文章编号: 1007- 757X (2005) 04- 0052- 04

汉语 TTS 系统中多音字问题的一种有效解决方案

刘景勇　①柴佩琪②姚秋明③

摘　要: 多音字现象的存在给汉语 TTS (Text to Speech) 系统增加了难度。本文旨在提出一种解决中文 TTS 系统中的多音字判决问题的统一方案。这种方案基于统计学习的思想。首先构造一个基于特征的词典, 该词典可以根据学习的语料动态更新。在有权值和无权值两种更新词典的方法中, 通过试验对比最终选择了无权值的方法。我们采取建立规则的办法作为对词典的补充, 分别用分类回归树(CART)、扩展的随机复杂度(ESC)进行了实验。通过实验, 最终以 CART 生成的局部规则对词典进行补充, 得到了较为满意的效果。

关键词　文语转换(TTS); 特征词典; 分类回归树(CART); 扩展的随机复杂度(ESC)

中图分类号: TP317　　**文献标识码**: A

1　引言

一字多音在汉语中是很常见的现象, 却没有统一的规则可循。通常我们判断多音字的读音往往是根据学习的经验, 或者是约定成俗的读法。追求中文字音转换的正确率, 是汉语 TTS 系统最基本的要求。否则, 谈自然度也是毫无意义的。

多音字在实际语料中出现的比例是相当高的。我们对试验中的语料进行了统计。总共的近万句句子中, 不同的汉字数有 3571 个, 其中多音字 867 个, 占 24 3%。其中包括语料中出现的部分变调。

为了解决多音字的判音问题, 实现一个字音转换系统, 我们提出了无权值的词典加规则补充的做法。无论对于以词组形式出现的多音字还是以单字形式出现的多音字都能由这种方案加以判定。这里的词典和规则都是对多音字所在的上下文特征的描述。

2　字音转换系统的结构

基于本文提出的方案最终形成一个字音转换系统。字音转换系统在 TTS 系统结构中属于文本分析的阶段。字音转换是直接从文本串 $W = w_1 w_2 \cdots w_n$ 到音序列 $C = c_1 c_2 \cdots c_n$ 的映射, w_i 代表一个字, c_i 代表该字对应的正确拼音。该系统最终采用的即是无权值的词典加规则补充的做法。

最终实现的系统可分为两个部分(如图 1):

上半部分的结构用于词典和规则的学习。大语料中的文本需要先通过分词系统切分, 文本切分的结果是常用词、韵律词和韵律短语, 同时进行人工或半人工拼音标注。这些结果转换成可用于学习的文件格式后, 就可以对词典进行更新和提取规则了。下半部分的结构用于自动标注拼音的过程。未知文本经过分词系统的处理, 由系统读入并进行判决, 最后输出已标注拼音的结构。当然要真正用于 TTS, 还要加入后期变调处理。例如"上声上声"变调成"阳平上声"(舞 wu2 蹈 dao3) 以及"一"、"不"两个字的变调(不 bu2 会, 不 bu4 管)。我们讨论的是通过一个怎样的学习机制来获取好的词典和规则, 在此过程中如何进行方法的取舍。

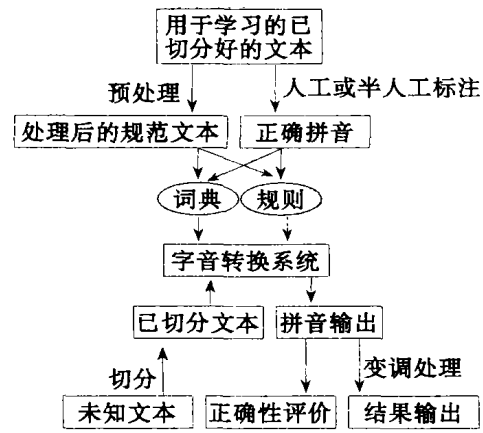

图 1　系统总体结构图

3　基于特征词典的多音判别

这里的特征意指多音字的上下文语言环境。基于统计特征的方法本来是为了解决文本正确性校对的问题, 同时, 特征提取给多音字的判音也带来了一种有效的思路。特征即是词典中的存储内容, 也是规则的描述对象。

