基于字典和统计的分词方法

陈 平,刘晓霞,李亚军
CHEN Ping,LIU Xiao-xia,LI Ya-jun

西北大学 信息科学与技术学院,西安 710127
Institute of Information Science & Technology,Northwest University,Xi'an 710127,China
E-mail:simaping2007@126.com

CHEN Ping,LIU Xiao-xia,LI Ya-jun.Chinese word segmentation based on dictionary and statistics.Computer Engineering and Applications,2008,44(10):144-146.

Abstract: Proposes a method based on dictionary and statistics.The method uses the changed dictionary structure that is able improve efficiency,then uses statistics to deal with the unregistered words left over in the first step,also can resolve most ambiguity.

Key words: word segmentation based on dictionary;word segmentation based on statistical method;crossing ambiguities;unregistered

摘 要:提出了一种基于字典与统计相结合的中文分词方法,该方法利用改进的字典结构能够快速切分,在其基础上进一步利用统计的方法处理所产生未登录词,并且能解决大部分交集歧义问题。

关键词:基于字典的分词;基于统计的分词;交叉歧义;未登录词
文章编号:1002-8331(2008)10-0144-03 **文献标识码**:A **中图分类号**:TP391

1 引言

中文分词是文本分类和信息处理的基础,歧义处理和未登录词识别是中文分词的两大难点。目前主要有三类分词技术:基于字典的分词、基于统计的分词、基于规则的分词。基于字典的分词切分简单,但正确率只有80%左右,速度慢;基于统计的分词能识别高频未登录词,也不易出现歧义,但准确率低;基于规则的分词正确率较高,但非常复杂。实际使用的分词系统都是把使用字典的机械分词作为一种初分手段,再利用其它分词方法来进一步提高切分的准确率,包括未登录词的识别。本文采用改进的字典结构来提高字典处理的速度,并用统计方法解决中文分词的两大难点:歧义问题和未登录词识别问题。

2 基于字典的处理

2.1 改进的字典存储结构

分词词典的查询速度在很大程度上制约着匹配算法的执行效率[1]。汉语词典一般有十几万词条,如果每次匹配都检索全部词典,效率是可想而知的。目前为提高词典的查找速度,许多研究者开始提出一些分词模型[2],简单常用的是多级索引的词典结构,但其实践并不尽如人意,文献[3]中计算表明,进一步做二级甚至三级索引所节省的空间远不够因数据结构复杂化而带来的开销。本文使用首字的区位码作为标识的一级数组索引,指向以此字为首的词的HashMap。汉字的区位码就是GB2312码中的汉字部分。包括单字词在内,词典中的首字有6 763个,声明一个大小6 763的数组就足够了,如dic[6 763]。每个汉字对应一个区位码,java中可以通过str.getBytes("GB2312")获得任意汉字的机内码,再对得到的两个字节分别减去A0得到区内码,比如"人"字,它的区位码是4043,那么以"人"为首的词都存在数组元素dic[4043]中,与每个首字对应的词有几个到几百个,它们以HashMap数据结构存放,key是词,value是词频。结构如图1所示。这种改进的词典存储结构解决了首字匹配的问题,从而有效地提高匹配效率。

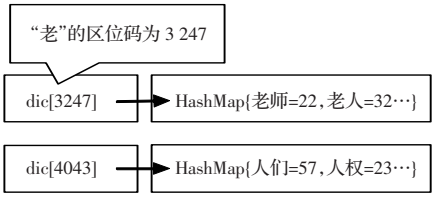

图1 词典的存储结构

2.2 匹配算法

一些统计表明[3,4],单纯使用正向最大匹配的错误率为1/169,单纯使用逆向最大匹配的错误率为1/245,逆向匹配的切分精度略高于正向匹配,遇到的歧义现象也较少。所以本文采取逆向匹配算法。

定义1 逆向最大匹配(RMM)对于文本中的字串$ABC,C\in$

基金项目:陕西省自然科学基金(the Natural Science Foundation of Shaanxi Province of China under Grant No.2006F50);航空科学基金项目(No.06ZC31001)。

作者简介:陈平(1982-),女,硕士研究生,主要研究方向:信息处理、Web挖掘;刘晓霞(1965-),女,副教授,主要研究方向:信息处理、图形图像处理;李亚军(1983-),男,硕士研究生,主要研究方向:信息处理。

