

# 基于混合字词网格的汉语音字转换问题的求解

章 森

(北京工业大学信息与计算科学实验室 北京 100022)

**摘 要** 汉语音字转换是中文键盘输入、汉语语音识别和中文信息处理的基础,也是一个非常具有挑战性的问题.文中分析了汉语音字转换的研究现状和存在的问题,提出了基于混合字词网格的汉语音字转换方法,给出了系统实现的架构,研究了混合 2-gram 模型的有关问题以及字词网格的求解算法,最后讨论了自动预测与系统学习功能的实现.在此基础上设计了原型系统并与 Windows XP 上的微软拼音输入系统进行了比较,在拼音到汉字的自动转换正确率方面有显著的提高.

**关键词** 汉语音字转换;  $n$ -gram 语言模型; Markov 模型; 字词网格; 用户行为

**中图法分类号** TP18

## Solving the Pinyin-to-Chinese-Character Conversion Problem Based on Hybrid Word Lattice

ZHANG Sen

(Information and Computational Sciences Research Laboratory, Beijing University of Technology, Beijing 100022)

**Abstract** The research and development of the Pinyin-to-Chinese-Character conversion is the core technique of Chinese Input system, Chinese speech recognition and Chinese information processing. First, the state-of-the-art of Pinyin-to-Chinese-Character conversion is briefly discussed, and its principles and shortcomings are analyzed. Then the conversion approach based on hybrid word lattice is proposed. The implementation of the main architecture is studied. The related problems with hybrid language model and the algorithms to solve the word lattice are investigated. Finally, the automatic prediction algorithm and the machine learning technology used in Chinese intelligent input systems are discussed. A prototype system realized based on the proposed approach is presented, and compared with the MS Pinyin input system in Windows XP. The experimental results show that the correct conversion rate from Pinyin to Chinese characters is significantly improved.

**Keywords** Pinyin-to-Chinese-Character conversion;  $n$ -gram language model; Markov model; word lattice; user's action

## 1 引 言

汉语音字转换技术主要实现从拼音编码到汉字内码的转换,它是中文键盘输入、汉语语音识别和中

文信息处理技术的基础.从本质上说,这种转换就是从一种语言到另一种语言的翻译过程,其困难在于语言固有的多义性,求解方法一般是用语言的上下文相关信息作为约束条件计算最优解.到目前为止,已经提出并实现了多种汉语音字转换技术,可以分

为如下几类<sup>[1-2]</sup>:基于语法规则、基于统计知识、基于模板匹配和基于上下文关联方法等.但是,这些方法并不是相互独立的,例如上下文关联技术在上述几类方法中都是不可或缺的.

基于语法规则的汉语音字转换是把汉语语法知识用计算机能够处理的形式表示出来,把汉语中的固定搭配关系用规则的形式表示出来;在进行汉语音字转换的时候,查找语法规则知识库,经过归约推理,得到转换的结果<sup>[3]</sup>.这种方法存在的问题是语法规则的完备性在理论上无法实现,其规范性在实现时难以保证.

基于统计知识的汉语音字转换利用汉语的统计语言模型和基于动态规划的算法来求解汉语音字转换问题,其目标是找到一条最佳路径.这种方法存在的主要问题是:语用知识不是  $n$ -gram 模型能够全部描述出来的,在理论上这样产生的语用知识是不完备的,在实际实现时也存在数据稀疏的问题.

基于模板匹配的汉语音字转换是把汉语短语作为“模板”.该方法首先查找模板库,进行匹配处理.如果匹配结果不唯一,则根据句法规则或概率信息进一步判定,选出一个最佳的可能结果作为输出.该方法的主要优点是易于实现系统的自学习功能,其缺点是汉语短语的定义不严格,且灵活性不够;另外,汉语短语的数量很大,可以达到几百万条,但实现匹配时要求实时处理,这是其困难所在.

基于上下文关联的汉语音字转换是利用上下文语境作为约束条件求解语言的多义性问题<sup>[4-6]</sup>.上下文关联的表示可以用规则的方法,也可以用统计语言模型的方法.上下文关联技术可以有效地利用在转换过程中产生的动态知识,这些知识对解决语言的多义性问题非常有用;但是,上下文关联技术也存在固有问题,例如上下文动态知识的获取、表示和利用,上下文搜索的范围问题,与其它技术结合的问题等.尽管如此,上下文关联技术仍是汉语音字转换技术发展的主要方向.

目前的汉语音字转换方法存在的主要问题在于:一是没有充分利用转换过程中产生的动态知识;二是没有充分利用专业领域的静态知识.无论上述的动态知识还是静态知识,其局部性都很强,可以看作是局部知识.利用这些局部知识往往可以求得问题的局部最优解,但不是全局最优解.针对上述问题,本文提出了基于混合字词网格模型的汉语音字转换方法,给出系统实现的架构,研究了基于混合

2-gram 模型和字词网格的求解算法,讨论了自动预测与系统学习功能的实现.我们将实现的原型系统与微软的拼音输入系统进行了比较,在拼音到汉字自动转换正确率方面有明显提高,基本实现了预期的目标.