我们总共抽取了如下 8 个特征(其中 a、b、d 各包含两项特征):

① 作者简介: 刘景勇, 同济大学计算机科学与工程系, 硕士研究生, 上海　200092
② 柴佩琪, 同济大学计算机科学与工程系, 教授, 博士生导师, 上海　200092
③ 姚秋明, 同济大学计算机科学与工程系, 硕士研究生, 上海　200092

(1) 词内左右邻接字

该特征实际是作为成词情况的一种判断, 用于处理多音字在不同的词语中读不同的音的情况。通式为: $x_{i-1}x_i$ 和 x_ix_{i+1}。x_i 是当前要判断读音的多音字。例如人参(shen1), 参(can1)加。

(2) 左右邻接词

该特征认为多音字在当前词中的读音和左右邻接词是有关的, 通式为: $W_{i-1}W_i$ 和 W_iW_{i+1}, 其中 W_i 表示多音字所在的当前词。这一对属性对多音字单独成词的情况有很好的作用。当多音字的左右词对其读音比较有约束力的时候, 该特征对多音的判断就会比较可靠。比如经过若干次学习后, 我们在"长(chang2)"的左邻接词中发现了"较"、"很"等副词, 这是很符合直观意义的。

(3) 当前词的词性

这个特征的作用很显然。比如"更"作名词时读作 geng1, 作副词时读作 geng4。"数"作名词时读作 shu4, 作动词时读作 shu3。

(4) 左右邻接词的词性

该特征认为当前词的词性和左侧两个邻接词性以及右侧两个邻接词性是有关的, 而当前词的词性又和多音字读音有关。通式为: $C_{i-2}C_{i-1}C_i$ 和 $C_iC_{i+1}C_{i+2}$, 其中 C_i 为当前多音字所在词的词性。比如: "长(chang2)的绳"、"长(zhang3)得美", 前者右邻接词性是助词、名词, 后者右邻接词性是助词、形容词。

(5) 边界条件

这里的边界包括语素, 韵律词, 韵律短语, 以及在句子中的首还是末。该特征对多音的作用主要体现在一些语气助词上面。比如"了"在句末时往往读作"le5"而不读作"liao3"。

下面以"重"字为例说明一下特征词典的结构, 如下图:

注: 图中用于词性表示的字母含义如下: a 形容词, d 副词, v 动词, n 名词, u 助词, ! 表示所有标点符号。用于边界表示的数字含义如下: 1 代表语素, 2 代表韵律词, 3 代表韵律短语, 4 代表句子。

特征词典对于一个多音字的每个读音, 都有一个包括以上 8 个特征的特征库。对于未知读音的多音字, 只要提取上下文的 8 个特征与库中的特征相匹配就可以判音了。我们试验了有权值和无权值两种方案。区别是: (1) 有权值的做法在学习时不仅要更新词典中的特征库, 而且要更新权值; 无权值的做法在学习中只更新词典中的特征库就可以了。(2) 在判音时, 有权值的做法是对所有的特征匹配情况作一个加权运算, 而无权值的做法无须如此。

3.1　有权值的特征词典

设当前多音字为 x, 它所在的上下文特征库是按照上述 8 项特征来提取的, 称之为活跃特征集, 记为 F。图 3 中每个椭圆是用来计算活跃特征集对于这一个拼音的得分的, 可以表示成 $\theta_i(F)$。上标 i 代表多音字 x 的第 i 个读音, $\theta_i(F)$ 的值通过对特征的匹配值乘上相应的权值得到。

