基于 HMM 的汉语整句拼音输入法研究 *

贾剑峰 , 史晓东 , 赖兴邦

(厦门大学信息科学与技术学院人工智能所,厦门 361005)

摘 要:介绍了一种基于 HMM 的汉语整句拼音输入转换为整句汉字的输入法,提出了引入语言知识后的一种音字选择方法,并给出了采用 N 元拼音文法时的选择模型。实验表明,该方法取得了较好的效果。

关键词:智能拼音输入法;隐马尔可夫;N 元模型

0 引 言

拼音输入法从输入的基本单位上又可以分为字、词、语句三级[2],字的输入技术特点是以字为输入单位,其主要问题在于汉字中多音字或近音字的问题使得重码很多,候选字多,影响输入速度;词级的输入技术是以汉语中词的单位来匹配拼音,这种输入方式符合人写作的习惯,常见的"联想"输入方式也可以归为此类,其实现和字一级的输入法相比效果提升不明显,其主要原因还是近音词识别的问题;基于语句的输入技术是以短语、句子为整个输入单位来进行转换,从操作心理来讲,操作人员一般是以具有一定意义的短语或短句作为持续的输入,其上下文具有很强的信息关联;从信息论的角度来讲,汉语的多维熵小于一维熵,因此候选词比字、词一级的候选表要少的多。

1 HMM 相关描述

HMM 是一种研究时间序列的随机方法,其描述的随机过程是状态转换和观察值产生的双重过程[3]。

1.1 Markov 链

Markov 链是 Markov 随机过程的特殊情况,数学上可以给出如下的定义:随机序列 Xn ,在任一时刻 n, 它可以处在状态 θ1,...,θN,且它在 m+k 时刻所处的状态为 qm+k 的概率,只与它在 m 时刻的状态 qm 有关,而与 m 时刻以前它所处状态无关,即有:

$$P(X_{m+k}=q_{m+k}|X_m=q_m,\cdots X_1=q_1 = P(X_{m+k}=q_{m+k}|X_m=q_m)$$

每一个状态之间都有转移概率,当上式中 k 为 1 时,称为一步转移概率,简称为转移概率,记为 aij,即有转移概率矩阵:

$$A=\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \cdots & & \cdots \\ a_{n1} & \cdots & a_{mn} \end{bmatrix} \& \begin{bmatrix} 0\leqslant a_{ij}\leqslant 1 \\ \sum_{j-1}^{n} a_{ij}=1 \end{bmatrix}$$

1.2 HMM 模型

在 Markov 链模型中,观察值和状态一一对应,在 HMM 模型中,观察值和状态不是一一对应的,观察者只能看到观察值,不能直接看到状态。

HMM 中的元素

N :模型中 Markov 链的状态数目

M :每个状态对应的所有可能出现的观察值的数目

π :初始状态概率矢量,其中:

πi = P(q1 =θi),1≤i≤N

A:状态转移概率矩阵

B:观察值概率矩阵

在拼音到汉字的识别过程中,汉字和该字的读音之间隐含着对应关系,即状态和对应的观察值之间的隐含关系;汉字之间的语言统计关系对应为 HMM 中状态之间的转移关系;求给定拼音指出汉字的过程即为 HMM 中给定观察序列得出状态序列的过程,因此,HMM 模型适应于描述拼音到汉字的转化过程。

1.3 改进的 HMM 模型

在 HMM 中每个状态 S_i 生成观察值 O_i 的生成概率 $P(O_i|S_i)$ 是独立的,而在一些情况下,独立性假设并

* 基金项目:863 项目(No.2006AA01Z139)、福建省重点科技项目(No.2006H0038)、福建省基金项目(No.2006J0043)
收稿日期:2007-12-18 修稿日期:2008-03-16