2 汉语音字转换问题

从系统设计的角度看,汉语音字转换的输入参数是字母或数字组成的代码串,输出的是汉字串,这些都是用户可见的;系统根据配备的语法知识将代码串转换成汉字串的过程是用户不可见的,但当系统配备的语法知识不足以确定唯一的输出时,系统就要借助于外部知识(如用户的选择等)来完成转换任务.显然,这是一个典型的人机交互系统.对于这样的一个系统,从用户的角度看,系统借助于外部知识的次数越少越好.从系统论的角度来看,如果系统不需要外部知识的支持,仅仅利用自身配置的语法知识进行处理,这个系统是一个封闭的系统.但是,这个封闭的系统要处理的问题集合是开放的,其可能的输入是无限的.因此,任何汉语音字转换系统,无论它的智能化程度多高,都要借助于外部知识,其差别只是多少而已.从信息论的角度看,给定一个代码串,其对应的汉字串是不确定的.因此,每一个代码串都有一定的信息熵,汉语音字转换是利用各种知识使得代码串的信息熵逐渐减少的过程.一般来说,汉语音字转换过程中给定的代码串越短,其信息熵越大,要确定它的输出(对应的汉字串)越困难.语句级汉语音字转换就是利用比较长的代码串,降低其信息熵,期望输出比较准确的结果<sup>[4]</sup>.

从数学的角度看,汉语音字转换就是求解约束优化的问题,其数学模型可以表示如下:

$$\begin{cases} C \Rightarrow W, & C = c_1 c_2 \cdots c_n, W = w_1 w_2 \cdots w_m, \\ & c_i \text{ 为代码, } w_j \text{ 为汉字} \\ & \text{目标函数为 } f(W, C), \text{ 约束条件为 } \Omega(W) \end{cases},$$

其中,  $C \Rightarrow W$  表示从代码串到汉字串的映射.一般情况下,这个映射不是一一映射,而是一对多的映射.对于每一个给定的代码串  $C$ ,它所有可能的映像形成了汉字串集合的一个子集,我们称为代码串  $C$  的映像集,也称为解空间,记为  $P(C)$ .显然,  $W \in P(C)$ ,  $P(C) \subset \Omega(W)$ .因为该优化问题的变量都是离散的,一个基本的求解算法是对于解空间中的每一个解,在给定的约束条件下计算相应的目标函数值,通过

比较目标函数的值,确定最优的解.显然,这个算法需要的计算量是很大的,特别是当解空间很大时.目前,一种常用的方案是利用概率统计的方法对上述数学模型进行简化,并利用基于动态规划的算法(如 Viterbi 算法等)进行求解.汉语音字转换的目的就是根据某些静态的和动态的语法、语用知识(约束条件),在解空间中快速找到一个解,它能够使得目标函数达到最优.

### 3 基于混合 $n$ -gram 的问题求解

#### 3.1 系统结构与字词网格

汉语音字转换系统主要包括如下几个步骤:检

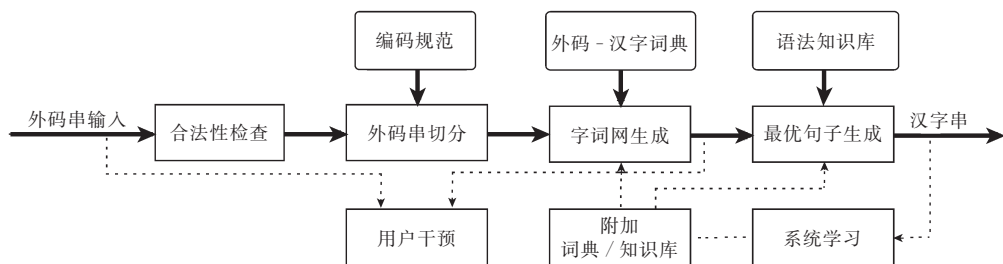


图 1 中文智能输入系统的实现架构

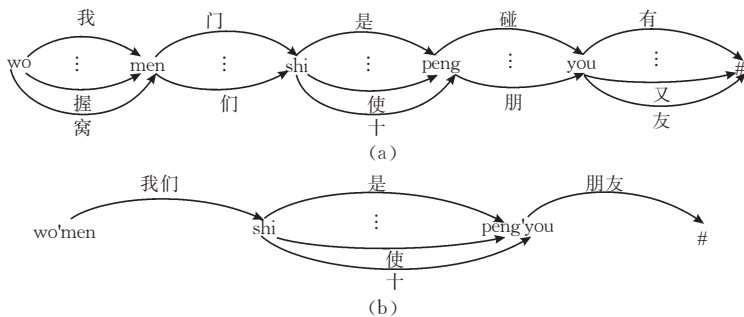


图 2 字词网格的有限状态图表示

#### 3.2 基于统计模型的字词网格求解

基于上述字词网格,如何快速求解出最优的汉字串(句子)是汉语音字转换问题的关键所在.从统计学的角度看,汉字输入中的字与字、词与词之间的相关信息可以用 Markov 过程进行近似.我们可以把拼音串的输入看作一个 Markov 过程.因此,实现汉语音字转换的过程实际上是利用 Markov 模型,在先求出汉字词先后之间的转移概率的基础上,再求得整个字词串的一个最大路径值.因此,基于统计的观点,我们可以把上述问题的求解等价的转化为最大概率的计算问题<sup>[7-8]</sup>:

$$W = \arg \max_w p(W|C) = \arg \max_w \frac{p(W)p(C|W)}{p(C)}$$

查输入的外码序列是否符合编码规范;外码序列分割;生成字/词网格;确定最佳转换结果(汉字串);反馈学习等.下面的图 1 描述了汉语音字转换系统的一般实现架构.

