语句拼音-汉字转换的智能处理

章森　宗成庆　陈肇雄　黄河燕
(北京 中国科学院计算技术研究所, 100080)

摘要　语句拼音-汉字转换是中文信息处理研究的一个重要方面，是键盘汉字输入和语音汉字输入的核心技术. 其主要特征是对动态输入的拼音串进行词法分析，给出所有可能的汉字句子，然后对这些汉语句子根据上下文环境进行句法分析和语义分析，动态调整句子中的字词，输出最佳结果. 近年来，语句拼音-汉字转换系统大量应用了人工智能技术和机器翻译的理论，以期提高系统转换的准确率和增强系统的智能处理功能. 本文分析了语句拼音-汉字转换系统所采用的核心技术，即知识支持、自动分词和动态调整等，讨论了计算机进行语句拼音-汉字转换的处理方法和过程，知识库的组成结构，用于拼音串自动分词的算法和实现，音字转换中动态调整的概率模型等. 本文还分析了现有语句拼音-汉字转换系统在拼音串自动分词和音字转换的动态调整中发生错误的原因，并提出了改进方法.

关键词: 自动分词，音字转换

Intelligent Approach: From PINYIN to Chinese Characters

Zhang Sen, Zong Chengqing, Chen Zhaoxiong, Huang Heyan
(IMT Research Center, Inst. of Computing Tech., CAS, Beijing, P.R.China, 100080)

Abstract: Transcription from PINYIN to Chinese characters is an important field of Chinese information processing and the kernel of Chinese input systems. In this paper, the knowledge supporting, automatic segment and replacing words on-line techniques were discussed. The intelligent approach based on statistic and rule method was presented and analyzed.

Key Words: Automatic Segment, PINYIN to Chinese Characters' Transcription

一、引言

语句拼音-汉字转换是中文信息处理研究的一个重要方面，是汉字进入计算机的基础，是键盘汉字输入和语音汉字输入的核心技术. 九十年代以来，拼音-汉字转换的研究已从单字转换和词语转换发展到了以句子为单位的转换[2]. 在实现方法上也大量应用了人工智能技术和机器翻译的理论，以期增强系统的智能处理功能. 近几年，国内外出现了一批这类系统，如台湾倚天公司的忘形输入系统，哈尔滨工业大学研制的音声汉字语句输入系统 INSUN，北京隆光威尔公司开发的低冗余序列中文输入系统 AUTOWAY 等.

语句拼音-汉字转换系统允许连续地输入一长串拼音代码，或一个句子的拼音代码．系统对输入的拼音代码串进行分析理解，并进而转换成相应的汉字串．这一处理思想符合人们的思维和记忆习惯．所以，人们期望语句拼音-汉字转换系统比以字词为单位的系统转换准确率更高，智能化程度更高．

由于计算机技术的迅速发展，存储量和响应速度已不是目前实现语句拼音-汉字转换的主要问题．现在对语句拼音-汉字转换的研究主要集中在如何提高系统转换的准确率上．但由于汉字是表意文字，其注音又采用西文的拼音文字，所以对汉字的分析理解必须是音形意等多维交叉信息的分析及综合．这也是实现语句拼音-汉字转换的根本问题所在．另外，还有如下一些问题：

．汉语的句子是由一串连续的汉字组成，词与词之间无空格或其它标志．所以对汉语句子的分析理解必须首先对句子进行词语切分处理．同汉语句子一样，汉语拼音串也存在词语的切分问题，而且其分词模糊性比汉语句子更大[1,3]．

．在汉语中，同音字和同音词很多，而且分布很不均匀．平均每个有调音对应 5~6 个汉字，但有些音对应的汉字达一百多个[1]．

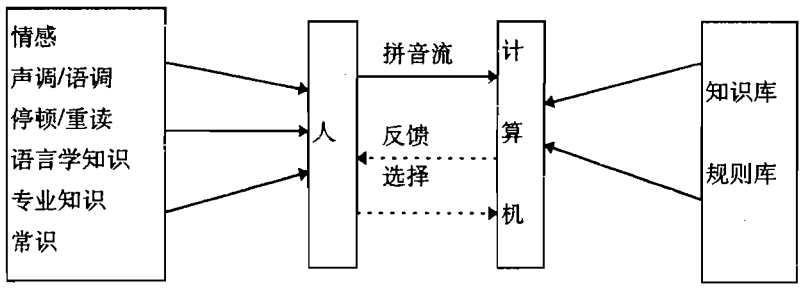
．汉语缺少形态变化，其语法尚未形成规范化，而且人们习惯使用非规范的句子．因此，汉语语法及语义的研究和应用对提高语句拼音-汉字转换系统的转换准确率影响较大．

