一种基于词树的高效解码算法

张志强[1],张太红[1,2],董 峦[1,3]
(1. 新疆农业大学 计算机与信息工程学院,新疆 乌鲁木齐 830052;
2. 中国农业大学 信息与电气工程学院,北京 100083;
3. 河海大学 计算机与信息工程学院,江苏 南京 210098)

摘 要:音字转换是汉语言信息处理的一个重要方面,在语音识别、汉语拼音输入等方面都有广泛的应用。为了找到一种行之有效的音字转换解码算法,在研究拼音分词与词树理论并分析词树求解过程的基础上,提出了基于语言模型实现音字转换的高效解码算法。该算法采用零概率重估、路径剪枝和多音字处理等多项技术,通过对词树进行的剪枝处理、对常用词的处理以及对解码过程中所产生多音字的处理,实现了普遍意义上的音字转换。为验证所提算法的有效性和可行性,基于新疆维吾尔自治区科技计划项目《多语种民族特色文化信息资源处理及共享服务平台》所提供的三组数据进行了对比实验。实验结果表明,提出的新算法取得了 97.78% 的转换准确率,优于其他传统算法。

关键词:拼音分词;词树;语言模型;n-gram 模型;音字转换
中图分类号:TP391.1 **文献标识码**:A **文章编号**:1673-629X(2017)08-0043-04
doi:10.3969/j.issn.1673-629X.2017.08.009

An Efficient Decoding Algorithm Based on Word Tree

ZHANG Zhi-qiang[1], ZHANG Tai-hong[1,2], DONG Luan[1,3]
(1. College of Computer & Information Engineering, Xinjiang Agricultural University, Urumqi 830052, China;
2. College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China;
3. College of Computer and Information Engineering, Hohai University, Nanjing 210098, China)

Abstract: Phonetic conversion is an important aspect of Chinese language information processing, which has been widely used in speech recognition, Chinese Pinyin input and so on. In order to find an effective syllable-to-character decoding algorithm, an efficient decoding algorithm is proposed based on the study of phonetic word segmentation, the word tree theory and the analysis of word tree solving. It uses zero probability reassessment, path pruning, processing of polyphonic words to realize the syllable-to-character conversion generally by pruning of word tree, processing of common words and processing of polyphonic words in the decoding process. In order to verify the validity and feasibility of the proposed algorithm, the contrast experiments on three sets of data provided by Xinjiang Uygur Autonomous Region Science and Technology Program, Multilingual Ethnic Cultural Information Resource Processing and Sharing Service Platform, have been conducted. The experimental results show that it has achieved 97.78% conversion accuracy, which is superior to other traditional algorithms.

Key words: phonetic word segmentation; lexicon tree; language model; n-gram model; Pinyin-Chinese character transform

0 引 言

语言模型(Language Model, LM)[1]是语音识别系统(Speech Recognition System, SRS)[2]的一个重要组成部分。语言模型,一般分为以统计为基础的统计语言模型(Statistical Language Model, SLM)和以规则为基础的规则语言模型(Rule-based Language Model, RLM)。在现有条件下,SLM 处于主流地位,通过对大量语料统计[3],获得词与词之间的连接信息,为评价一个词串是否有意义提供依据。

n-gram 语言模型是统计语言模型中比较典型的模型[4],它的结构简单,易于构建和应用。但是,应用 n-gram 语言模型时,需要解决训练语料稀疏而引起的

收稿日期:2016-07-26 **修回日期**:2016-10-27 **网络出版时间**:2017-07-05
基金项目:新疆维吾尔自治区科技计划项目(2015X0106)
作者简介:张志强(1986-),男,硕士研究生,研究方向为数据库技术;张太红,博士,教授,硕士生导师,通讯作者,研究方向为数据库技术、农业信息化技术。
网络出版地址:http://kns.cnki.net/kcms/detail/61.1450.TP.20170705.1650.026.html

零概率问题[5]。为了解决该问题,提出了一种基于词树的音字转换算法,通过拼音分词,对词树进行搜索和剪枝,对常用词以及对多音字进行处理。

1 拼音分词理论

为了提供更为准确的词特征,在此利用拼音分词。拼音分词的任务就是把通过键盘输入的汉语拼音串,切分成拼音词单元。例如“zhong guo ren min yin hang”可以切分为“zhong guo/ ren min /yin hang”。