这里,外码就是拼音编码,输入的拼音串一般没有边界标记,例如,“womenshipengyou”是一个输入的拼音串.基于此输入,一般有两种方法生成字词网格:一种是逐字转换法,也就是每次转换时考虑一个拼音,如图 2(a)所示,这种方法的转换效率和准确度都不高,现在的汉语音字转换已经很少使用;另一种方法是先进行拼音分词,然后再转换,如图 2(b)所示,这种方法的转换效率和准确度都有所提高,也是目前广泛采用的技术.

$$= \arg \max_w [p(W)p(C|W)] = \arg \max_w [p(W)],$$
其中,  $W = w_1 w_2 \cdots w_n$  表示汉字串,  $C = c_1 c_2 \cdots c_m$  表示外码串,  $p(C|W)$  表示给定汉字串  $W$  其外码串为  $C$  的概率,这个概率一般情况下可以近似认为是 1,因为给定汉字串时外码串一般是确定的.因此,上述问题的求解可以转化为计算汉字串  $W$  出现的概率  $p(W)$ .

假设汉字串  $W = w_1 w_2 \cdots w_n$  中的当前汉字  $w_i$  只与其前  $k$  个汉字有关,即  $w_i$  只与  $w_{i-1}, w_{i-2}, \cdots, w_{i-k}$  有关,那么,概率  $p(W)$  的计算问题可以用  $k$  阶 Markov 过程进行近似,即

$$p(W) = p(w_1)p(w_2|w_1)\cdots p(w_n|w_1w_2\cdots w_{n-1})$$

$$\approx \prod_{i=1}^n p(w_i | w_{i-k} \cdots w_{i-1}).$$

实际上,上述问题是汉字的  $k+1$  阶统计语言模型问题,而统计语言模型可以通过统计分析大量的语料获得.但是,由于当  $k$  较大时,计算  $k+1$  阶统计

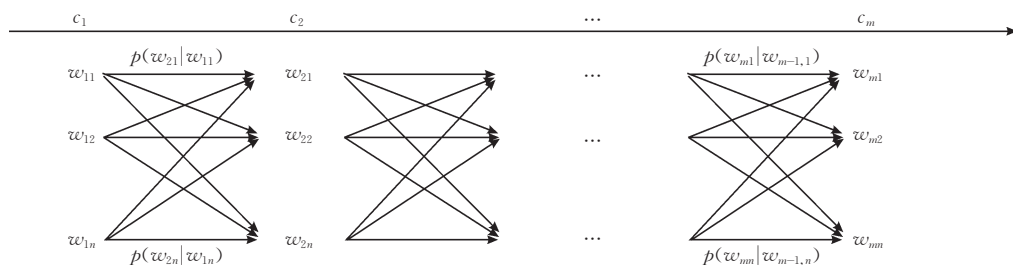


图 3 字词网格的求解过程

图 3 中,  $c_1 c_2 \cdots c_m$  表示外码串,  $w_{11} w_{12} \cdots w_{1n}$  表示外码  $c_1$  对应的可能的  $n$  个汉字,  $p(w_{21} | w_{11})$  表示汉字  $w_{11}$  到汉字  $w_{21}$  的转移概率,即统计模型 2-gram 中  $w_{11}$  出现情况下  $w_{21}$  出现的条件概率,其余符号的意义类似.显然,对于上述字词网格,直接计算需要的计算量为  $O(n^m)$ ,但可以用 Viterbi 算法求出一条最大路径值,其计算量为  $O(n^2 m)$ ,回溯求出的最大路径可以得到相应的汉字串  $W$ .

我们对上述基于统计方法的求解做一个简要的理论分析.

(1) 在给定外码串  $C = c_1 c_2 \cdots c_m$  的条件下,我们假设了所求的汉字串  $W = w_1 w_2 \cdots w_n$  满足条件  $W = \arg \max_W p(W | C)$ . 实际上,这个假设对出现概率较小的汉字串是不公平的;因为根据上述求解方法,它们不是概率意义上的最佳解,但它们可能是真正要求的解.例如,某些专用名称(如人名、地名等)的出现概率可能很小,它们往往被错误转换;一个极端的例子是,当输入的拼音外码串只有一个拼音时,拼音对应的汉字可能多个,只有出现概率最大的才是概率意义上的最佳解,如果用户实际需要的不是这个汉字,就会产生转换错误.这样的错误是很常见的,这是因为平均每个无声调拼音对应 16.2 个汉字,拼音重码太多.

