹配到的词语之后的文本。此时进行分词的歧义检测，第一步从 C_2 开始，匹配词表中最长的词语之后的文本，到达语句末尾时 $D_n=n+L_n-1$；第二步求出 $m=\max(D_i)$，$i=2\cdots n$，如果 $m>n$，就认为出现了歧义，对于出现的歧义，取文字片断 $C_1C_2\cdots C_nD_1D_2\cdots D_{m-n}$ 进行逆向词语匹配和分词，如果没有出现歧义，就顺序检测以后的词语，开始下一轮的正向扫描匹配分词。

使用这种正向和逆向混合的算法的主要依据是，正向匹配的词表容易维护，算法速度快，但是逆向匹配产生的歧义较少。使用这种机械分词算法，具有系统资源占用少，分词速度快的优点，但是存在如下问题：容易漏判分词歧义；没有得到词语的有意义的词类信息，对后续处理造成困难。

由于以上问题的存在，文章采用基于二元文法的统计概率的文本分词模型。分词后的熟语料中标注了词类，容易获得词类与词类的同现频度数据，其算法基本思想是：

设 $C=c_1c_2\cdots c_n$ 是一次输入的含有个连续汉字的字符串。$W=w_1w_2\cdots w_m$ 是将 C 进行切分后的 m 个词组成的词串，w_1、w_m 分别表示 C 的串首和串尾的界定标志；$T=t_1t_2\cdots t_m$ 是对标注后得到的词类标记串，t_1、t_m 分别为给界定标志赋予的虚词类。

设 P(W|C)是在给定字符串 C 后产生结果得到词串 W 的概率。为得到一个词串 W′，使得其概率达到最大：

$$P(W'|C)=\max_{W}P(W|C) \tag{1}$$

根据贝叶斯公式可以得到：

$$P(W|C)=\frac{P(W)P(C|W)}{P(C)} \tag{2}$$

其中，P(C)和 P(C|W)为确定值。则可以得到：

$$P(W'|C)\propto\max_{W}P(W) \tag{3}$$

所以，词的切分过程可以用寻找最大概率值的词串的过程来实现，使用词语的二元同现统计模型。但是，汉语中的词语数

基金项目：国家 863 高技术项目和国家自然科学基金资助

作者简介：王玮，1973 年生，博士后，主要研究领域为数据挖掘，神经网络。蔡莲红，1945 年生，教授，博士导师，主要研究领域为多媒体信息处理。

量非常大，很容易由于统计的语料覆盖面不够而造成统计矩阵极其稀疏，进而影响分词的准确度。

由于汉语语句的语法具有一定的稳定性，词类在句子中的分布情况具有一定的普遍意义。由于汉语中每个词语可能对应多个词类。因此，建立词语和词类的对应关系表示如下：

$$P(W)=\frac{P(T)P(W|T)}{P(T|W)} \tag{4}$$

采用二元统计，对公式(3)和(4)进行合并，得到：

$$P(W'|C)\propto\max_{W}P(t_1)\prod_{i}\frac{P(t_i|t_{i-1})P(w_i|t_i)}{P(t_i|w_i)},\ i=2\cdots m \tag{5}$$

公式(5)把词与词的同现概率统计问题转换为词类与词类之间的同现概率，某个词属于特定的词类概率，以及某个词类中的特定的词语在语料中出现的概率，与(3)相比，对语料中出现的概率，和式(3)相比，对语料的要求显然低了。但是，在词表较大的情况下，$P(w_i|t_i)$和$P(t_i|w_i)$的得出仍然是需要有足够的语料支持的。