拼音分词采用的是 Trigram 模型并且采用绝对平滑(Absolute Smoothing)算法。和汉语分词过程相似,构造的拼音网格如图 1 所示。

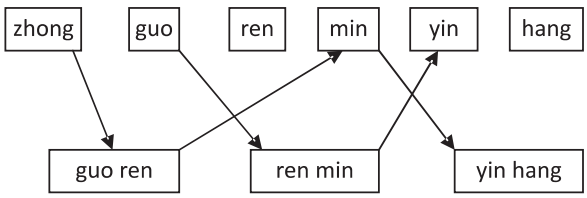

图 1 拼音分词网格

拼音分词就是指在图 1 的网格中搜索出最优切分路径。在 Trigram 模型中,可寻找满足式(1)的切分:

$$S^* = \arg\max_s \prod_{i=1}^{n} p(p_i \mid p_{i-2}p_{i-1}) \tag{1}$$

其中,$p(p_i \mid p_{i-2}p_{i-1})$ 表示拼音串中首个拼音词出现的概率。

和音字转换相比,汉字到拼音转换过程比较简单,所以大规模获取拼音汉字转换的语料也较为容易。利用大规模拼音分词语料单独训练拼音分词模型,同时利用这个模型对音字转换模型的训练语料和测试语料进行重新切分。训练与测试均采用相同的系统处理,这样可以尽量弥补切分错误带来的影响。

2 词树理论

音字转换,就是利用语音识别中的语言模型,并用一定的解码算法进行处理,将一串没有音调的拼音串 $S = S_1S_2\cdots S_t$,转换为词串(或语句) $W = W_1W_2\cdots W_m$[6],或字串 $C = C_1C_2\cdots C_t$。其中,$W_i = C_{i1}C_{i2}\cdots C_{il} = C_{il}^{i1}$[7](用 C_{il}^{i1} 表示 $C_{i1}C_{i2}\cdots C_{il}$,下同)。解码过程,就是寻找最好词串 W^* 的过程[8],使得:

$$W^* = \arg\max_w P(W \mid S) = \arg\max_w P(W)P(S \mid W) = \arg\max_w P(W)\prod_{i=1}^{m} P(S_{k(i)}^{l(i)} \mid W_i) \tag{2}$$

其中,

$$P(S_{k(i)}^{l(i)} \mid W_i) = P(W_i \mid S_{k(i)}^{l(i)}) = \begin{cases} 1, W_i = C_{k(i)}\cdots C_{l(i)} \text{ 且 } C_j \in \mathrm{dhz}(S_j), j = k(i)\cdots l(i) \\ 0, \text{其他情况} \end{cases} \tag{3}$$

其中,$\mathrm{dhz}(S_j)$ 表示与 S 相对应可能汉字的集合。

由于在汉语中存在着许许多多的同音字词,求解式(2)的过程是在一个比较繁琐的词树上进行的,词树的路径数量因 S 的不同而不同。例如,当 S = “zhong guo ren min yin hang”时,它产生的词树可能如图 2 所示。

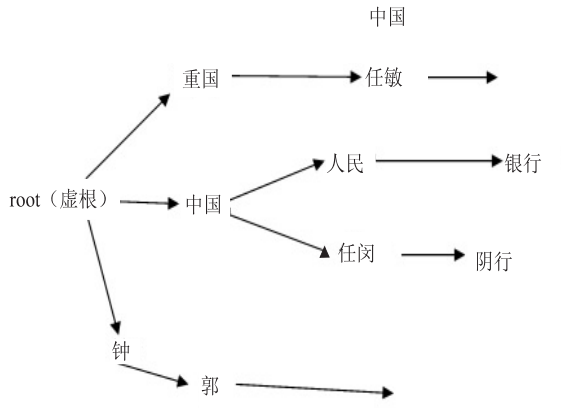

图 2 词树

式(2)的求解就是在图 2 中找出一条使 $P(W)$ 最大的路经。

3 求解 $P(W)$ 的过程

n-gram 模型是基于如下假设:第 n 个词只和它前面的 $n-1$ 个词有关,而和其他的词没有关系[9]。根据上面的假设,可以把 W_n 出现的概率写成如下条件概率的形式:$P(W_n \mid W_1^{n-1})$[10],此处的 W_1^{n-1} 代表词串 $W_1W_2\cdots W_{n-1}$(下同)。假设一个句子是由 m 个词构成,那么这个句子的先验概率[11]可以表示为:

$$P(W) = P(W_1)P(W_2 \mid W_1)\cdots P(W_{n-1} \mid W_1^{n-2})\prod_{i=n}^{m} P(W_i \mid W_{i-n+1}^{i=1}) \tag{4}$$