(2) 我们在计算汉字串  $W$  出现的概率  $p(W)$  的时候,假设了汉字串  $W = w_1 w_2 \cdots w_n$  中的当前汉字  $w_i$  只与其前  $k$  个汉字有关.这个假设考虑了句子中汉字之间的局部性关系,而忽略了全局性关系,在计算概率  $p(W)$  的时候必然会出现误差.这个计算误差实际上类似于函数逼近中的截断误差,它将导致最后求出的解可能有误差.在实际的汉语句子中,每

语言模型需要的语料数量非常巨大,现有的技术难以实现,所以,现在一般使用 2-gram 和 3-gram 较多,更高阶的很少使用.假设我们已经有了一个统计语言模型 2-gram,那么,一个字词网格的求解过程如图 3 所示.

个汉字之间都有联系,而不论它们之间的相对距离有多远;但是,这种长距离的相关性用统计的方法现在难以描述,一种常用的方法是用语法规则来描述.

基于统计方法的汉语音字转换的准确率上限是可以估计出来的.我们首先考虑输入的句子是单字情况,这时其准确率上限是 6.2% 左右;接着考虑输入的句子中包含  $n$  个汉字的情况.假设  $C$  表示  $n$  个拼音连接而成的串,所有这样的拼音串组成的集合为  $\Omega(n)$ ,  $\Omega(n)$  的基数记为  $|\Omega(n)|$ ,拼音串  $C$  对应的汉字串可能多个,它们的个数记为  $O(C)$ ,这时其

准确率上限是:  $\overline{P(n)} = \frac{1}{|\Omega(n)|} \sum_{C \in \Omega(n)} \frac{1}{O(C)}$ ; 因为  $\Omega(n)$  是一个无限集合,实际计算  $\overline{P(n)}$  的时候可以采用抽样估计的方法.假设我们得到了  $\overline{P(n)}$ ,那么基于统计方法的汉语音字转换的准确率上限是:  $\bar{P} = \lim_{N \rightarrow \infty} \frac{1}{N} \sum_{n=1}^N \overline{P(n)}$ . 这个准确率上限是在统计意义下的平均值.如果再考虑到外码串在分词过程中可能产生误差(拼音串分词的误差大约在 5% 左右),基于统计方法的汉语音字转换的准确率上限  $\bar{P}$  一般不会超过 95% 左右.

### 3.3 混合 $n$ -gram 模型

基于输入的外码串转换成的字词网格一般有两种:一种是单字网格,这是将外码串切分成一个个的外码而生成的.另一种是先对外码串进行分词处理,然后再转换成字词网格,这种网格是字词混合的.对于单字网格,可以直接使用统计语言模型  $n$ -gram (我们称这种语言模型为正规  $n$ -gram 模型)进行求解,但效果并不好,主要原因有两条:(1) 计算量大;(2) 转换准确度低.对于字词网格,不能直接使用正

规  $n$ -gram 进行求解,而是要采用混合统计语言模型,我们称之为混合  $n$ -gram 模型. 这种模型中的语言对象 gram 可以是单字,也可以是词. 基于这种混合语言模型的字词网格求解与基于正规  $n$ -gram 模型的字网格求解比较起来,其优势在于减少了网格中的节点数目,降低了计算量,提高了转换准确度<sup>[9-10]</sup>.

对于一个基于拼音串生成的字词网格,我们讨论利用混合  $n$ -gram 模型进行求解的问题. 假设字词网格有  $m$  个节点,第  $i$  个节点有  $n_i$  个对象,它们分别是  $w_{i1}w_{i2}\cdots w_{in_i}$ . 在用 Viterbi 算法计算最大路径的过程中,对  $m$  个节点进行分段处理,即处理完前面  $i-1$  个节点后,再考察第  $i$  个节点(当前节点)的对象. 假设已经处理完前面  $i-1$  个节点,且第  $i-1$  个节点的  $n_{i-1}$  个对象分别对应一个  $i-1$  阶的最佳路径值  $\alpha_{i-1}(1), \cdots, \alpha_{i-1}(n_{i-1})$ ,那么第  $i$  个节点的第  $j$  个对象的  $i$  阶的最佳路径值为  $\alpha_i(j) = \max_{1 \leq k \leq n_{i-1}} [\alpha_{i-1}(k)p(w_{i,j}|w_{i-1,k})]$ ,其中  $p(w_{i,j}|w_{i-1,k})$  可以由混合  $n$ -gram 模型给出. 在计算最佳路径值  $\alpha_i(j)$  的时候,可能遇到如下几种情况:

(1) 所有的  $p(w_{i,j}|w_{i-1,k})$  全不为零 ( $1 \leq j \leq n_i, 1 \leq k \leq n_{i-1}$ ). 这是正常情况,按照上述方法可以计算最佳路径值  $\alpha_i(j)$ .

(2) 所有的  $p(w_{i,j}|w_{i-1,k})$  全为零 ( $1 \leq j \leq n_i, 1 \leq k \leq n_{i-1}$ ). 按照算法的正常处理,路径将在第  $i$  个节点处断开,得不到一条关于  $m$  个节点的最佳路径. 这时,为了使得考察的路径能够继续扩展,一般有两种方法对  $p(w_{i,j}|w_{i-1,k})$  进行改造:一种是静态处理方法,对语言模型中为零的项进行平滑处理,例如常用的 Backoff 方法等;另一种是动态处理方法,即对  $p(w_{i,j}|w_{i-1,k})$  赋予一个很小的不为零的数值. 当只有部分  $p(w_{i,j}|w_{i-1,k})$  为零的时候,可以忽略之.