收稿日期:2007-07-23 **修回日期**:2007-10-18

$W, BC \in W, ABC \notin W, W$ 为字典,那么就取切分 A/BC。

鉴于逆向最大匹配算法简单易行,这里不再作赘述。

3　基于统计的处理

歧义词问题和未登录词识别问题是中文分词的两大难题,本文用统计的方法来识别未登录词,发现并处理大部分交集歧义词。

3.1　歧义的处理

从歧义字段的切分结果来看,歧义字段可以分为真歧义字段和伪歧义字段。根据文献[6]的统计,伪歧义字段占总歧义字段的 94%,其中 84.10%属于交集型歧义,据文献[7]统计,交集型歧义字段的切分结果形式虽然多样但仍具有一定的规律:(1)链长为 1 的歧义字段的切分结果为 A/BC 和 AB/C 的歧义字段占到了链长为 1 的歧义字段数的 89%;(2)链长为 2 的歧义字段 ABCD 中,切分结果为 AB/CD 的占约 98%。二者合计占到了歧义字段的 97.61%,若以此规律为切词基础,那么将链长为 2 的歧义字段转换为链长为 1 的歧义字段处理将同样会达到上述的效果。

本文将依据此规律对链长小于 3 的歧义字段进行以下处理:

(1)在切词过程中,切出一个词 W_{i-1};

(2)指针回退一个字长,在剩余短语中继续切分;

(3)如果能在最右端切分出一个词 W_i,则发现歧义。然后统计 W_{i-1} 和 W_i 的词频,取较大的作为切分结果;

(4)若词频没有太大差别,以逆向最大匹配算法切词。

3.2　未登录词识别

未登录词包括新词(如一些领域出现的通用词或专业术语)和专有名词(如人名或地名)。在许多资料和研究中对专有名词的处理是通过建立专有名词词典,在切词过程中匹配词典。这样做虽然直接有效,但其不足之处是所依据的大规模语料库的建设非常繁琐,且不易有穷列举。文献[8]中提出了一种自适应的分词算法来识别未登录词,在逆向匹配的基础上依次切分出 W_n、W_{n-1}、…、W_i,对剩下的字符串跳过 n 个字长从右至左如果能再切分出一个词 W_{i+1},则认为 W_i 和 W_{i+1} 之间的 n 个字可能组成一个未登录词,但再次切词时还是会留有盲点,甚至有可能会将词组分开。本文提出了一种改进的识别未登录词的算法,不需要大规模语料库和专有名词词典,也无需借助词类信息,可以有效地识别一些未登录词。

算法原理:设待处理的字符串 S,则 S 由若干个词组成,记为 $S=W_1W_2\cdots W_i\cdots W_n$。其中 W_1、W_2、…长度不等。设 W_i 是分词词表中没有收录的未登录词,那么组成分三种情况:(1)由若干个单字组成(如,驻马店,李小龙;这部分地名和人名较为多见)(2)由若干个词组组成;(如,国际米兰)(3)由单字和词组成。(如,粗糙集)。这种情况下需要判断正在处理的词和刚切分出的词的相关性,因此本文采用类似 2-gram 模型的方法。先了解以下 2-gram 模型:给定句子 $S=W_1W_2\cdots W_n$。由链式规则:$P(W_i)=P(W_i|W_{i-1})$,对 $P(W_i|W_{i-1})$而言,W_{i-1} 即为 W_i 的历史。考虑前面 1 个词构成历史的模型即为 2-gram 模型。采用此模型的优点是计算代价小,且无需太多训练语料。令 $c(W_1,\cdots W_i)$表示词串 $W_1W_2\cdots W_i$ 在训练语料中出现的次数,则由最大似然估计,$p(W_i|W_{i-1})=c(W_{i-1}),W_i)/c(W_{i-1})$。

具体作以下处理:

(1)把预处理分出的词按分隔符分成 String 数组,创建主 HashMap。

(2)取两个词组成一个 2-gram,pre 为前一个词,word 为随后的词,若 pre 不在 HashMap 中则将其存入,紧接着为其创建子 HashMap,将 word 存入对应的子 HashMap 中。