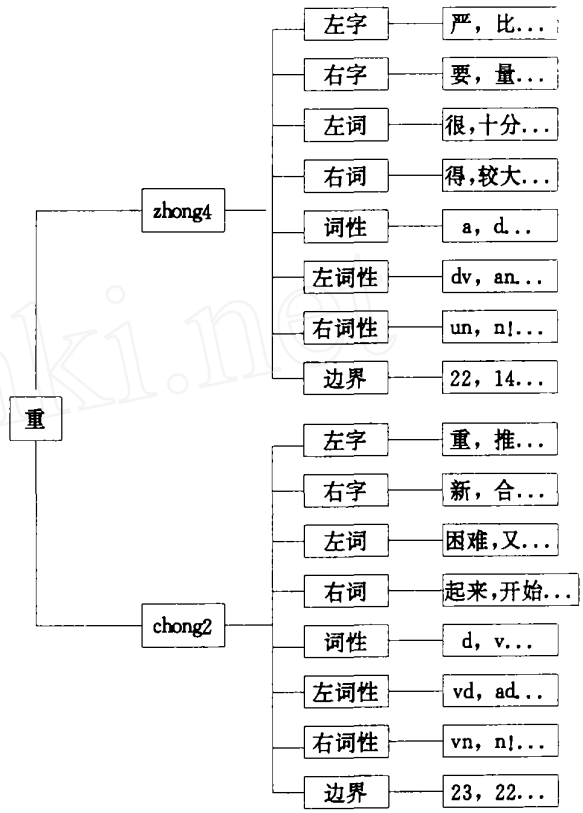

图 2　"重"的特征词典示意图

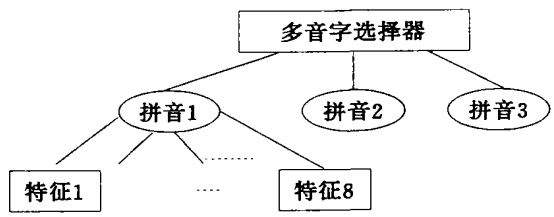

图 3　多音字选择器结构图

F 去匹配词典中的特征时, 有一个匹配分, 记为 $v_i^k(F)$。它的取值为 0 或 1, 0 表示 F 匹配不到该拼音的特征词典, 1 表示可匹配到。下标 i 指的是该字第 i 个读音, 上标 k 表示第 k 个特征, k= 1, 2, …, 8。

特征集 F 依照"寻找交集"的原则去匹配特征词典。这是一个简单的查找过程, 很容易实现。

最后计算 $\theta_i(F)$ 的公式为:

$$\theta_i(F) = \sum_{k=1}^{8} w_{i,x}^k v_i^k(F)$$

其中, $w_{i,x}^k$ 就是多音字 x 的第 i 个读音的第 k 个特征的权值。最后找出最大的 $\theta_i(F)$。取到 i, 就得到了可靠的拼音。公式描述如下:

$$i = \arg\max_i \theta_i(F)$$

在学习过程中, 特征词典本身是要动态更新的。对于有权值的词典来说更新包括两方面, 新的特征的加入和权值的调整。

3.2 无权值的特征词典

无权值的特征词典不考虑权值的影响, 所以学习过程只要动态更新特征就可以了。

基本的判决方法是把 8 项特征分为四组, 按照经验上的优先顺序排列。用活跃特征集 F 对每一组特征依次去作匹配。如果匹配到某一组特征的时候已经有一个读音占有优势地位, 即它的匹配值 $v_i^K(F)$ 已大于另几个拼音的值, 就取这个读音作为判断结果而不必再进行下去。这里 $v_i^k(F)$ 的最终结果取值范围是{0, 1, 2,. . . m},m 是所有特征组中的特征项数。F 匹配到某一组中的一项特征 $v_i^k(F)$ 的值就加 1, 依次类推。无权值的做法简单明了, 只考虑局部效应。

3.3 有权和无权的试验结果及分析

从表 1 可以看到, 虽然有个别多音字有权值占优, 可是总体表现仍然是无权值占优。我们有更多的试验说明了无权值是占优的, 这里不再给出。

表 1　学习语料 sample(7096 字次, 多音 1844 字次)
测试语料 test(5885 字次, 多音 1471 字次)

	有权值	无权值
sample 正确判定字次	7048	7083
sample 多音错误率	0.0097	0.0026
sample 整体错误率	0.0068	0.0018
test 正确判定字次	5685	5688
test 多音错误率	0.049	0.046
test 整体错误率	0.034	0.033
长正确率	0.823	0.909
了 正确率	0.972	0.970
为 正确率	0.796	0.817
重 正确率	0.967	0.972

