

最大熵算法在汉语拼音标注中的研究与实现

张丽青, 寿永熙, 马志强

(内蒙古工业大学 信息工程学院, 内蒙古自治区 呼和浩特 010080)

摘 要: 经过对最大熵模型的研究, 找到一种适合汉语拼音标注的最大熵模型算法. 利用该算法解决了多音字单字成词的情况, 从而使得所有包含多音字的词都是两字或多字词. 使用该算法随机抽取“读者文摘”中的一篇文章进行标注实验, 实验表明拼音标注正确率达到了 96.6% 以上.

关键词: 最大熵模型; 多音字; 拼音标注; 统计; 特征; 分词

中图分类号: TP391

文献标识码: A

文章编号: 1000-7180(2012)08-0120-03

The Research and Implementation of Maximum Entropy Algorithm in Phonetic Annotation

ZHANG Li-qing, SHOU Yong-xi, MA Zhi-qiang

(College of Information Engineering, Inner Mongolia University of Technology, Huhhot 010080, China)

Abstract: Through maximum entropy model study, a algorithm for maximum entropy model that is for pinyin marked must be founded. Using the algorithm put an end to the situation that polyphone word is considered to be a word, so that all words with multiple pronunciations are two or more words. Using the algorithm mark the article in Reader's Digest, the results show that pinyin marked rate has reached 96.6 percent or more.

Key words: maximum entropy model; polyphony; phonetic annotation; statistics; features; segmentation

1 引言

当今使用计算机来实现汉语拼音标注的方法主要有两类:(1)直接给汉字标注拼音;(2)分词后标注拼音. 直接给汉字标注拼音很难解决一字多音的问题. 如果没有上下文关系, 我们在拼音标注的时候是无法确定读音的. 而如果先分词就可以以词为单位解决多音字读音的问题了. 但是这样的标注依然有其局限性, 因为很多词是以单字为词的, 或者很难给予拼音标注有利的分词结果. 缺失了上下文关系, 就无法拼音标注了.

为了解决上述方法标注拼音时出现的问题. 本文提出了一种改进的适合于拼音标注的分词算法.

2 最大熵算法原理

汉语分词是将一个汉字序列切分成一个个单独的词. 分词就是将连续的字序列按照一定的规范重

新组合成词序列的过程. 在英文中, 单词之间是以空格作为自然界符的, 而中文只是字、句和段能通过明显的分界符来简单划界, 唯独词没有一个形式上的分界符. 由于中文在基本文法上有其特殊性, 因此引入分词技术.

汉语分词是文本挖掘的基础, 对于输入一段中文, 将其成功进行分词, 可以达到电脑自动识别语句含义的效果. 如何让电脑自动识别语句含义? 其处理过程就需要分词算法.

分词算法本质上可以划分三类:(1)基于规则的分词方法;(2)基于统计的分词方法;(3)人工智能的分词方法^[1-5].

其中基于统计的分词方法得到了越来越多的应用, 最终分词的效果也是很明显的. 它是一种通过统计计算字与字之间相邻共现频率或概率来反映两个字之间成为词的可信度的方法. 这其实就是分析一个字上下文关系的方法. 而且统计来源于现实世界

真实的语料,最终的分词结果更具有可信度.本文采用最大熵法来进行分词,并将分词之后的结果进行拼音标注.

最大熵方法是目前解决分词算法中一种比较有效的算法^[6].最大熵模型是一种广泛应用于自然语言处理中的概率估计方法,可以综合观察各种相关或不相关的概率知识,对许多问题的处理结果都达到或者超过了其他方法的最好结果.贾丽洁^[7]对最大熵模型在分词中的技术进行了研究.构建了一个中文分词系统,使用语料库进行最大熵模型的训练,并对文本进行词性标注,最终得到文本的分词结果.

下面结合一个例子来说明本文如何在汉语拼音标注中使用和建立最大熵模型.例如在汉语中“落”这个字有四个读音,分别是 luò、là、lào、luō.

(1)他在有了生活体验后才落笔画画.落字读 luò.

(2)我忙着出来,把票落在家里了.落字读 là.

(3)他因睡觉时受寒或枕头头的姿势不合适,以致脖子疼痛.落枕了.落字读 lào.

(4)她大大落落走上讲台.落字读 luō.