(3)统计 $c(W_i)$、$c(W_{i-1})$、$c(W_i,W_{i-1})$的词频,计算出 $p(W_i|W_{i-1})$。

(4)若 p 大于一定的阈值,则判定 pre 是未登录词,否则,将其切分。

在算法中重要的是建立 2-gram 模型的数据结构,本文采用了二级索引的 HashMap 的结构,它可以提供常数时间的查找,在存储时,每一个 key 值对应一个在训练语料中出现过的词语,每个 key 之对应得 value 值又是一个子 HashMap。建立数据结构的相关算法如下:

```
public void add(String pre,String cur)
{
    String key=pre;
    String word=cur;
    boolean b=HM.containsKey(key); //HM 是一个 HashMap,这里判断 pre 是否在主 map 中
    if(b==false)
     {                                    //若主 map 中没有,则添加进去
        HashMap hm=new HashMap(); //创建子 HashMap
        Hm.put(key,newInteger(1)); //存储主 KEY 的频率
        Hm.put(word,newInteger(1)); //存储主 KEY 后那个词的频率
        HMap.put(key,hm); //将主 KEY 和对应得子 Map 放入主 Map 中
      }else
          {                                    //若主 Map 中含有该词,则对其值进行修改
            HashMap temp=HM.get(key);
            int count=((Integer)temp.get(key)).intValue()+1;
            temp.put(key.new Integer(count));
            //判断子 Map 中是否有该词,若无,则添加进子 Map,若有,则修改其值
            if(temp.containsKey(word))
             {
               int value=((Integer)temp.get(word)).intValue()+1;
               temp.put(word.new Integer(value));
             }else
                 {
                   temp.put(word,new Integer(1));
                   HM.put(key,temp);
                 }
            }
}
```

4　实验结果及分析

采用 Sogou 语料库,抽取了 1 000 个样本进行测试。包括新闻、体育、科技、医学、军事、旅游、文学 7 个方面。评价的性能参数主要是:

(1)召回率=切分正确的词/切分出的词总数;

(2)分全率=切出的词数/原文应有词的个数;

(3)准确率=切分正确的词/完全正确切分时的词数;

实验结果如表1所示。

表1 分词效果

	新闻	体育	科技	医学	军事	旅游	文学
数据规模	32 K	32 K	32 K	36 K	40 K	32 K	28 K
准确率/%	96.62	97.54	96.89	96.96	97.21	96.93	97.05
分全率/%	98.25	97.85	98.03	97.92	98.19	97.94	98.89
召回率/%	94.52	95.36	95.24	94.64	95.31	94.89	95.42

把本算法与天津市海量科技发展有限公司开发的海量分词系统比较,以一篇新闻类文本进行测试,文本规模是4 275个字节,海量分词系统分词历时2 266 ms,本文算法用时1 781 ms。词的切分也有进步之处。结果如表2。

表2 分词对比

测试用例	海量分词系统	本文算法	备注
这家公司叫做快钱,说这句话的是快钱的CEO关国光	这/家/公司/叫做/快/钱/,/说/这/句/话/的/是/快/钱/的/CEO/关/国光	这家/公司/叫/做/快 钱/,/说/这/句/话/的/是/快钱/的/CEO/关国光	可以识别频率出现高的“快钱”和人名“关国光”
“快钱的后端绝不做 应用。”关国光信誓旦旦/地表示	“快/钱/的/后/端/绝不/做/应 用/”/关/国光/信誓旦旦/地/表示	快钱/的/后端/绝不/做/应用/”/关国光/信誓旦旦/地/表示 全国/已/有/641/占/	可以识别低频但连续的单字成词
全国已有641占全国 人口20%的城市开展 试点工作	全国/已/有/641/占/全国/人口/20%/的/城市/开展/试点/工作	全国/人口/20%/的/城市/开展/试点/工作	可以识别数字和百分比

此外,还采集了一些歧义字段集进行测试,结果表明:本算法可以正确识别大部分链长为1或2的歧义字段,对于链长大于3的歧义字段则不能正确切分。

5 结论

本文作者创新点:提出了结合字典与词频的分词方法,通过改进的词典存储结构可以快速切分,又通过词频统计解决了未登录词问题和大部分歧义问题,由实验结果可知,此算法的分词系统的分全率在97%以上,但在歧义处理方面还不能处理多链歧义和组合歧义,这将是以后的研究重点。

参考文献:

[1] Li Qing-hu,Chen Yu-jian,Sun Jia-guang.A new dictionary mechanism for Chinese word segmentarion[J].Journal of Chinese Information Processing,2003,17(4):13-18.