二、知识支持

分析人对汉语句子的理解过程对实现计算机的语句拼音-汉字转换是有启示的．人在处理一句话时所接收的信息是多维的，有语音和语调信息，有句子的停顿信息和重读信息，可能还有外部表情和身体语言，甚至可能还有某些暗示信息等．人对这些信息，利用个人已有的语言学知识、专业知识和常识等进行分析综合，来推断句子的意义．在处理过程中，随着新信息的加入，人在不断学习，对前一步的结果不断进行修正，从而得到最后的结果．

计算机的语句拼音-汉字转换过程与人对自然语言的理解过程是类似的．但计算机所接受到的信息只是一串有调或无调的拼音代码，没有其它更多的信息．计算机对拼音串分析处理时所能依据的只是大容量的规则库和词语库等．但计算机缺少对外部世界多种信息的感知和接受能力，缺少世界知识和常识等．所以，虽然计算机有比人快得多的计算能力和大得多的存储能力，但由于它接受的信息少，缺乏人所具有的许多知识，其语句拼音-汉字转换的输出往往与人们期望的不一致．

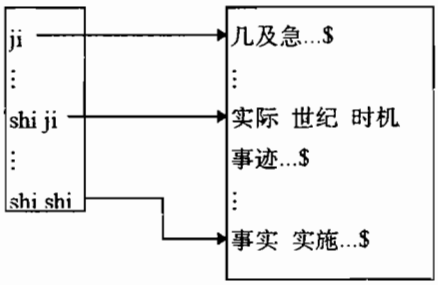
计算机的语句拼音-汉字转换是人和计算机相互交互的过程．首先是人输入给计算机的拼音代码流，然后是计算机反馈的候选汉字并等待人的选择．人工选择表明计算机仍然把人作为重要的外部知识源．整个交互处理过程如下：



由上述计算机进行语句拼音-汉字转换的处理方法和过程可以看出，词语库和规则库是系统进行音字转换的基础. 词语库主要用于对拼音流的分词阶段，而规则库主要用于对音字转换的所有句子进行句法分析和语义分析阶段，排除不合理的句子，给出最佳结果. 词语库和规则库都由静态库和动态库两部分组成. 动态库是为加强系统的机器学习功能而设置的. 因为语句拼音-汉字转换系统将处理的都是开放语料. 无论系统的静态库有多大，它总是不完备的. 所以，动态库对提高系统的性能是必要的.

2.1 词语库的组成和结构

静态词语库主要包括国标单字(6763 个)和常用词语(2 万左右)等. 动态词语库主要包括系统运行中动态切分出的而不包括在静态词语库中的词语. 这两个库的结构有所不同，主要是为方便系统的存取而考虑的. 在静态库中，每个拼音串词语后面跟一个指针，指向它对应的一串汉字词语. 静态库的结构如下所示:



其中，$表示一个拼音串对应的汉字词语集合的结束位置. 在静态库中，一般限定词语的最大长度. 据文献[1] 统计，汉语中词语的最大长度可限定为 7 字长. 所以静态库中的词语长度一般定为 1~7，而两字词所占的比例最大，约大于 70%.

动态词语库比静态词语库一般要小得多，其结构也比静态词语库简单，拼音串空间和汉字串空间一般都是定长的. 这样设计主要是从系统对动态词语库的存取速度方面考虑的. 在动态库中也存在一个拼音串对应多个汉字串的情况，这时，用 HASH 表比较好. 下图是动态词语库的结构示意图(其中****表示指向下一个的指针，FFFF 是结束标志):

拼音串	汉字词语串	标志
shi xian	实现	****
****	事先	FFFF

2.2 规则库的组成

规则库主要包括各种语法规则，在系统实现时，为了便于规则的存取、修改和新规则的加入，一般分为动态规则库和静态规则库两部分. 建立规则库的目的是把汉语句子中的单字和词语与语法和语义信息联系起来，从而对句子中的单字和词语加上一定的约束和限制. 汉语的特点是词语无形态变化，词语组句的方式非常灵活，有些词语可以放在句子中的任何位置. 所以，汉语的规则比较复杂，规则的形式与系统进行句法分析和语义分析时采用的

具体文法有关. 一般形式如下:

关键字--->语境测试条件|结果

对汉语句子进行分析处理时, 依次取句子中的词语与规则中的关键字进行匹配. 如匹配成功, 则检查当前语境是否满足语境测试条件. 若满足, 则给出该规则的结果; 否则, 该规则匹配失败, 取下一条规则.