我们看到“落”字的发音有四个,想要正确标注起来很困难.我们已知的落的发音只有以上四种.那么就假设这四种发音存在如下关系.

$$P(S_1|w) + P(S_2|w) + P(S_3|w) + P(S_4|w) = 1$$

$P(S_i|w) (1 \leq i \leq 4)$ 表示“落”在汉语句子中标注为 S_i 的概率.在这种情况下我们认为每个发音的概率是相同的,也就是均等一致的概率分布.但在实际过程中,发现这四种发音出现的概率是不相同的,luò 的发音最多.也就是说根据这个现象,我们在标注拼音的时候标注为 luò 时正确率会大一些,可这样标注的效果并不好,因为其他发音总是会出现的,一旦出现就会标注出错.为了能更好地标注,我们继续对句子观察研究后发现,当“落”字后面是“在”字的时候发 là 字的概率要比其他时候大很多.这说明该随机过程是受到上下文信息的影响的,出于这种现象就可以引入一个二元函数来表达这种现象:

$$f(x, y) = \begin{cases} 1 & \text{if } y = \text{luò} \wedge \text{next}(x) = \text{'在'} \\ 0 & \end{cases}$$

x 表示上下文环境,上面的公式表达一个含有“落”字的句子, y 表示输出,对应着“落”字的发音,可看作是上下文环境 x 的一个函数.表示 x 中跟在“落”字后面的一个字是“在”.我们称这一函数是一个特征函数^[8-10],简称特征.

根据以上关于特征函数的概念,在文本中需要

标注的字 w 和在 w 周围出现的字都可以作为一个特征,但是这并不是说只考虑 w 后面的一个字,我们考虑的也可能是 w 前面或后面的一个字或者 n 个字.

在实际应用中发现符合这样的特征集合会非常大.而我们需要的特征只是有用的特征.所以必须再引入一种约束,使得有用的特征保留下来,无用的特征去除.约束的相关概念如下:假设有 n 个特征,则模型属于约束产生的模型集合,即

$$C = \{p \in P \mid p(f_i) = \tilde{p}(f_i), i \in \{1, 2, \dots, n\}\},$$

式中, $p(f_i)$ 为模型概率, $\tilde{p}(f_i)$ 为经验概率.

模型的目标是产生约束集下最均匀分布的模型,而条件概率 $p(x|y)$ 均匀性的一种数学方法为条件熵,定义为

$$H(p) = - \sum_{x,y} \tilde{p}(x) p(y|x) \log p(y|x)$$

其中 $0 \leq H(p) \leq \log |y|$

所以最大熵模型可以描述为:若在允许的概率分布 C 中选择模型,具有最大熵的模型 $p_* \in C$ 就认为是所选模型,

$$p_* = \arg \max H(p).$$

3 算法描述

要利用最大熵原理实现一个适合拼音标注的分词算法.要对文本分词的地方只关心含有多音字的部分,即一切的分词行为都是围绕着多音字展开的.其他非多音字不进行分词处理.为了建立一个最大熵模型,需要有大量的实例文本来训练我们的模型,目的是在实例中寻找多音字及其上下文信息,建立一个特征集合,然后把不满足约束条件的特征去掉,计算满足约束条件下的概率分布,最后形成具有最大熵模型的特征库.

实例文本需要从训练语料库中得到,我们选用了《人民日报》汉语拼音标注语料库,该语料库是一个已经拼音标注的熟语料.算法建立过程可分为 4 个步骤:(1)扫描语料库中的文本,找到所有的多音字及其出现的次数及在文本中的出现位置.这样就形成了一个多音字集合 $W = \{w_1, w_2, \dots, w_m\}$;(2)构建特征集合.上下文特征函数为 $f_{(w,y)}(x,y)$, x 是输入的多音字上下文环境, y 是输出的分词结果. x 是出现 w 且包含 n 个字符的文本串.由此看出,样本中出现多音字周围的所有包含 n 个字符的串均可作为模型的特征,但要注意 x 中 w 读音一致的才认

为是同一特征. 根据约束条件筛选合适的特征. 我们的特征都存在于语料库文本中, 筛选出的特征满足样本中出现的事件. 此时的特征仍然是比较粗略的. 这里使用基于频次和近似增益算法来进一步筛选特征. 其中增益计算公式为

$$G_{s,f}(\lambda) \equiv - \sum_s \tilde{p} \log Z_{\lambda}(x) + \lambda \tilde{p}(f)$$