[2] 李双龙,刘群,王成耀.基于条件随机场的汉语分词系统[J].微计算机信息,2006,22(10):178-180.

[3] 张国煊,王小华,周必水.快速书面汉语自动分词系统及算法设计[J].计算机研究与发展,1993(1):63-67.

[4] Ma Yu-chun,Song Han-tao.Research of Chinese word segmentation based on the Web[J].Computer Application,2004,24(4):134-136.

[5] 邹海山,吴勇,吴月珠.中文搜索引擎中的中文信息处理技术[J].计算机应用研究,2000,17(12):21-24.

[6] He Ke-kang,Xu Hui.Design of an expert system of automatic word segmentation in written Chinese text[J].Journal of Chinese Information Procesing,1991,5(2):1-14.

[7] Yan Yin-tang,Zhou Xiao-qiang.Study of segmentation strategy on ambiguous phrase of overlap type[J].Journal of the China Society for Scientific and Technical Information,2000,19(6):637-643.

[8] 黄水清,程冲.基于既定词表的自适应汉语分词技术研究[J].现代图书情报技术,2006,5:13-17.

(上接122页)

交叉熵,同时在给定微小区软切换概率前提下,对数交叉熵随微小区软切换用户切换到宏小区的概率 p_{mM} 的变化而变化,且在 p_{mM}=0.35时达到极小值,这一结果反应了 p_{sho} 和 p_{mM} 的优化程度。

将图2和图3的结果相结合,并考虑到仿真中误差影响,微小区软切换用户切换到宏小区的概率 p_{mM} 一般控制在30%以下,既符合宽带宏小区的性能指标,又合理地降低了微小区系统软切换呼叫的阻塞率,进而提高了整个CDMA系统的性能。因此,对数交叉熵和微小区软切换用户切换到宏小区的概率 p_{mM} 的引入,更好地分析了3G及B3G移动通信系统的相应性能,并合理地解决了微小区软切换过程中移动用户由于微小区资源紧张而造成的呼叫终止情况,特别在城市的商业区或人口密集区,这个特点更为突出。

4 结论

本文针对一个宏小区覆盖一个微小区的重迭小区结构软切换性能进行了定量分析和仿真。为体现仿真结果的可靠性,引入对数交叉熵概念加以分析。从分析结果看出,合理地选择微小区软切换用户切换到宏小区的概率,可相应降低重迭小区结构CDMA系统微小区软切换呼叫的阻塞率,进一步改善重迭小区结构系统整体性能,提高CDMA系统的通信质量,该分析结果对3G及B3G的CDMA移动通信系统的参数设置及性能优化具有重要的参考价值。

参考文献:

[1] Kang C S,Cho H S,Sung D K.Capacity analysis of spectrally overlaid macro/microcellular CDMA systems supporting multiple types of traffic[J].IEEE Transactions on Vehicular Technology,2003,52(2):333-346.

[2] Liu Zhi-ping,Wang Ya-feng,Yang Da-cheng.Effect of soft handoff parameters and traffic loads on soft handoff ratio in CDMA Systems[C]//Proceeding of ICCT2003,2003:782-785.

[3] Steele R,Lee C C,Gould P.GSM,cdmaone and 3G Systems[M].[S.l.]:John Wiley&Sons Ltd,2001.

[4] Homnan B,Bebjapolakul W.QoS-controlling soft handoff based on simple step control and a fuzzy inference systems with the gradient descent method[J].IEEE Transactions Vehicular Technology,2004,53(3):820-834.

[5] Gilhousen K S,Jacobs I M.On the capacity of a cellular CDMA systems[J].IEEE Transactions Vehicular Technology,1991,40(2):303-311.

[6] Shapira J.Microcell engineering in CDMA networks[J].IEEE Transactions Vehicular Technology,1994,43(4):817-825.

[7] Shore J E,Johnson W.Properties of cross entropy minimization[J].IEEE Transactions on Information Theory,1981,27(4):472-482.

[8] de Boer P T,Kroese D P,Mannor S,et al.A tutorial on the cross-entropy method[EB/OL].http://web.mit.edu/6.454/www/www_fall_2003/gew/CEtutorial.pdf.