(3) 第  $i$  个节点只有一个对象. 这样的节点相当于有向图中的汇节点(sink node),如果字词网格中的每个节点都是这样的节点的话,计算量显然会大大降低,准确度会大大提高. 在这样的节点上,我们可以把路径断开,使得整个路径的计算转化成部分路径的计算,最后再合并起来. 这样处理的结果可以使得计算量降低,但转换准确度不受影响.

建立混合  $n$ -gram 模型需要统计分析大量的语料. 我们以  $n=2$  为例,汉语中常用词大约有 5 万条,建立混合 2-gram 模型至少需要数十亿级规模的语

料;而建立正规的汉语 2-gram 模型需要千万级规模的语料,它们之间相对差了 2 个数量级. 另外,混合 2-gram 模型占用的存储空间也是很大的. 假设混合 2-gram 中的每一个条目(其中至少包括 2 个汉字词和一个概率数值)占用 8 个字节,那么一个这样的混合 2-gram 至少需要 2GB 的存储空间. 这么大的语言模型无论是查找还是存储都是不方便的,在实际应用中必须进行压缩. 语言模型的压缩技术主要利用其稀疏性特征,即语言模型中大量的(90%左右)条目的概率值为零,对于这样的条目我们可以从语言模型中去掉,而在用到它们的时候用动态平滑的方法生成. 这样就可以把语言模型压缩到原来规模的 10% 左右. 这样的模型还可以采用编号的方法进一步压缩. 例如,对模型中用到的 5 万词条进行编号,从 1 到 5 万(采用 16 进制,每个标号最多 4 个字节),在混合 2-gram 的条目中,不是存储的汉字词本身,而是它们的标号. 在对词进行编号的时候,可以用霍夫曼编码方法,即出现最多的、最长的词给予最小的标号,其余的依次类推. 这样还可以把语言模型再压缩 10 倍以上,在数十兆的规模.

建立统计语言模型的过程中,我们可以粗略地估计出至少需要多少语料,但多少训练语料才是足够的? 从统计理论上说,根据大数定理,训练语料无限多时才能得到统计特征. 但实际中能够使用的训练语料总是有限的,机器的能力也是有限的. 因此,我们需要估计一下多少训练语料是合适的. 一般地,可以采取粗估计和精估计两种方法:粗估计方法是先统计出汉语基本  $n$  字词的词汇量大约是多少并把它们列出,然后逐渐增加训练语料,当训练所得的  $n$ -gram 中已经包含了列出的基本  $n$  字词时,可以认为训练语料已经合适了. 另一种精估计方法是逐渐增加训练语料,判断  $n$ -gram 中某些条目的概率数值是否已经收敛,如果已经收敛了,可以认为训练语料已经合适了. 在  $n$ -gram 的所有条目中,可以挑选其中的概率值较大的条目(例如,选出 10%)作为考察对象,计算在每一次增加训练语料的迭代过程中考察对象的概率值的变化率. 如果总体变化率很小,表明已经收敛,停止训练.

### 3.4 动态预测与系统学习的实现

汉语音字转换的过程是一个人机交互的过程,在这个交互过程中产生许多用户行为. 通过对用户行为的跟踪与分析,在汉语音字转换过程中可以获取与用户有关的动态知识,进而提高其智能化程



度<sup>[11]</sup>. 我们以拼音输入为例, 讨论在汉语音字转换过程中系统的动态预测与学习功能的实现问题.

我们需要注意和分析以下几类对象: 第一类对象是已经转换出来的字词, 这些结果都是经过用户

确认的; 第二类对象是余下的还没有转换的拼音外码串. 在输入的拼音串较短时, 我们可以利用它们进行动态预测, 实现过程如图 4 所示.

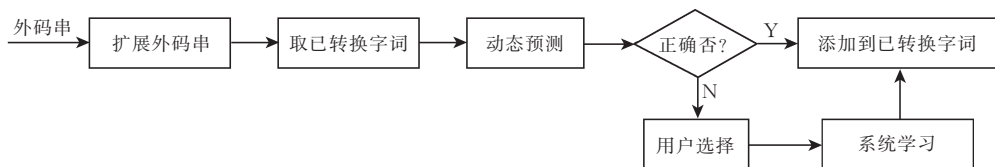


图 4 动态预测的实现

假设  $w_1, w_2, \dots, w_n$  分别是根据余下的外码串中待处理的外码串扩展出的汉字或词, 基于上述两类对象以及混合语言模型 2-gram, 可以实现在汉语音字转换中的一步预测如下 (如图 5): 取已转换的汉字串  $W$ , 抽取  $W$  中的最后一个词  $u$ , 则一步预测的字词  $w_*$  为  $w_* = \arg \max_{w_i} [p(w_i | u)]$ . 显然, 如果  $u_1, u_2, \dots, u_k$  分别是  $W$  中的最后  $k$  个词, 我们还可以使用混合语言模型  $n$ -gram 实现一步预测如下:

$$w_* = \arg \max_{w_i} [p(w_i | u_1, \dots, u_k)].$$

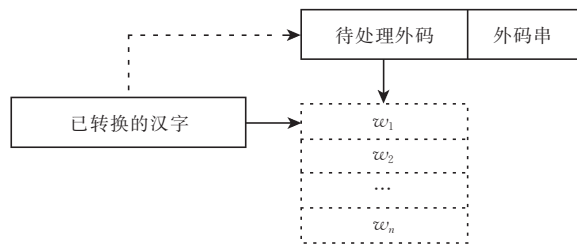


图 5 基于混合 2-gram 的一步预测技术

在动态预测过程中可能出现预测错误, 需要用户选择或更正. 当在预测过程中出现同样的情况时, 用户不希望系统还出现同样的预测错误. 因此, 系统应该具有自动学习功能, 添加系统原来没有的知识, 改进系统已有的知识, 调节系统的知识库以便更好地适应不同背景用户的需求. 在汉语音字转换过程中, 系统学习的行为主要是指分词的学习和转换的学习, 学习方法主要是跟踪并记录用户的选择以及上下文.

我们首先讨论系统对输入的拼音串的分词问题. 因为在自动分词过程中往往出现歧义问题, 所以可能出现分词错误, 而分词错误往往导致转换错误. 产生分词错误的情况主要有两种: 一是系统把该切分成一个词的拼音串没有切分成一个词, 例如系统词库中未登录的专用名称 (人名、地名、术语等); 二是系统把不该组成一个词的外码串切分成了一个

词, 例如系统往往把某些组词能力较强的字与其左右结合, 错误切分成一个词. 假设输入的拼音外码串为  $c_1 c_2 \dots c_m$ , 系统在自动分词过程中将  $c_i \dots c_j$  分成了一个词, 但用户认为正确的分词结果应该是  $c_i \dots c_k$ , 并且用户确认了  $c_i \dots c_k$  对应的汉字串为  $w_i \dots w_k$ . 这时系统就应该在动态词库中记录正确的分词结果为  $c_i \dots c_k$ , 其对应的汉字串为  $w_i \dots w_k$ , 其上下文为  $c_{i-1}$  和  $c_{k+1}$ . 在外码转换过程中的学习与此类似, 不再赘述. 可见, 这种系统学习是一个机械的过程, 容易实现. 但是, 系统对未登录词的处理并非易事, 而未登录词的处理是影响汉语音字转换性能的一个主要因素.

系统对未登录词的处理方法主要有两种: 一种是离线处理方法, 它一般要求用户在动态词典中手工加入未登录词及其相关的知识, 这种方法不在我们讨论的系统自动学习的范畴; 另一种是在线处理方法, 其主要思想是系统自动识别未登录词并加入系统的动态词库, 其困难在于怎样使系统自动确定未登录词的边界, 并且不需要增加用户额外的操作负担, 对此问题一般采用规则而不是统计的方法进行处理. 假设输入的外码串为  $c_1 c_2 \dots c_m$ , 其中  $c_i \dots c_j$  是一个未登录词, 应该切分成一个词. 但是, 一般情况下, 系统在自动分词的时候, 会把  $c_i \dots c_j$  分成若干词, 典型的情况是把  $c_i \dots c_j$  中的每个外码  $c_k$  单独分成一个词, 用户根据系统的提示对每个外码  $c_k$  选择一个合适的汉字  $w_k$ ; 当系统处理完整个的外码串  $c_1 c_2 \dots c_m$  的时候 (即外码串  $c_1 c_2 \dots c_m$  中的每个外码都转换完毕并得到了用户的确认), 其转换的结果为汉字串  $w_1 w_2 \dots w_m$ . 这时, 系统要对分词和转换过程中用户的行为进行跟踪学习, 处理规则如下:

(1) 如果系统把  $c_i \dots c_j$  中的每个外码  $c_k$  单独分成一个词, 而且  $c_{i-1}$  与  $c_{j+1}$  都是词, 那么, 将把  $c_i \dots c_j$  合并成一个词, 在动态词库中添加词条记录:  $c_i \dots c_j$ ,

$w_i \cdots w_j, c_{i-1}, c_{j+1}$ .

(2) 如果系统把  $c_i \cdots c_j$  中的每个外码  $c_k$  单独分成一个词, 而  $w_i$  (或  $w_j$ ) 是最常用字, 则把  $c_{i+1} \cdots c_j$  (或  $c_i \cdots c_{j+1}$ ) 合并成一个词, 在动态词库中添加词条记录:  $c_{i+1} \cdots c_j, w_{i+1} \cdots w_j, c_i, c_{j+1}$  (或者添加:  $c_i \cdots c_{j-1}, w_i \cdots w_{j-1}, c_i, c_{j-1}$ ).

(3) 如果外码串  $c_i \cdots c_j$  的长度  $j-i+1$  超出了系统规定的词的最大长度  $N$ , 则把外码串  $c_i \cdots c_j$  切分成  $M$  个词条, 其中  $M=(j-i+1)/N$ , 最后一个词条的长度为  $(j-i+1) \bmod (N)$ .