作者简介:贾剑峰(1984-),男,新疆伊宁人,硕士研究生,研究方向为自然语言处理

不准确，比如说在出现“生长”一词时，“长”读“zhang”的概率应该与“生”字有关，故将 HMM 中的生成概率改为 $P(O_i|S_{i-1}S_i)$的二元生成概率，通过上文的影响来修正生成概率，从而选择更加准确的序列。

2 音字转换语言理解模型

本文的主要研究模型：

$S=(S_1,S_2,\cdots S_n)$ 为输入文本的一串拼音字符序列，即观察值序列；

$W=(w_1,w_2,\cdots w_n)$为模型的输出汉字序列，即状态值序列，其中对于每个 $S_i\rightarrow W_i\in C_i, C_i=\{C_{ij}|C_{ij}$ 为 W_i 候选字$\}$，若$|C_i|$=m 则有 m^n 种的候选汉字序列，要从中选择出最符合汉语语言规律的一种排列即：

$$W=\arg\max P(W/S)$$

这是典型的 HMM 中选择最佳状态转移序列问题：

状态转移概率矩阵：$A=(a_{ij}), a_{ij}=(W_j|W_i)$

观察值概率分布：$B=(b_j(k)), b_j(k)=P(S_k|W_j)$

初始状态：$\pi=(\pi_i), \pi_i=P(q1|Wi)$

基于 HMM 的拼音识别过程如下：

$$W=\arg\max P\{(W|S)\}=\arg\max\{P(S/W)P(W)\}$$

其中第一项是语音模型，第二项是汉字的语言模型。

具体分析如下：

(1)N-gram 语言模型

根据概率成绩公式：

$$P(W)=\prod_i P(W_i|W_1,W_2,\cdots,W_{i-1})$$

$P(A_i|A_1,A_2,\cdots,A_{i-1})$表示一句话中前 i-1 个拼音是 $A_1,A_2,\cdots,A_{i-1}$ 第 i 个拼音是 A_i 的概率，该式是准确的概率值，一般情况下用 N-gram 模型来估算，即第 i 个拼音只于前 N-1 个拼音有关：

$$P(W)=\prod_i P(W_i|W_{i-n-1},W_{i-n-2},\cdots,W_{i-1})$$

在统计语料库中假设$(w_1,w_2,\cdots,w_{n-1},w_n)$出现次数为 $C(w_1,w_2,\cdots,w_{n-1},w_n)$，$w1,w_{n-1}$ 出现的次数为 $C(w1,w_{n-1})$，则 N 元组的相对频率为：

$$f(w_n|w_1,w_2,\cdots,w_{n-1})=\frac{C(w_1,w_2,\cdots,w_{n-1},w_n)}{C(w_1,\cdots,w_{n-1})}$$

根据统计知识，在语料库足够庞大的前提下有：

$$P(W_n|W_1,W_2,\cdots,W_{n-1})=f(w_n|w_1,\cdots,w_{n-1})$$

在实际运用中，由于语料库总是有限的，在某些频率上会出现 0 概率，采用平滑算法将概率给予补偿，这方面有很多的平滑方法，本文采用 SGT[4]来进行数据概率的平滑，即：

$$r^*=(r+1)\frac{n_{r+1}}{n_r}$$

$$P_{GT}=\frac{r^*}{N}$$

n_r=出现 r 次的 n 元组数目

N=所有 n 元组数目

SGT 中为了避免对出现次数很多的 n 元组无法采用 GT 的情况，用一条曲线 S 来拟合(r,Nr), (Nr=arb (b<-1),)

取 N 为不同的值可以得到不同的模型，然后利用简单的线性插值法将不同的模型整合：

$$P(w_n|w_{n-2},w_{n-1})=\lambda_1P_1(w_n)+\lambda_2P_2(w_n|w_{n-1})+\lambda_3P_3(w_n|w_{n-1},w_{n-2})$$