考虑到大部分词的词类是唯一的，如果想进一步简化此模型，可以只考虑分词歧义而忽略在同样的切词方案下的词类标注的歧义，从而把$P(w_i|t_i)$和$P(t_i|w_i)$看作常数，忽略其影响：

$$P(W'|C)\propto\max_{W}P(t_1)\prod_{i}P(t_i|t_{i-1}),\ i=2\cdots m \tag{6}$$

需要指出的是，由于t_1为句首标记，显然$P(t_1)$是一固定的常数，因此可以得到：

$$P(W'|C)\propto\max_{W}\prod_{i}P(t_i|t_{i-1}),\ i=2\cdots m \tag{7}$$

这个公式只涉及了词类的二元同现概率，对于中小规模的语料库来说，是一个可以进行实际计算的模型。

3 关联规则的描述

关联规则描述的问题是：在给定的数据库中，每个交易包含一个数据项集，关联发现函数作用在这个交易集上，返回各项集之间存在的密切关系。找出这样的一些规则对于确定市场策略是很有价值的，它表示用户在购买某些物品时会有多大的倾向购买其他种类物品，决策者可以根据诸如“90%的用户在购买面包和牛油的同时也会购买牛奶”的规则把存放牛奶的货架和存放面包、牛油的货架挨在一起。这样牛奶的购买力就会增强。这种密切关系可以这样说明：“包含X，Y的交易中的60%也包含Z，包含X，Y，Z的交易占整个交易集的15%”，其中的百分比分别定义为关联的信任度和支持度。X，Y，Z既可以是单个数据项，也可以是数据项集，但是其交集一定是空集。上述的关联规则形式化描述是：(X，Y)⇒Z|(60，15)。

关联规则的模型定义如下：$I=\{i_1,i_2,\cdots,i_m\}$是m个不同项目交易的集合。对于一个给定的交易数据库D，其中的每个交易T是I中的一组项集，即T包含I。每一个交易都与一个唯一的标识符TID相对应，如果I中的每一个子集X，存在X包含T，就表示一个交易包含X。关联规则蕴含式形如X⇒Y形式，X和Y是项集，X包含I，Y包含I，且$X\cap Y=\emptyset$。若用P(X)表示交易发生的概率，关联规则X⇒Y的支持度因素和信任度因素就可以分别定义如下：P(XY)和P(XY)/P(X)。挖掘的关联规则就是发现那些满足用户指定的最小支持度和信任度的规则。

关联规则的算法过程描述如下：

```
Begin:
  input minsupport;
  L1={frequency items};
  S=0;
  For(k=2;Lk-1≠Ø;k++)
  Ck=NewCandidate;
  For(all items t in D)
  For(all k sub-item sets in t)
  if(s∈Ck)s=s+1;
  Lk={c∈Ck|c≥minsupport};
  All frequency item sets=⋃k Lk;
End
```

4 实验结果

目标测试集是“为”在不同的情况下数据描述，通过将“为”的文本特性(如“为”的词性可以分成介词、助词、副词)、语音特性(如音调发二声、四声、五声)及其位置特性(句首、句中、句末)等特征作为“为”的属性描述。训练测试集如下：

作为	认为	成为	为了	君为轻	取名为
十二人为一组为祖国健康工作					
这是为农民养鱼起示范作用					
今年上海市新建的超市以连锁经营科学管理便民服务和规模效益为目标					
员工们都被王青伟这种不为一家为万家的精神所感动					
还因为汪英俊被人们称为奇人					
尽管已有完成乌拉圭回合协定为标志的显著成绩					
八宝乡沼气服务站去年为户农民建池					
云南市场上的微型车多为重庆长安柳州五菱车系列					
上海家庭电暖器拥有量约为百分之十					
命名为春华沟					
位于北京西罗园小区的全聚德西罗园烤鸭店集餐饮娱乐住宿为一体					
被中东阿拉伯地区奉为朝圣佳品					
以辽宁中部城市群和东北三省及内蒙古东部为腹地					
以小汤山中心区为最高点					
并被选为河南省工商联副主任					