上述给出的对未登录词的处理规则比较简单, 而实际采用的规则可能远远不止这些. 但是, 无论使用多少规则, 对未登录词的处理也不可能彻底解决. 虽然可以使用比较复杂的句法分析的方法, 将句子的骨干部分抽出, 其余部分进行归类合并, 但输入的拼音外码串往往不是整个句子而是词语或短语, 且往往不合乎语法. 因此, 考虑到对整个系统性能的影响, 句法分析往往得不偿失.

## 4 实验结果

汉语音字转换系统对实时性和准确度具有很高的要求, 在用户输入外码串完毕之后, 用户希望立刻看到转换的结果; 当有多个转换结果时, 用户期望得到的结果能出现在最前面. 因为我们使用混合 2-gram 作为求解字词网格的知识库, 而混合 2-gram 含有的条目在 100 万以上的规模, 所以, 对混合 2-gram 的搜索算法是影响系统实时性的一个重要因素; 在混合 2-gram 中, 如果  $p(w_2 | w_1)$  不为零, 其结构组织如下:

$$w_2, w_1, p(w_2 | w_1),$$

其中  $w_2$  是当前词,  $w_1$  是  $w_2$  的前趋词,  $p(w_2 | w_1)$  是统计概率值. 这种以当前词为关键词的组织结构有利于分块建立索引, 提高搜索速度. 这样建立的索引表中大约有 5 万项, 每个索引项平均对应混合 2-gram 中的大约 20 个条目; 索引表中的索引项按词典顺序排序, 采用二分查找算法进行搜索, 对每个关键词在索引表中最多进行大约 16 次查找即可; 这样, 即使对混合 2-gram 中相应块进行顺序搜索, 对每个词的平均搜索次数在 36 次左右, 系统实现可以实时.

我们使用的混合 2-gram 是通过网络搜索引擎的机器人程序自动采集网页信息并进行清洗过滤、

分类、统计分析而得到的, 含有的条目大约是 120 万, 采用非压缩格式时候占用大约 21 兆的存储空间. 系统自动学习过程中得到的知识存储在动态词库中, 其结构组织如下:

$c_1 c_2 \cdots c_m, w_1 \cdots w_m, c_L, c_R, w_L, w_R, Freq(W | C)$ , 其中,  $c_1 c_2 \cdots c_m$  是外码串,  $w_1 \cdots w_m$  是对应的汉字符串,  $c_L, c_R$  是  $c_1 c_2 \cdots c_m$  的外码上下文 (可能为空),  $w_L, w_R$  是汉字上下文 (可能为空),  $Freq(W | C)$  是计数器. 系统词库大约含有 5 万个词条, 其组织结构与动态词库类似, 都是以外码串作为关键词; 同一个外码串可能对应不同的汉字符串, 所以词库中以同一个外码串开始的条目可能有多条; 字词网格的生成和求解都是以系统词库和动态词库为基础的.

基于本文提出的方法, 我们在 Windows XP 系统下开发了一个汉语音字转换的原型系统, 采用了本文提出的混合 2-gram 语言模型, 实现了动态预测和自动学习功能等. 为了测试原型系统的转换正确率, 我们从《人民日报》的新闻、社科等方面的文章中随即抽取了 960 句 (包含 12689 个音节) 作为测试语料, 尔后我们对 Windows XP 的微软拼音输入系统也用同样的测试语料进行了测试, 以便比较它们在转换正确率方面的差别. 在测试中, 我们分首字正确和首页正确两种情况进行了统计 (每页最多显示 10 个可选结果), 如表 1 所示.

表 1 转换正确率测试结果比较

系统	首字正确率/%	首页正确率/%
原型系统	92.1	96.2
微软拼音	87.6	94.9

在微软拼音系统中, 有些转换错误的原因是没有考虑用户最近输入的历史记录; 例如, 当用户输入了拼音串 “boshishenglunwen”, 并选择了 “boshi” 对应的汉字符串 “博士” 之后, 系统可以根据语言模型预测到 “博士” 之后的 “sheng” 对应的是 “生”, 而不是 “声”; 还有些错误是外码串切分不当引起的, 如 “lingyigeren”, 微软拼音系统给出的转换结果是 “凌夷个人”; 前一种错误在我们的原型系统中较少, 后一种错误出现在两种系统中的概率几乎一样. 对后一种错误, 系统经过一段时间的学习, 分析用户的切分信息, 可以使正确率有明显提高, 这说明系统的自动学习功能是有重要作用的.

## 5 结束语

本文分析了汉语音字转换技术的现状, 在此基

础上提出并讨论了基于混合字词网格求解的汉语音字转换方法,给出了系统实现的架构,研究了混合 2-gram 模型的有关问题以及字词网格的求解算法,最后讨论了自动预测与系统学习功能的实现.我们实现了一个汉语音字转换原型系统,并用小规模语料进行了拼音到汉字的自动转换测试,并用同样的测试语料在微软拼音系统上进行测试;结果表明,在转换正确率方面我们提出的方法有一定的优势,但是,测试的结果也说明了系统存在一些问题,特别是对系统出现的转换错误进行分类研究之后,我们发现混合 2-gram 模型还不够精确,用于训练该模型的语料选取太偏重于新闻、社科等领域,不够全面;另外,对未登录词的处理规则不够,是导致转换错误的一个主要原因.我们相信,采用混合 3-gram 模型将会显著提高汉语音字转换的性能,系统自动学习功能是提高系统适应不同用户的有效途径.