通过试验可以得出, 使用无权值方法有以下优点:

(1) 决策方法简便, 效率高于有权值方法;

(2) 较好的避免了次要特征对多音判决的负作用。

(3) 同等学习条件下无论是对学习文本还是对未知文本自身错误率都要低;

4 规则的结构及生成算法

每一项规则都由条件(condition), 类别(class), 概率(possibility)三项构成(如图 4)。系统对规则的处理是一致的, 规则可以由 CART、ESC 或其他的算法生成。使用不同算法生成的规则, 达到的准确率也不同。

图 4　规则的例子

```
character= 长 thisproc= a
- - - > > > pinyin= chang2 possibility= 0.857
character= 长 thisproc= v border= 42
- - - > > > pinyin= zhang3 possibility= 0.947
```

4.1 分类回归树(CART)算法

CART[1,2](Classification and Regression Tree)是模式识别中一种简单有效的分类方法, 大多用于决策和聚类问题。CART 主要有训练和决策两种功能。

对确定语境下的多音字判定读音从实质上讲就是一个分类问题。如果已知了一个多音字针对其不同读音的分类树, 我们就可以方便地通过提取的特征在树中找到一条路径, 最终获得叶结点的读音。

我们的问题可定义为: 对于给定特征和读音的样本集 D, 寻找分类树 T, 使得错误率 $R_p(T)$ 最小。得到分类树 T 就得到了规则, 在程序中使用规则只要搜索分类树 T 就可以了。

下图是"长"字的 CART 的例子:

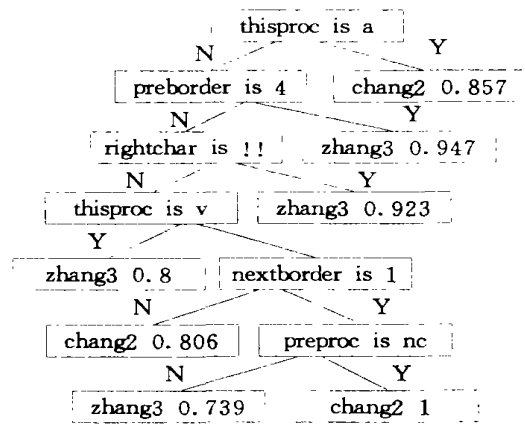

图 5　"长"的分类树示意图

上图的属性集是{thisproc, border, rightchar, preproc}, 类别集合为{chang2, zhang3}。每一个内结点都有一个特征等式。对应满足等式与否有不同的路径。从根结点开始, 一直沿着所提的问题遍历到叶结点, 就可以得到最终读音。叶结点的结构是类别和取到该类别的概率。

对于"长"字, 如果现在通过文本特征的提取得到{thisproc= v, border= 42}, 则最终会到达 zhang3 0.947 的那个结点。通过特征进行有限次的路径选择都可以判断出结果, 没有哪种情况是落在路径之外的。

4.2 基于扩展的随机复杂度(ESC)算法

输入的学习文本是已有正确的拼音标注的文本。经过预处理后, 对每一处多音就可以形成 $D_i=(d_i, c_i)$ 的结构, 其中 d_i 是通过分词系统获取的上下文特征, c_i 是该词正确的拼音标注。

我们的问题就定义为: 如何通过样本集 $D=\{D_1,\ldots D_m\}$ 求得一个 $H=\{f(X)\}$, 使得 $ESC(D^m)$ 最小。X 是未知拼音的多音字的上下文特征, H 就是最终要求得的规则的集合。

ESC 生成算法[4,5]分为特征选取、条款生成、规则生成、裁剪规则四个步骤。限于篇幅, 我们对算法本身不做过多叙述。

4.3 规则的试验

我们进行了两种实现方式的比较, 结果如表 2。

表 2 中的样本集不包含测试集, 已删去规则中的默认读音。以后默认读音由词典作统计得到, 所以包含默认读音的规则集对于我们的系统意义不大。

表 2　两种实现的单字测试

	ESC	CART
长 样本集	120/141= 0. 85	1127/141= 0. 901
长 测试集	76/90= 0. 844	79/90= 0. 878
为 样本集	233/301= 0. 774	232/301= 0. 771
为 测试集	158/218= 0. 725	162/218= 0. 743
重 样本集	94/109= 0. 862	98/109= 0. 899
重 测试集	85/105= 0. 810	85/105= 0. 810