其中：

$$\sum_{i=1}^{n}\lambda_i=1$$

(2)语音模型

求解语音模型即为求解 $P(S_k|W_j)$的概率问题。根据原始的 HMM 模型，每一个 $P(S_k|W_j)$是相对独立的有：

$$P(S|W)=P(S_1,\cdots,S_n|W_1,\cdots,W_n)=\prod P(S_i|W_i)$$

本文采用二元的生成概率模型，如上文所述，有：

$$P(S|W)=P(S_1,\cdots,S_n|W_1,\cdots,W_n)=\prod P(S_i|W_{i-1}W_i)$$

对于每一个可以从语料库中进行统计，由于缺少相关的拼音汉字对照语料，本文采用对词的拼音匹配来自动生成语料。

汉字中多音现象相当普遍，但词的多音却不常见，而且分词系统的准确率已经达到 97%以上，因此，对经过分词之后的文本进行词的拼音匹配有较高的准确率[5]。

由此，可建立整个理论模型。

3 实验具体实现

利用上述的理论模型对非特定的拼音序列进行实验，由于有调拼音需要的语料库庞大，因此本实验针对无调拼音进行相关的实验。

3.1 实验数据

实验语料库采用《人民日报》2000 年 63M 的语料

库,针对二元、三元模型进行了训练,并调用史晓东老师的 segtag 程序进行分词并标注拼音,得到相应的音字语料库。

测试语料为截取 1999 年《人民日报》50 句以及《读者》50 句内容。

二元模型直接采用 SGT 估算, 三元模型采用线性插值即:

$$P(w_n|w_{n-2},w_{n-1})=P_1(w_n)/6+P_2(w_n|w_{n-1})/3+P_3(w_n|w_{n-1},w_{n-2})/2$$

三元模型的比重较大。

另外,采用后向的最大分词匹配输入串,减少输入量,对于有固有奇异性的拼音用空格或'符号将可能出错的词对分割开来。

3.2 实验结果

实验中由准确率来评价系统的性能:

$$\text{准确率}=\frac{\text{正确的输出字数}}{\text{总的输出字数}}$$

表 1

	二元	三元模型	传统 HMM 三元模型	改进 HMM 三元模型
封闭测试	0.8122	0.8470	0.8692	0.9040
开放测试	0.6820	0.7229	0.7531	0.7566

正确输入输出例子:

jintianxiawuwocanjialeyichangkaoshi

今天下午我参加了一场考试

nanhangkeyunliangweijudiyi

南航旅客运输量位居第一

错误的举例:

huiyoufagengdadeshigu

会有发更大的事故

lieningyouyubujue

列宁由于不绝

对比文献 6 中的效果:

表2

候选数	1	2	3
Nogram	80.17%	91.37%	94.62%
Unigram	83.46%	92.63%	95.64%
前向 Bigram	85.08%	93.74%	96.56%
Bigram	88.02%	95.11%	97.33%
前向 Trigram	87.88%	94.86%	97.09%
Trigram	92.77%	96.90%	98.14%

3.3 实验分析

通过对比二元和三元的效果可知,三元模型对于输入的语料还是具有较好的适应性的, 通过笔者实验,三元模型对于成语(四字)的输入和二元的比较效果明显;

有些词的二元效果比三元的效果好,究其原因还是因为语料库中三元组比二元组稀疏的缘故,如果考虑增大语料库,可能会包括较多的三元组,但也有可能加入更多的二元组,这也是 N-gram 的一个特点;

对于改进的 HMM 模型,可以在一定程度上解决多音字的问题,但由于大多数的汉字与拼音之间还是一对一的关系,即生成概率为 1,因而转移概率还是影响效果的主要因素, 所以在测试期间发现传统 HMM 和改进的 HMM 中出错的地方基本一致。

4 结 语

通过将拼音识别问题转换为 HMM 模型问题来解决, 得到了一种基于整句的统计音字转换模型,通过测试具有一定的可行性。为解决语音模型的统计问题, 采用二元的生成概率来改善多音字的发音概率,得到了优化的总体结果, 但也有改差的个体结果,得出生成概率对于 HMM 的估算还是有相当的意义的。