基于以上数据集可以得到如下规则：

```
tone=2& 词性=介->location=E [Coverage=0.476(10);
                             Strength=0.400;]
tone=2 & location=M->词性=助 [Coverage=0.524(11);
                              Strength=0.273;]
location=M& 词性=介->tone=4 [Coverage=0.524(11);
                             Strength=0.364;]
tone=4 -> 词性=介 [Coverage=0.238(5);Strength=1.000;]
词性=介 -> tone=4 [Coverage=0.762(16);Strength=0.313;]
tone=2 -> location=E [Coverage=0.714(15);Strength=0.267;]
location=E -> tone=2 [Coverage=0.190(4);Strength=1.000;]
tone=2 & location=M -> 词性=副 [Coverage=0.524(11);
                                Strength=0.182;]
location=E -> 词性=介 [Coverage=0.190(4);Strength=1.000;]
tone=4 & location=M -> 词性=介 [Coverage=0.190(4);
                                Strength=1.000;]
词性=介 -> location=E [Coverage=0.762(16);Strength=0.250;]
tone=2 -> 词性=助 [Coverage=0.714(15);Strength=0.200;]
词性=助 -> tone=2 [Coverage=0.143(3);Strength=1.000;]
location=H -> tone=4 [Coverage=0.048(1);Strength=1.000;]
tone=4 -> location=H [Coverage=0.238(5);Strength=0.200;]
```

(下转58页)

计算机5个方面的题材，共约21万的训练语料。从训练语料获取的一模式为出现的单个词和词性。按在语料中出现的次数排序，取前十位为N(37149)、V(26863)、U(8869)、D(6619)、的(6372)、M(6277)、P(4721)、R(4623)、Q(4120)，由于一模式未利用上下文信息，因此，不构成搭配规则。其它搭配模式候选搭配规则集与搭配模式规则集的部分结果见表1与表2。

表1 候选搭配模式

模式	左2	左3	左4	左5	右2	右3	右4	右5
规模	10750	50010	74291	240201	11411	96487	96487	228824

表2 搭配模式规则集

模式	左2	左	左4	左5	右2	右3	右4	右5
规模	7683	40537	64892	92892	7892	41097	70280	102081

表1与表2为搭配模式候选搭配规则集与搭配模式规则集中规则的规模的统计结果，其搭配模式规则集中的规则如："左2模式：A N (0.92)；右2模式：N 和 (0.852)、V 国有(0.862)；左3模式：V D V (0.924)；右3模式：N 和 N (1)、V 老师 D (0.846)；左4模式：她 D V U (1)；右4模式：V 说 的一句话 (0.791)"等等，其中带下划线的为目标词的标注词性。其中"A N (0.92)"这条规则的含义为：if (词1,A) then (词2,N)，它表示前一个词的词性为"形容词"，后一个词的词性为"名词"。

作者对20万字中126篇新闻语料已分词但未标注词性的语料，采用二元统计模型[3,7]进行标注后，再利用搭配模式规则，对其兼类词进行二次标注。将前5篇的词性标注结果列在表3中。表3中的搭配模式规则对应的词性标注正确率为在统计标注的基础上，再利用搭配模式规则对兼类词的词性给予修正后的词性标注正确率。

另外对1万词开放测试：基于二元模型的方法测试的准确率90.3%，其中兼类词标注准确率为78.5%；混合方法测试的准确率为94.8%，其中兼类词标注准确率为83.6%。

表3 5篇语料的统计结果

文章号	词条数	统计标注		搭配模式规则	
		全部词	兼类词	全部词	兼类词
1	1139	92.1	86.4	93.9	89.4
2	1251	94.0	88.0	95.3	90.1
3	1544	94.8	90.8	96.3	91.3
4	1350	94.3	81.6	95.7	87.5
5	9568	93.3	80.2	95.4	86.6
汇总	20448	93.7	85.4	95.3	88.9