如果有非常多的语料作保证,那么可以根据最大似然度规则得到:

$$P(W_i \mid W_{i-n+1}^{i-1}) = \frac{C(W_{i-n+1}^{i})}{C(W_{i-n+1}^{i-1})} \tag{5}$$

其中,$C(W_{i-n+1}^{i})$ 和 $C(W_{i-n+1}^{i-1})$ 分别表示词串 $W_{i-n+1}\cdots W_i$ 和 $W_{i-n+1}\cdots W_{i-1}$ 在训练样本中出现的次数。

实际应用 n-gram 模型时,n 一般取得很小,目前最常用的是 $n=3$,称 3 元模型。这时式(4)和式(5)可以分别写成:

$$P(W) = P(W_1)P(W_2 \mid W_1)\prod_{i=3}^{m} P(W_i \mid W_{i-2}^{i=1}) \tag{6}$$

$$P(W_i \mid W_{i-n+1}^{i-1}) = \frac{C(W_{i-2}^{i})}{C(W_{i-2}^{i-1})} \tag{7}$$

假设与模型相配的词库大小是 10 000,则 3 元词串的所有组合可以达 $10\ 000^3$ 项,这是个很大的数目,用 2 亿字来训练这个算法,也将会出现很多 $C(W_{i-2}^{i})$

为 0 的情况,也就是零概率的问题。$C(W_{i-2}^{i})$ 为 0 有以下两种情况:第一种是 W_{i-2}^{i} 在语义上虽然正确,但是没有被统计到;第二种是如果 W_{i-2}^{i} 在语义上就不正确,不可能在语料中出现 3 元词串。对于前者而言,如果按照式(7)计算它的概率,将会导致不可靠的结果,因此,图灵估计概率的算法就在这种情况下产生了,但是,图灵算法的条件苛刻[11],在现在的计算机技术条件下,产生的语言模型想要满足其条件,是非常困难的;基于这种原因,产生了许许多多“修正”、“退化”的派生化算法,从而能获得较好的结果[12]。但是,“修正”、“退化”这些操作将使之丢失一部分信息。用不同于上述算法的方法来解决这个问题,且不丢失任何信息,效果比较理想。具体而言,式(6)中各项采用如下算法:

$$P(W_1) = C(W_1)/N \tag{8}$$

$$P(W_2 \mid W_1) = \begin{cases} C(W_1^2)/C(W_1), C(W_1^2) > 0 \\ \alpha C(W_2), C(W_1^2) = 0 \end{cases} \tag{9}$$

$$P(W_i \mid W_{i-2}^{i-1}) = \begin{cases} C(W_{i-2}^{i})/C(W_{i-2}^{i-1}), C(W_{i-2}^{i}) > 0 \\ \beta C(W_{i-1}^{i})/C(W_{i-1}), C(W_{i-2}^{i}) = 0 \text{ 且 } C(W_{i-1}^{i}) > 0 \\ \alpha C(W_i), C(W_{i-2}^{i}) = 0 \text{ 且 } C(W_{i-1}^{i}) = 0 \end{cases} \tag{10}$$

其中,N 为训练语料总词数;α 、β 为两个经验数据。实验表明:α 取 10^{-11}、β 取 10^{-3} 是一组合适取值。

4　对词树的剪枝处理

如果在图 1 的词树中求解式(2),随着音节串 S 的增长,路径数目将会迅速膨胀,当音节数目大于 10 的时候,路径数目将会达到上千条,如果不增加剪枝技术,时间复杂度和空间复杂度是无法容忍的[13]。利用剪枝技术,把路径搜索限制在有限的范围内,是整个算法不可缺少的部分[14]。在上千条路径中,期望(正确)的路径只有一条,其他都是多余的,所以理想的剪枝技术应当是:

(1)不会发生错误剪枝;

(2)尽量多地剪去不是所期望的路径[15]。

为了实现剪枝功能,定义如下的数据结构,用来记录相关信息[16]:

```
#define ITM 16
struct tab
{
node * point;  /* 指向叶子节点 */
int d;  /* 路径达到拼音串的具体位置 */
float loggl;  /* 从根到叶的概率乘积的对数值 */
} path [ITM];
```