## 参 考 文 献

- [1] Chen Yi-Fan, Hu Xuan-Hua. Chinese Character Input Technology and Related Theories. Beijing: Tsinghua University Press, 1994(in Chinese)  
(陈一凡, 胡宣华. 汉字键盘输入技术与理论基础. 北京:清华大学出版社, 1994)
- [2] Chen Yi-Fan, Zhu Liang. A general statement for the intelligent input software of Chinese characters. Journal of Chinese Information Processing, 2003, 17(2): 60-65 (in Chinese)  
(陈一凡, 朱亮. 汉字键盘输入智能处理软件综述. 中文信息学报, 2003, 17(2): 60-65)
- [3] Wang Xiao-Long. The research of machine learning in Chinese syllable-character transformation. Chinese Journal of Computers, 1993, 16(5): 370-337 (in Chinese)  
(王晓龙. 音字转换中的机器学习研究. 计算机学报, 1993, 16(5): 370-377)
- [4] Xu Zhi-Ming, Wang Xiao-Long, Jiang Shou-Xu. A sentence-

level Chinese character input method. High Technology Letters, 2000, 10(1): 51-56 (in Chinese)

(徐志明, 王晓龙, 姜守旭. 一种语句级汉字输入技术的研究. 高技术通讯, 2000, 10(1): 51-56)

- [5] Yu Shi-Wen. The application of grammatical analysis technique in Chinese input. Journal of Chinese Information Processing, 1988, 2(3): 20-26 (in Chinese)  
(俞士汶. 中文输入中语法分析技术的应用. 中文信息学报, 1988, 2(3): 20-26)
- [6] Zhang Sen, Zong Cheng-Qing et al. Analysis on the intelligent processing mechanisms in the sentence-level pinyin to Chinese characters conversion. Journal of Chinese Information Processing, 1998, 12(2): 37-43 (in Chinese)  
(章森, 宗成庆等. 语句拼音-汉字转换的智能处理机制分析. 中文信息学报, 1998, 12(2): 37-43)
- [7] Guo Jin. Statistical language modeling and some experimental results on Chinese syllable-to-words transcription. Journal of Chinese Information Processing, 1993, 7(1): 18-27 (in Chinese)  
(郭进. 统计语言模型及汉语音字转换的一些新结果[J]. 中文信息学报, 1993, 7(1): 18-27)
- [8] Zhao Yi-Bao, Sun Sheng-He. A word-self-made Chinese phonetic-character conversion algorithm based on Chinese character bigram. Acta Electronica Sinica, 1998, 26(10): 55-58 (in Chinese)  
(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): 55-58)
- [9] Chen Zheng, Lee Kai-Fu. Spelling correction in pinyin input. Chinese Journal of Computers, 2001, 24(7): 758-763 (in Chinese)  
(陈正, 李开复. 拼写纠正拼音输入法中的应用. 计算机学报, 2001, 24(7): 758-763)
- [10] Chen S F, Goodman J. An empirical study of smoothing techniques for language modeling. Computer Speech and Language, 1999, 13: 359-394
- [11] Gao Jian-Feng, Goodman Joshua, Li Ming-jian, Lee Kai-Fu. Toward a unified approach to statistical language modeling for Chinese. ACM Transactions on Asian Language Information Processing, 2002, 1(1): 3-33



**ZHANG Sen**, born in 1963, Ph.D. .

His research interests include multimedia signal processing and computational linguistics.

## Background

The Pinyin-to-Chinese-Character conversion is the fundamental and core technique in Chinese Input system, Chi-

nese speech recognition and Chinese information processing. The research and development in this area have made great



progress and promoted Chinese information processing theory and technology significantly since 1980s, i. e. , the conversion accuracy which is the most important evaluation factor can reach 90% or higher in some environments. However, the conversion accuracy still can be improved by exploring the users' input activities and exploiting dynamic knowledge yielded in the process of users and systems interaction. The purpose of the authors' s work is to provide high performance Pinyin-to-Chinese-Character conversion approach based on large scale hybrid language models and the word lattice decoding algorithm. Hence, the proposed approach tried to integrate the dynamic information such as the recently selected context, the user' s recent profile, the automatic prediction algorithm and the machine learning technology to improve the performances of the Pinyin-to-Chinese-Character conversion. This report not only contributes some new techniques for Pinyin-to-Chinese-Character conversion research, but also

tests their effectiveness in Chinese Input system and Chinese speech recognition system.

The research is partly supported by the National Natural Science Foundation of China under project "The Study of Non-Linear Features of Speech Based on Re-producing Kernel" with grant No. 60572125. This project is to investigate new features of speech and exploit them in speech recognition with the hope of promoting the performance, especially the accuracy, of Mandarin speech recognition systems. In the early 1990s, the author started the research on Chinese information processing. In last few years, many works have been finished by his research groups. Some of their papers and reports have been published by some important domestic and oversea publications or proceedings, such as Journal of Chinese Information Processing, Journal of Software and Int. Conference on Audio, Speech and Signal Processing.