由于汉语兼类现象分布不均匀，对一些用法灵活、兼类较多的词，不管采用哪一种统计模型，只要是单纯的统计模型都不能有效地反映出词本身的一些确定语法特点，这时应根据其语法功能以及统计现象归纳出词性规则。基于规则的方法，可以将上下文的信息包含在规则中，这样有助于理解和简化标注系统，而且不需要一个较大的统计数据表；而基于概率统计的方法虽然需要26×26的同现矩阵，对于兼类词也必须计算出词汇概率，但概率统计方法能够解决规则不能解决的问题，它是一种柔性的方法。比如，有些训练实例与其它训练实例的重叠性较差，使得训练语料中部分搭配没有加入到搭配模式规则集，为了解决这一问题，规则的方法只能靠扩大语料库的规模来增强规则的适应范围，但又会产生规则的数量增加，出现系统处理的速度减慢和规则间的相互影响明显加剧等新的问题，因此，规则的规模应加以限制。若词性规则规模不能覆盖所有的语言现象，则可以把基于规则的方法作为基于概率统计方法的补充，从而构成一种高效的词性标注方法。实验表明这种混合的方法能将单纯基于概率统计方法标注词性的正确率提高3%左右。(收稿日期：2000年10月)

参考文献

1.Eric Brill, Philip Resnik.A Rule-Based Approach to Prepo-sitional Phras Attachment Disambiguation.http:// ww.cs.jhu.edu/~brill/ acad-pubs.html

2.Eric Brill.Transformation-Based Error-Driven Learning and Natural Language: A Case Study in Part of SpeechTagging.Computational Linguistics, 1995; 21(4): 543-565

3.刘开瑛，郑家恒，赵军.语料库词类自动标注算法研究.机器翻译研究进展，电子工业出版社，1992.8：378—385

4.温锁林.中文文本兼类词的标注技术.中文信息处理国际会议论文集，清华大学出版社，1998：194-199

5.王素格，苗夺谦，刘开瑛.基于 Rough Set 自动获取词性标注规则初探.Beijing International Conference on Machine Translation & Computer Language Information Processing, 1999.6: 469-474

6.孙杰，林鸿飞，姚天顺.一种获取机器翻译系统词类搭配规则的机器学习方法.模式识别与人工智能，1999，6；12(2)：157-163

7.王素格.汉语词性标注知识获取方法研究 .山西大学硕士学位论文，2000.6

(上接18页)

词性=助 -> location=M [Coverage=0.143(3); Strength=1.000;]

location=M -> 词性=助 [Coverage=0.762(16); Strength=0.188;]

tone=2 -> 词性=副 [Coverage=0.714(15); Strength=0.133;]

词性=副 -> tone=2 [Coverage=0.095(2); Strength=1.000;]

第一条规则表示的是当"为"的声调是二声、词性是介词时，其出现在句或词末的可能性为47.6%，这条规则的连接强度为0.4，第二条规则表示的是当"为"的声调是二声、在句或词中时，词性是介词时的可能性为52.4%，这条规则的连接强度为0.273，以后的规则含义可以类推得到。

5 结束语

在合成系统中，文本分析正确与否直接影响其合成质量，文章提出了应用数据挖掘领域中关联规则的发现方法用于汉语词中的自身属性的规则化描述，通过获取这样的规则，可以对以后的分词产生指导，提高分词的正确率。文章下一步的研究工作是结合基于规则和统计方法，合理选择最有可能影响语义和发音的属性。(收稿日期：2000年11月)

参考文献

1.胡其炜.中文文本表音与韵律分析研究.清华大学硕士学位论文，2000

2.Rakesh Agrawal, Ramakrishnan Srikant.Fast Algorithm for Mining Association Rules.In Proceedings of the 20th VLDB Conference, Santigo, Chile, 1994

3.Ramakrishnan Srikant, Rakesh Agrawal.Mining Generalized Association Rules.In Proceedings of the 21st VLDB Conference, Zurich, Switzerland, 1995

4.Rakesh Agrawal, Tomasz Imielinski, Arun Swami.Mining Association Rules between Sets of Items in Large Databases.In Proceedings of the 1993 ACM SIGMOD Conference, Washington DC, USA, 1993.5