该算法描述见图 1; (3) 计算满足最大熵条件的概率

$p(x|y)$ 具有 Gibbs 分布: $p(x|y) = \frac{1}{z^{\lambda}(y)} \exp \sum_{i=1}^n \lambda_i f_i$

(x, y) , 其中 $z^{\lambda}(y) = \sum \exp \sum_{i=1}^n \lambda_i f_i(x, y)$ 为归一化因子. 这种概率分布唯一, 符合最大似然分布; (4) 根据计算出最大熵条件的概率 $p(x|y)$, 输出带有唯一拼音的特征模型库, 以后标注拼音时, 只要在文本中匹配到已给特征便可以标注出汉语拼音. 流程图见图 1.

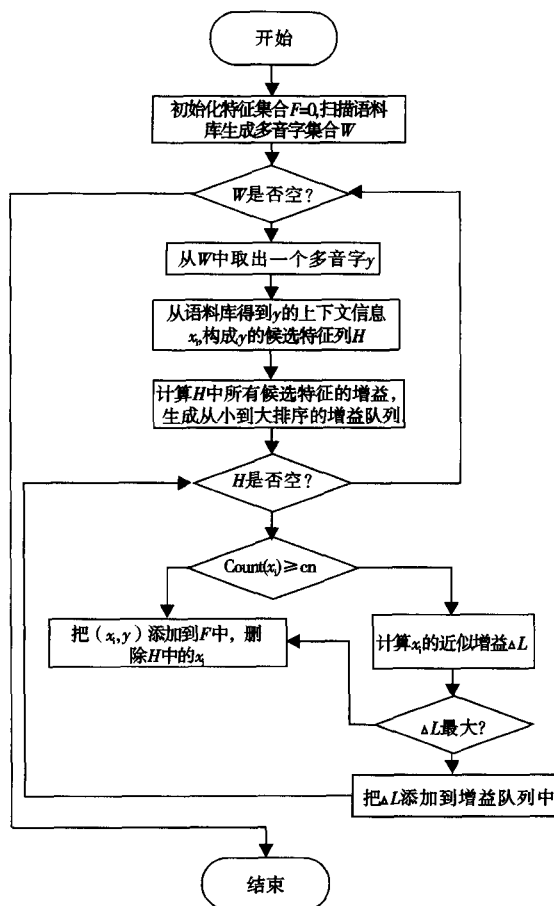


图 1 特征选择算法流程图

4 实验测试与数据分析

4.1 评测方法和标准

拼音标注是基于分词系统的, 分词的优劣程度可直接反映标注结果的好坏, 但是我们最终关注的

是拼音标注的结果, 而不关心切分出的词是否符合自然语言的语义和语法规则. 所以评测标准可描述为以下形式:

$$\text{准确率: } \alpha = \frac{\text{正确的多音字标注结果数}}{\text{所有的多音字标注结果数}} \times 100\%$$

用不同容量大小的语料库训练最大熵模型, 最终对同一段内容进行分词标注, 观察上述的评测指标. 查看训练语料库的大小对最终的标注结果产生的影响. 我们分别选用了包含 5 MB、7 MB、10 MB、12 MB 大小的语料库文本对系统进行训练, 测试所得的实验结果如表 1 所示

表 1 不同容量大小语料库实验对比结果

语料库容量评测指标/MB	标注准确率/%
5	94.1
7	96.4
10	96.6
12	96.6

实验表明, 随着训练库的容量增加准确率会有所提高, 但是相差不是很大, 但是 10 MB 容量的训练库已经基本满足实用需求, 大于 10 MB 容量的训练库已经没有变化.

4.2 评测数据与结果

使用 Java 语言进行算法实现, 实现了汉语拼音转换的软件. 在该软件中输入“读者文摘”原创版第十一卷合订本总第 21-22 期一篇文章“藏了 66 年的甜蜜爱情”, 实验对比结果如表 2 所示.

表 2 标注对比实验结果

实验名称	多音字出现次数	错误次数	错误率
单字标注拼音	842	194	23.0%
基于分词标注拼音	842	28	3.3%

5 结束语

用计算机自动为文本进行高质量标注拼音, 在拼音读物排版系统、对外汉语教学, 特别是中小学和幼儿语文教学中具有实际意义. 汉语拼音标注属于自然语言信息的处理, 本文就其中实现的原理, 算法进行了研究, 通过优化改进分词算法, 提高了标注拼音的准确度. 但是仍然存在错误标注的情况, 将来继续优化构造特征算法, 来进一步提高准确度.