动态规则库是为存储系统建立的新规则. 因为任何描述自然语言的规则库都不可能是完备的, 所以规则库必须在不断地使用中加以完善. 规则库中同类规则的优先级可以根据约束的强弱来确定, 也可以根据使用的频度来确定. 规则库的结构可按如下算法建立:

1). 从规则中提取关键字 KEY;

2). 计算关键字 KEY 的散列值 HASH(KEY);

3). 将规则的右部, 即语境测试条件|结果存入 HASH(KEY)所指的单元.

三、自动分词

自动分词包括两个方面: 一是对汉语文本的自动分词; 另一是对拼音代码串的自动分词. 语句拼音-汉字转换系统用到的是后一种分词. 这两种分词从表面上看相当于对两种不同语言的分词, 但实际上它们具有内在的联系. 文本分词是拼音代码串分词的依据和基础, 而拼音代码串分词可以作为文本分词在某种意义下的抽象模型. 由于拼音代码串的模糊度比汉语文本的更大, 所以, 拼音代码串分词的复杂度要比汉语文本的更大. 无论是对拼音分词, 还是对文本分词, 目前最关心的是分词的准确度, 时间和空间已不重要.

实际上, 汉语文本的自动分词的问题到目前为止还没有彻底解决[4], 对人名、地名、新词和歧义字段的处理还不完善. 拼音代码串的自动分词不仅具有汉语文本自动分词的所有问题, 而且歧义现象更严重, 拼音代码之间错误组词的概率远远大于汉字间组词的概率.

分词算法对分词精度有较大影响. 任何分词算法都不可避免地会发生分词错误. 应用于拼音串自动分词的算法主要有: 最大匹配法(MM)[1]、最少分词法[1] (FWM)和逐词遍历法[1]. 根据扫描方向的不同, 最大匹配法又分为正向最大匹配法(FMM)和反向最大匹配法(BMM). 最大匹配法只给出唯一的分词结果, 这对拼音串分词是不合适的. 据[4] FMM 和 BMM 的切分正确率都在 90%左右. 但由于 MM 法算法简单, 容易实现, 所以仍得到广泛应用. 在语句拼音-汉字转换系统中, 可以同时利用 FMM 法和 BMM 法切分同一拼音串, 当二者的切分结果不一致时, 利用概率统计的方法进行选择. 这样做可以使切分正确率达到 99%左右. 以下例子是 MM 法不能正确处理的情况.

例 1. FMM 和 BMM 法切分的结果均错, 约占 0.05%.

(正) yuan zi/ jie he/ cheng/ fen zi 原子/结合/成/分子.

(误) yuan zi/ jie he/ cheng fen/ zi 原子/结合/成分/子. (FMM)

(误) yuan zi/ jie he/ cheng fen/ zi 原子/结/合成/分子. (BMM)

例 2. FMM 和 BMM 法切分的结果之一错, 约占 9.24%.

(正) jin tian/ shi/ xing qi/ wu. 今天/是/星期/五.

(误) jin tian/ shi xing /qi wu. 今天/是星/期五/. (FMM)

例 3. FMM 和 BMM 法切分的结果相同，但均错，约占 0.41%.

(正) fan ying/ le /yi ge/ ren /de/ jing shen /mian mao 反映/了/一个/人/的/精神/面貌

(误) fan ying/ le /yi/ ge ren /de/ jing shen /mian mao 反映/了/一/个人/的/精神/面貌

最少分词算法是以切分后得到的词数最少为原则的，这符合大多数拼音串的分词情况. 显然，该算法给出的分词结果可能是不唯一的，但它并没有给出所有可能的情况. 最少分词法可以通过比较多个分词结果的词频，正确处理许多最大匹配法无法处理的交集歧义. 以下例子是 FWM 法不能正确处理的情况.

例 4. (正) zhe/ shi/ yi/ suo/ da xue 这/是/一/所/大学.

(FWM) zhe/ shi yi/ suo/ da xue 这/是一/所/大学.

例 5. (正) ran er shi you yi yi de 然而/是/有/意义/的.

(FWM) ran er shi you yi yi de 然而/是有/意义/的.