从表 2 中可以看出, 总体上 CART 的表现比 ESC 稍占优。

4. 4　词典加规则补充的试验

现在我们拿已学习好的特征词典, 加上上述实验生成的"为"的规则作为补充。因为仅用词典判断,"为"的正确率相当低, 那么我们看一下加入规则后的效果。测试文件是语料中抽取的含"为"的句子, 总共 7374 字次, 多音 1548 字次,"为"字 326 字次。结果如表 3。

表 3　词典和规则结合的测试

	无权	+ ESC	+ CART
正确判定字次	7008	6991	7010
多音错误率	0. 069	0. 072	0. 066
整体错误率	0. 050	0. 050	0. 049
为 正确率	0. 790	0. 757	0. 801
	有权	+ ESC	+ CART
正确判定字次	6916	6938	6958
多音错误率	0. 087	0. 082	0. 079
整体错误率	0. 062	0. 059	0. 056
为 正确率	0. 713	0. 755	0. 794

可以看到, 虽然规则和有权值的做法相结合也能使其正确率有所提高, 但总体上不如无权值的表现。所以我们最终选择了无权值的特征词典加规则补充的做法。

5　总结

无权值的方法训练完所有的语料后, 错误主要出现在变调和部分难判断的字, 例如"为"、"曾"、"解"等。因为"为"的高低频度的读音相差不是很大, 语言环境都类似, 造成词典中无法明显地区分, 所以我们暂时仅加入了由 CART 生成的"为"的规则。再加入后期"三三"变调以及"一不"变调等处理, 最终该系统获得了 96. 6% 的多音字正确率和 97. 6% 的总体正确率。具体见下表:

表 4　系统加入规则和变调以后的评价

	特征词典训练完后	加入'为'的规则后	加入变调处理以后
正确判定字次	122435	122500	123437
多音错误率	0. 046	0. 045	0. 034
整体错误率	0. 032	0. 031	0. 024

这说明无权值词典加规则补充的方案是有效的。当词典面临更多易混淆的语境时, 规则的重要性将体现得更加明显, 对正确率的贡献也会更加可观。后面的工作将针对其他难以判别读音的多音字(例如多音姓氏"曾"、"谢"、"单"、"华"等)研究并添加更多的规则。词典和规则的可扩充性使得该字音转换系统可以继续完善下去。

参考文献:

[1] Roger J. Lewis, M. D., Ph. D. An Introduction to Classification and Regression Tree (CART) Analysis [A]. Presented at the 2000 Annual Meeting of the Society for Academic Emergency Medicine in San Francisco[C]. California.

[2] Alin Dobra. Classification and Regression Tree Construction[D]. Department of Computer Science, Cornell University, Ithaca NY. Nov. 2002

[3] Hang Li and Kenji Yamanishi. Text Classification Using ESC-based Stochastic Decision Lists[J]. Information Processing & Management, 2002

[4] Zirong Zhang, Min Chu and Eric Chang. "An Efficient Way to Learn Rules for Grapheme-to-Phoneme Conversion in Chinese"[A]. International Symposium on Chinese Spoken Language Processing [C] (ISCSLP2002). Taipei, 2002

[5] 刘松汉. 从汉语多音字的现状看多音字的整理[J]. 南京社会科学, 1995 年 12 期

(收稿日期: 2004-12-30)

verified by emulation results, which can also be used to solve similar combinational optimization problems.
Keywords CGA TSP deceptive problems local optimum global optimum

Precise Timing of Windows2000 ····· P(46)
Xu Liping Zhang Jian (Information College, Shanghai Fisheries University, Shanghai 200090; *Library, Shanghai Fisheries University, Shanghai* 200090)
Abstract Windows2000 Operation System is widely used in the field of computer application for its excellent property. This article analyzes the real-time capability of Windows2000, and brings forward some methods for fulfilling the task of precise timing. Among others, this article mainly talks about the method of setting 8253 by directly programming and programming interrupt processing by means of VxD(Virtual Device Driver) technique.
Keywords real-time precise-timing information VxD