(下转第 126 页)

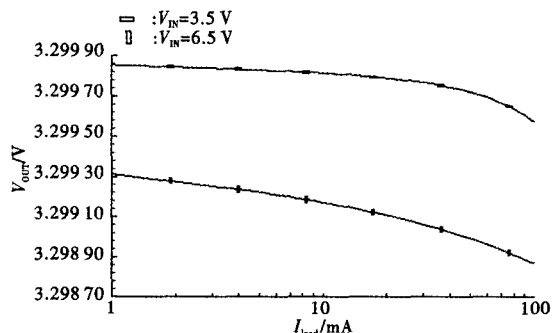


图7 LDO 的负载调整率

输出电压的过冲,从而很好的改善了 LDO 的负载瞬态响应. 由于第二级采用的是低输出电阻的 buffer,这样既可以把第二个极点推得很远,节省了补偿电容的面积,也能够很好地驱动调整管的栅极大电容,减少了 LDO 的静态功耗. 仿真结果显示,线性调整率和负载调整率分别为 0.023% 和 0.5% 左右,达到了期望的值.

参考文献:

- [1] 宁志华,何乐年,刘磊. 一种低功耗、高稳定性的无片外电容线性稳压器[J]. 电路与系统学报,2009,4(2):84-

(上接第 122 页)

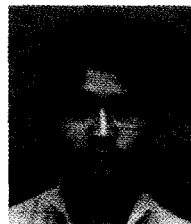
参考文献:

- [1] 罗万,彭泽润. 高浓缩汉语拼音拼写方案及其原理[J]. 北华大学学报,2006,(6):1-10.
- [2] Ma Zhiqiang, Su Yila, Ma Yao. The study of chinese dictionary mechanism based on the suboptimal search tree[C]// The International Conference on Computer and Automation Engineerin (ICCAE 2010). [S. L]: ACM Press,2010:65-69
- [3] 马志强. 一种给批量汉字加注有声调拼音的方法[J]. 微电子学与计算机,2008,25(4):185-187.
- [4] 梁南元. 汉语计算机自动分词知识[J]. 中文信息学报,1990(2):29-33.
- [5] 郭翠珍,朱巧明,李培峰,等. 基于信息网格的分词服务的研究 [J]. 微电子学与计算机,2006,23(5):121-123.
- [6] 冯尚友. 信息熵与最大熵原理[J]. 水利电力科技,1995,22(3):24-29.
- [7] 刘培玉,贾丽洁. 基于最大熵模型的分词技术研究[J].

89.

- [2] Jun Choi,Jungeui Park,Wooju jeong,et al. Design of LDO linear regulator with ultra low-output impedance buffer [J]. IEEE J Solid-State Circuits,2009,37(9):1323-1972.
- [3] Sansen, Willy M C, 陈莹梅. 模拟集成电路精粹[M]. 北京:清华大学出版社,2008.

作者简介:



杨洁 男,(1987-),硕士研究生. 研究方向为 CMOS 模拟集成电路设计.

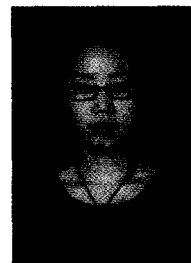


曾云 男,(1956-),教授,博士生导师. 研究方向为新型和高性能电子器件、电子器件应用、CMOS 图像传感器.

计算机软件与理论,2007(4):18-21.

- [8] 张仰森,曹元大,俞士汶. 最大熵方法中特征选择算法的改进与纠错排歧[J]. 北京理工大学学报,2006,26(1):36-40.
- [9] 宋国杰,唐世渭,杨冬青,等. 基于最大熵原理的空间特征选择方法[J]. 软件学报,2003,14(9):1544-1550.
- [10] Yu Shiwen, Duan Huiming, Zhu Xuefeng, et al. The specification of basic processing of contemporary chinese Corpus [J]. Journal of Chinese Information Processing, 2002.

作者简介:



张丽青 女,(1984-),硕士,研究生. 研究方向为人工智能、中文分词.