通过结果比较,如果从更加准确的音字语料的角度出发来估算生成概率,以得到较好的统计模型;

我们利用的 HMM 模型只是单纯的一层统计模型,如果能将句法层次的 HMM 模型再加入进来得到一个多层的 HMM 可以预计其效果肯定更好。

比较文献[6]可知,增加候选词可以有效的提高命中率,因此,对于 HMM 还可以用多元的查找来进行。

参考文献

[1]陈原主编. 汉语语言文字信息处理.上海教育出版社,1997

[2]王晓龙,王幼龙. 语句级汉字输入技术. 中文信息学报, 1996,10(4):50~59

[3]李元祥,丁晓青,刘长松. 基于 HMM 的汉语文本识别后处理研究. 中文信息学报, 1999, 13(4): 29~34

[4]William A. Gale and Geoffrey Sampson, Good-Turing Frequency Estimation Without Tears, Journal of Quantitative Linguistics, 1995.Volume:2

[5]陈正,李开复. 拼写纠正在拼音输入法中的应用. 计算机学报,2001,24(7)

[6]杨浩荣,王作英,陆大. 汉语语音识别中的拼音多候选问题,电子学报,1999,27(4):58~62

(下转第 19 页)

参考文献

[1]黄拙,张健. 由一阶逻辑公式得到命题逻辑可满足性问题实例[J]. 软件学报,2005,16(3):329~335

[2]张德富,黄文奇,王厚祥. 求解 SAT 问题的拟人退火算法[J]. 计算机学报,2002,25(2):148~152

[3]贺毅朝,王彦祺,寇应展. 一种求解 3-SAT 问题新方法. 计算机工程与应用,2006,42(16) 70~72

[4]徐宗本. 计算智能—模拟进化计算. 北京:高等教育出版社,2005

[5]贺毅朝,王熙照,寇应展. 一种具有混合编码的二进制差分演化算法. 计算机研究与发展,2007,44(9),1476-1484。

[6]刘勇,康立山等. 非数值并行算法(二)—遗传算法. 北京:科学出版社,2003

[7]耿素云,屈婉玲,王捍贫. 离散数学教程. 北京大学出版社,2002

[8]康博创作室. Visual C++6.0 高级编程. 北京:清华大学出版社,2000

Solving Satisfiability Problem Based on Modified Genetic Algorithm

CAO Guo-sheng , HE Yi-chao , LI Min , WANG Zhi-gong , WEI Tian-xing

(Information Engineering School, Shijiazhuang University of Economics, Shijiazhuang 050031)

Abstract: Genetic algorithm (GA) is a kind of intelligence algorithm based on the theory of evolution proposed by Holland, which is suitable for solving optimization problem. For solving the Satisfiability problem (SAT) using SGA, combines SGA with Local Search Algorithm, proposes a modified hybrid genetic algorithm (MHGA) to solve 3-SAT. First, transformes 3-SAT to the computation problem of polynomial function on $\{0, 1\}^n$, whether exists a zero point. The results of numerical computation show that MHGA is feasible and effective for solving hard SAT.

Keywords: Genetic Algorithm; Satisfiability Problem; De Morgen Law; Local Search

(上接第 6 页)

Research on HMM-Based Chinese Whole Sentence Pinyin Input

JIA Jian-feng , SHI Xiao-dong , LAI Xin-bang

(Institute of Artificial Intelligence of Information Science and Technology College, Xiamen University, Xiamen 361005)

Abstract: Introduces a HMM-based Chinese Pinyin input method which takes whole Chinese pinyin sentence into the characters, improves the selection of characters using language knowledge and gives a model using N-gram Pinyin grammar. The experiments show that the method of achieving better results.

Keywords: Intelligent Pinyin Input Method; Hidden Markov Model; N-gram Model