整个搜索算法都包含剪枝技术,描述如下[17]:

(1)path[ ]对路径初始化;

(2)生成词树的第一层上节点;

(3)从 S 取出 $S_1 \sim S_l$ (l 最大为 4);

(4)按照可能的词进行组合,按照 1 ~ 4 字词的规模生长;

(5)取 6 条概率较大的路径并且将它们存入 path[ ],并且先按 d 进行排序,后按照 log_gl 进行排序;

(6)进行循环处理;

(7)生长;

(8)从 path[ ]中取出 d 最小的路径,并取出 $S_d \sim S_{d+1}$ (l 最大为 4);

(9)按步骤(4)-(5)行算法扩展后一层节点;

(10)剪枝;

(11)对于相同的 d ,保留概率大的两条路径,其他全部剪去;

(12)当第二条的 log_gl 比第一条小 2(小 100 倍)时,亦剪去;

(13)判结束;

(14)若所有路径到达 t (词串尾);

(15)则{期望路径 = $\arg\max_{path[\,]} \log\_gl$;

(16)输出期望路径;

(17)转(20);

(18)}

(19)否则转(6);

(20)结束。

系统任何时候最多保留 8 条路径。

5　对常用词的处理

在不同领域有着不同的常用词。例如,计算机领域的常用词如图 3 所示。

计算机常用词
中央处理单元、主板、随机存储器(内存)、只读存储、监视器、键盘、鼠标、芯片、光盘驱动器(光驱)、硬盘、软盘、光盘刻录机、集线器、调制解调器、即插即用、不间断电源、基本输入输出系统、安装、卸载、向导、操作系统

图 3　计算机常用词表

为了提高音字转换速度,可以建立一个或多个常用词表。例如对有关计算机的语音进行音字转换解码,可以关联计算机常用词表。一旦通过前面的解码处理得到“中央”二字,可以查询计算机常用词表,那么“中央处理器”的概率肯定是很大的。通过此方法,在一定程度上可以有效提高解码的速度。

6　对多音字的处理

汉语中不仅含有大量同音字,而且含有不少多音字[18],如:“长”,有时念 chang(如“长度”),有时念

zhang(如“长大”);“落”,有时念 luo(如“落后”),有时念“la”(如“落下”),等等。在音字转换中,如果不解决这个问题,有时会造成不可逆转的错误,不仅出错音节所对应的汉字会出错,而且还会影响前后一大串,从而对整句造成灾难性的后果[19]。如:音节串“wo men dou shi zhong guo ren”,得到的正确结果是:“我们都是中国人”。此句第 3 个音节“dou”是多音字,如果变为“du”,那么由于该音节错了,破坏了句中相关词之间的连接关系[20],于是产生如下错误结果:“我们毒誓种过任”。

对多音字处理的方法是:在解码的同时给多音字增加一个候选项。其中,如果把“dou”念成“du”,系统除了按照原来的音节串进行检索外,还会自动将“du”替换成“dou”再次检索一遍词树,重复检索的范围是 $S_{i-3} \sim S_{i+3}$ (i 是多音字的下标)。根据从语言模型中词的先后连接信息所得概率,系统会自动判别应该取“dou”这个读音,还是取“du”这个读音,以便获得正确的结果,所以具有对多音字的容错能力,实际操作表明,这是一种非常有效的方法。

7 实验测试

7.1 实验设备

实验设备配置见表 1。

表 1 实验设备配置表

配置	型号
电脑型号	联想 Lenovo G480 20149 笔记本电脑
操作系统	Windows 7 旗舰版 64 位 SP1 (DirectX 11)
处理器	英特尔 第二代酷睿 i5-2520M @ 2.50 GHz 双核
主板	联想 31900005WIN8 STD PRC (英特尔 HM76 Express)
内存	8 GB
主硬盘	三星 SSD 750 EVO 250 GB(250 GB/固态硬盘)
显卡	英特尔 HD Graphics 3000(64 MB/联想)
显示器	奇美 CMN1470(13.6 寸)