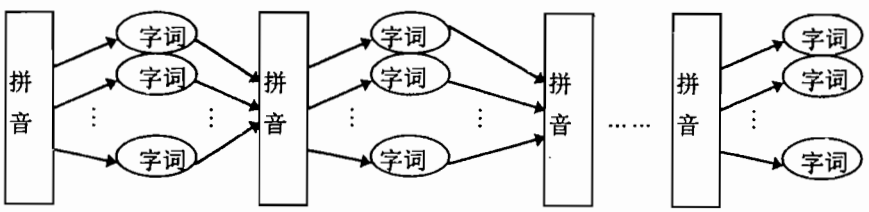
逐词遍历法是为了克服最大匹配法和最少分词法都会丢失某些分词结果的缺点而提出来的. 给出拼音串的所有可能的分词结果是为了在句法平面和语义平面作进一步处理，以便给出正确的结果. 易见，逐词遍历法是一个 NP 问题.

对分词算法的改进以及对知识库的完善不能期望在分词阶段完全解决分词问题. 计算机自动分词问题的解决依赖于更高层次的知识的应用，如语法知识和语义知识等. 计算机分词与人分词的过程不一样. 人是先理解后分词，或边理解边分词，而计算机是先分词后理解. 分词问题的彻底解决还要建立在对原文句子理解的基础之上. 这正是计算机自动分词的困难所在.

四、动态调整

影响拼音串到汉字串精确变换的主要因素是汉语同音字词的大量存在. 汉语中无调音节约有 400 个，有调音节约有 1300 个，而 GB2312-80 规定的汉字数量是 6763 个. 文献[1] 中统计的汉语常用二字词为 28000 多条，三字词约有 3700 条，四字词约有 3600 条. 一般地，每个汉语有调音节对应 5~6 个汉字，每个无调音节对应约 16 个汉字. 因此，拼音-汉字转换实际上是一对多的映射，而动态调整的目的是依靠句子中的信息将这个一对多的映射转化为一对一的映射.

语句拼音-汉字转换的动态调整算法一般可分为两类：一类是基于概率统计的方法，另一类是基于规则的方法. 概率统计方法首先对拼音串中的每个拼音给出所有相应的字词. 根据对应的字词的概率，选择其中概率最大者作为候选，再考虑句子的上下文信息，调整候选的字词，给出变换的结果，如下图所示：



由上图可知，整个变换过程就是选择一条从始点到终点的最佳路径. 求解这条最佳路径可采用动态规划算法. 如果把上述每步变换都作为一个状态来考虑的话，就可以应用MARKOV 模型的方法求解最佳路径. MARKOV 模型主要考虑前一状态的选择对其后的状态选择的影响，也就是前一字词选定时其后字词被选择的概率. 该方法的缺点是明显的:它只是利用了部分上文信息. 另外，该方法不仅存在盲点(概率小的字词可能永远不被选取)，而且对边界初值比较敏感，可能发生一错就错一串的情况.

以下给出一些错误变换的实例.

例 6: qing ba shou ju qi lai!

(正) 请把手举起来! (误) 轻把手句起来!

例 7: ta ji cong ming you piao liang.

(正) 它及聪明又漂亮. (误) 它几聪明邮票量.

对以上系统中发生错误调整的地方，系统要进行人机交互操作才能解决. 人们总希望这样的交互操作越少越好.

基于规则的动态调整方法是对整个句子进行操作，根据语法规则库先对变换来的汉语句子进行句法分析，排除不合语法的句子. 如果剩下的句子不是唯一的，则再对句子进行语义分析，排除不合语义的句子. 实际上是根据每个词的词性和词义检测该词在特定上下文环境中是否合理. 如果剩下的句子还不是唯一的，则对句子中不确定的字词选择使用概率最大者，确定唯一的句子作为输出结果. 此时，如果使用条件概率，即 HMM 方法，则比简单地应用字词概率的选择方法会更好.

五、结束语

语句拼音-汉字转换系统的智能处理越来越重要，其中的许多问题还处于研究探索阶段，不能期望语句拼音-汉字转换系统的精度能达到百分之百. 知识库和规则库不可能完备，拼音串的自动分词不可能百分之百正确，动态调整不可能百分之百正确，因此，最后的输出结果也不可能百分之百正确. 智能处理是提高语句拼音-汉字转换系统精度的有效途径，基于规则的动态调整算法和 HMM 方法的结合值得进一步在实践中检验.

参考文献

[1] 刘源 等，信息处理用现代汉语分词规范及自动分词方法，清华大学出版社,1994

[2] 王小龙 等，声音语句输人的研究，计算机学报，Vol.17, No.2,1994

[3] 梁南元，汉语计算机自动分词知识，中文信息学报，Vol.4, No.2,1990

[4] 黄昌宁，中文信息处理中的分词问题，语言文字应用，No.1, 1997