Application of Hybrid Programming Between VC++ and Matlab in Ht-7 Data Diagnosing ····· P(49)
Qi Na Luo Jiarong Shen Jie (Institute of Plasma Physics, Chinese Academy of Sciences, Hefei 230031)
Abstract For actual need, according to the individual characteritics of VC++ and matlab we adopt the technology of hybrid programming between VC++ and Matlab in HT-7 data diagnosing. Two methods of hybrid programming realized by Matlab Engine and Matcom are presented in this paper, and by comparing Matlab and hybrid programming in implementation, this article illustrates the superiority of hybrid programming and also points out the respective limitations of the two methods.
Keywords engine Matcom hybrid programming data diagnosing

An Effective Solution to Polyphone Problem in Mandarin TTS ····· P(52)
Liu Jingyong Chai Peiqi Yao Qiuming (Department of Computer Science and Engineering, Tongji University, Shanghai, 200092)
Abstract The phenomenon of polyphone characters in Chinese increases the difficulty of Mandarin TTS (Text to Speech) system. This thesis is aimed to propose a unified approach to the polyphone decision in Mandarin TTS. The method is based on the thinking of statistical learning. First, we construct a lexicon based on multi-features, which can update automatically according to the corpus in learning. Both of the weighted and unweighted methods are used to update the lexicon. Eventually we choose the unweighted one due to its higher accuracy. We make experiments with classification and regression tree (CART) as well as extended stochastic complexity (ESC). Through experiments, we achieve a relatively satisfactory result using CART to create partial rules as the complement to the lexicon.
Keywords TTS feature lexicon CART ESC

Restricting or Denying BT Transfer Traffic with NBAR ····· P(56)
He Junjie(Network Center, Ningbo University, Ningbo 315211)
Abstract The BitTorrent(BT) application is a peer-to-peer software application that facilitates audio, video and image file-sharing between clients. Unchecked use of this software may lead to the huge waste of the bandwidth. This article introduces how to restrict or deny BT transfer traffic through Network-Based Application Recognition(NBAR) on Cisco.
Keywords BT NBAR Racket Description Language Module(PDLM) Cisco

Learners' Garden

Implementation of Dynamic Properties Configuration with .NET PropertyGrid ····· P(58)
Ding Hao Zhao Zhengwen Li Zhenye (Software College, Northwestern Polytechnical University, Xi'an 710065; *School of Computer Science, Northwestern Polytechnical University, Xi'an* 710072; *Information Center, Affiliated Hospital of Ningxia Medical College, Yinchuan* 750004)
Abstract This paper introduces an implementation method of dynamic properties configuration with .NET PropertyGrid. With this method, you may add, delete and edit object's properties dynamically, and will not be confined to the modification of its property value.
Keywords PropertyGrid dynamic properties property configuration

Introduction of I²C Bus Technology and Application Examples ····· P(61)
Zhao Hui Dong Decun (Traffic and Transportation Institute, Tongji University, Shang Hai, 200331)
Abstract This article introduces the advantage of Inter-Integrated Circuit(I²C) Bus technology in hardware system designing and analyzes the bus technology standard and electric characteristics. Meanwhile, the technical characteristics, the operation, the application examples, the hardware designing and the software application of RX-8025, which is an I²C bus chip made by EPSON company, are described in this article.
Keywords I²C bus RX-8025 real-time clock module

Address: Room 1504, Floor 15, Bao Zhaolong Library, Shanghai Jiaotong University
1954 Huashan Road, Shanghai 200030
Tel: 86-21-62933230 Fax: 86-21-62933230 Email: smcaa@online.sh.cn
URL: http//www.smcaa.online.sh.cn IP: 202.96.210.198
Publisher: Shanghai Microcomputer Application Association Code Number: 6329M
Distributor: China International Book Trading Corporation (P.O. Box 399, Beijing)