7.2 词 库

词量有 50 000 条,0 ~ 39 号是常用的全角符号,40 ~ 50 000 号是汉字词条,长度为 1 ~ 4 字,以 2 字词居多。

7.3 语言模型

由 2 亿字的中文语料训练形成,含有 50 本电子书、2 年的人民日报,内容涵盖范围非常广,包含外交、政治、经济、文化、民生等众多领域。

7.4 测试集

新疆维吾尔自治区科技计划项目《多语种民族特色文化信息资源处理及共享服务平台》提供的 3 组数据,共 2 000 句,内容包含政治、外交、体育、民俗、文化和日常生活等方面。

7.5 测试结果

(1)准确性。

句数:2 000;字数:20 000;错字:444;准确率:97.78 %。

(2)转换速度。

4.4 字/s,所使用电脑核心部件配置见表 2。

表 2 核心配置

配置	型号
处理器	英特尔 第二代酷睿 i5-2520M @ 2.50 GHz 双核
内存	8 GB
主硬盘	三星 SSD 750 EVO 250 GB(250 GB/固态硬盘)

这个核心配置只能算是计算机的中等配置水平,因此导致计算机的运算速度不够高,如果提高计算机配置,音字转换的速度势必大大提高。

8 结束语

音字转换包括两个重要指标:准确率和转换速度。准确率与零概率重估算法、剪枝技术、多音字处理等因素存在着密切的联系。在诸多因素中,零概率重估算法是最重要的一项。基于以上原因,提出了以语言模型为基础的音字转换算法,并将算法应用于仿真系统。对词树进行搜索和剪枝,随后对常用词、多音字进行处理,得到的准确率达到 97.78%。仿真实验表明:该算法具有很好的有效性和可行性。引入 α、β 两个参数来计算概率并处理零概率事件,使转换速度达到 4.4 字/s,满足了实时处理要求。若能提高计算机性能,则可以达到更为理想的效果。

参考文献:

[1] 陈雅兰,胡小华,涂新辉,等. 基于位置语言模型的中文信息检索系统的研究[J]. 计算机科学,2015,42(7):265-269.

[2] Aubert X L. One pass cross word decoding for large vocabularies based on a lexical tree search organization[C]//Proc. of Eurospeech'99. [s. l.]:[s. n.],1999:1559-1562.

[3] 任光辉,茅旭初. 多约束条件的全球定位系统单频单历元短基线定向技术与实现[J]. 上海交通大学学报,2014,48(3):335-340.

[4] 李春生. 一种体现长距离依赖关系的语言模型[J]. 科技视界,2014(5):55-56.

[5] Bacchiani M,Ostendorf M. Joint lexicon,acoustic unit inventory and model design[J]. Speech Communication,1999,29(2):99-114.

[6] 艾山·吾买尔,早克热·卡德尔,买合木提·买买提,等. 基于 C#的语言模型计算工具[J]. 电脑知识与技术,2013(33):7590-7592.

[7] Chao Y R. Tone contour[EB/OL]. 1979. http://en. wikipe-

(下转第 51 页)

4　结束语

针对信用评估问题,对已有的影响信用数据进行处理与建模,提出了一种最速下降法的改进方法,能够在建模过程中更高效地运算。另外,将一步转移概率应用到信用的评估预测中,实现了对影响信用数据不足的用户所进行的评估以及对未来一段时间后的用户信用所进行的评估。

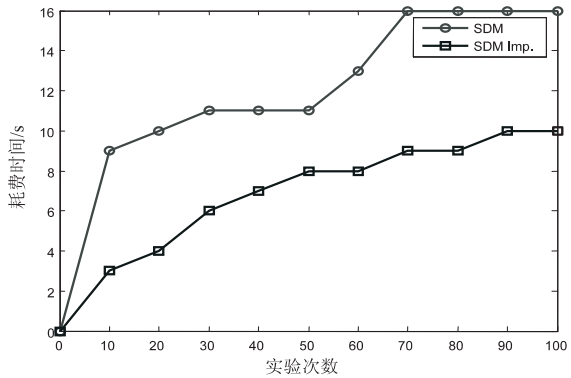

图 4　改进前后的运算时间对比

参考文献:

[1] 陈永明,周　龙,李双红. 基于 AHP 和 DEMATEL 方法的农户信用评级研究[J]. 征信,2012(5):20-24.

[2] 孙玲芳,祁　军,徐　会,等. 面向交易型虚拟社区的信用评价模型研究[J]. 信息技术,2014,38(7):74-77.

[3] Lu Jianchang, Wu Jipeng. The fuzzy comprehensive evaluation on credit risk of power customers based on AHP[C]//Second international symposium on information science and engineering. Shanghai:[s. n.],2009:148-151.

[4] 李俊丽. 基于层次分析法的农户信用评估[J]. 商业研究,2009(10):125-127.

[5] Qiu Y. An importance sampling method based on variance minimization with applications to credit risk[C]//Proceedings of the 29th Chinese control conference. Beijing:[s. n.],2010:3176-3179.

[6] 吴　锋,李秀梅,朱旭辉,等. 最速下降法的若干重要改进[J]. 广西大学学报:自然科学版,2010,35(4):596-600.

[7] 李鸿仪. 理想化最速下降法及其逼近实例[J]. 上海第二工业大学学报,2011,28(1):8-13.

[8] 池光辉,刘建伟,李卫民,等. 权核 Logistic 回归模型的分类和特征选择算法[J]. 计算机工程与应用,2013,49(9):41-44.

[9] 王　鹏,孙继银,郭文普,等. 前视红外目标匹配中的图像质量建模[J]. 计算机应用研究,2012,29(12):4797-4800.

[10] 郑兰祥,万　雪. 基于 Logit 法的我国农村小额贷款公司信用风险评分模型构建研究[J]. 安徽农业大学学报:社会科学版,2014,23(4):49-54.

[11] 姜　盛. 基于 Logistic 的信用卡套现侦测评分模型[J]. 计算机应用,2009,29(11):3088-3091.

[12] Mastin A, Jaillet P. Loss bounds for uncertain transition probabilities in Markov decision processes[C]//51st IEEE conference on decision and control. Maui, HI: IEEE, 2012: 6708-6715.

[13] 冯学伟,王东霞,黄敏桓,等. 一种基于马尔可夫性质的因果知识挖掘方法[J]. 计算机研究与发展,2014,51(11):2493-2504.

[14] Hu Yuting, Xie Rong, Zhang Wenjun, et al. Prediction of tourists flow distribution based on transition probability matrix[C]//8th international conference on information science and digital content technology. Jeju Island, Korea:[s. n.],2012:636-640.

(上接第 46 页)

dia. org/wiki/Tone_contour/.

[8] Cremelie N, Martens J P. In search of pronunciation rules[C]//Proceedings of the ESCA tutorial and workshop on modeling pronunciation variations for automatic speech recognition. [s. l.]:[s. n.],1998:23-27.

[9] 何　莉,林鸿飞. 分布式检索中基于主题的语言模型集合选择策略[J]. 微电子学与计算机,2009(9):78-81.

[10] 刘海娟,张佳骥,陈　勇. 语言模型在话题跟踪中的应用[J]. 无线电工程,2008,38(9):20-23.

[11] 姜　维,关　毅,王晓龙,等. 基于支持向量机的音字转换模型[J]. 中文信息学报,2007,21(2):100-105.

[12] 章　森. 基于混合字词网格的汉语音字转换问题的求解[J]. 计算机学报,2007,30(7):1145-1153.

[13] 李明琴,王作英,陆　大. 语音识别音字转换中的快速容错算法[J]. 中文信息学报,2002,16(5):38-43.

[14] 张瑞强. 关于汉语音字转换中语言模型零概率的问题[J]. 电子学报,1998,26(8):43-46.

[15] 赵以宝,孙圣和. 一种基于单字统计二元文法的自组词音字转换算法[J]. 电子学报,1998,26(10):55-59.

[16] 章　森,宗成庆,陈肇雄,等. 语句拼音-汉字转换的智能处理机制分析[J]. 中文信息学报,1998,12(2):37-43.

[17] 梅　勇,王群生,徐秉铮. 将词类信息融入三元文法统计模型的汉语音字转换方法[J]. 电子科学学刊,1998,20(5):625-630.

[18] 梅　勇,徐秉铮. 一种基于马尔可夫模型的汉语语音识别后处理中的音字转换方法[J]. 中文信息学报,1997,11(4):66-72.

[19] Downey S, Wiseman R. Dynamic and static improvements to lexical baseforms[C]//Proceedings of the workshop on modeling pronunciation variations. [s. l.]:[s. n.],1998:157-162.

[20] 庞春雷,赵修斌,卢艳娥,等. 短基线约束条件下的整周模糊度二维搜索算法[J]. 中国空间科学技术,2012,32(3):43-48.