一个拼音汉字自动转换系统的设计与实现

成　华　　尹宝林

(北京航空航天大学计算机科学与工程系)

摘　要：同音词识别是音字转换研究中的主要问题．本文建立了一个带调拼音语句到汉字语句的自动转换系统．它由五个部分组成：自动分词模块、词法分析模块、句法和语义分析模块、语义修饰模块和显示模块．其中句法分析是系统的核心，我们在比较各种方法的基础上选择了扩充转移网络(ATN)模型，使ATN的优点和音字转换的特点得到较好的结合．测试结果表明，系统的正确转换率达到97%．

关键词：拼音汉字转换，扩充转移网络模型．

THE DESIGN AND IMPLEMENTATION OF A PINYIN—CHINESE WORD CONVERSION SYSTEM

Cheng Hua　Yin Baolin

(Dept. of Computer Science & Engineering, Beijing Univ. of Aeronautics and Astronautics)

ABSTRACT: The recognition of homonyms is the main problem in Pinyin to Chinese word conversion research. This paper builds a system which can convert Pinyin sentences with tones to Chinese character sentences. It is composed of five parts: auto-split module, morphology analysis module, syntax analysis module, semantic analysis module and display module. Syntax analysis is the most important part in the system. After studying various methods carefully, we choose Augmented Transition Network(ATN) model as the base. We combine the trait of Pinyin to Chinese word conversion and the advantages of ATN model relatively well. The test result show that the system has a correct conversion rate of 97%.

Key words: Pinyin to Chinese word conversion, Augmented Transition Network(ATN) model.

一、引　　言

拼音汉字转换是连续汉语语音识别项目的一个组成部分，用于实现带调拼音语句到汉字语句的转换。由于汉语中音节数远远少于汉字个数，因而存在着严重的重音现象，同音词识别就成了音字转换中的关键技术。

模仿人与人之间的语音交流方式，在计算机汉语语音识别过程中，从声音到文字的输入过程分为语音识别（把声音信号转换为汉语拼音形式）和语音理解（把拼音语句转

换为汉字语句）两个阶段。可见音字转换作为汉语语音输入的后处理过程，是一个必不可少的部分。音字转换的另一个用途是作为智能化键盘汉字输入系统的核心部分。这种输入系统能在用户输入一句拼音码后给出正确的汉字句子，而不需用户过多地参与选择，是一种与传统的编码和拼音选择输入方法相比更加快捷方便的输入方法。不难看出，研究拼音汉字转换具有较大的实用价值。

目前这方面的研究已广泛开展，类似的系统主要有：杨(1987)[1]的系统由两万词库和若干条词法及短语规则组成，用ATN网络进行句法分析，系统可识别100多种句型，识别率达92%。汤(1989)[2]在ATN模型的基础上对较复杂的汉语句子进行了分析，但系统词典只含8000个常用词汇，识别率达96%。王(1989)[3]的系统提出了语音代码的最优分词算法FWF，语法分析以词法、短语和句法规则为基础，并通过机器学习来提高转换正确率，对于带调语音代码的正确转换率达98%。万(1992)[4]的FPY系统采用规则的自上而下分析技术，并利用句法相关的语义识别技术，使动宾、动补和数量短语得到很好的识别，同音词识别率为97.5%。虽然目前各种系统纷纷涌现，但还没有一个能达到真正实用，这一领域还有许多问题需要解决。

二、拼音汉字自动转换系统PCACS的总体设计

我们认为，就拼音汉字转换研究目前所能达到的转换正确率，用于实际应用领域是不够的。但如果加上良好的交互式界面，则可以成为一个比较实用的智能化键盘汉字输入系统。因此，PCACS由系统词典、词法规则库、汉语ATN语法网络和五个相对独立的处理模块组成。这五个模块是：拼音串自动分词模块、汉语词法分析模块、句法和语义分析模块、语义修饰模块以及显示模块。系统的总体流程示意图见下页。下面对系统的关键部分作详细介绍。

1 词典的组织

系统词典以北航“现代汉语词频统计”工程[5]所收录的词为基础，从中抽取近四万较常用的词条，加以分类整理，按词长分为四个词典。每个词典内词条按其拼音码排序，拼音相同的按词频从大到小排序。

为了加速对词典的查找，为词典建立了一级索引。

词典中每一个词条都具有下列一些信息：汉字码、词频、带调拼音码、词性、词的语义特性。

我们把汉语词汇分为十九类，即：

名词(N)、时间词(Nt)、方位词(Nl)、一般动词(V)、能愿动词(Vn)、趋向动词(Vq)、一般代词(P)、指示代词(Pd)、疑问代词(Pi)、数词(M)、量词(Q)、形容词(A)、副词(D)、介词(R)、结构助词(Hs)、时态助词(Hm)、连词(C)、句子连词(L)、叹词(I)。

对于其中的一些词类又进行了细化，作为词的语义特征。这对于进行汉语词法、句法和语义分析非常重要。

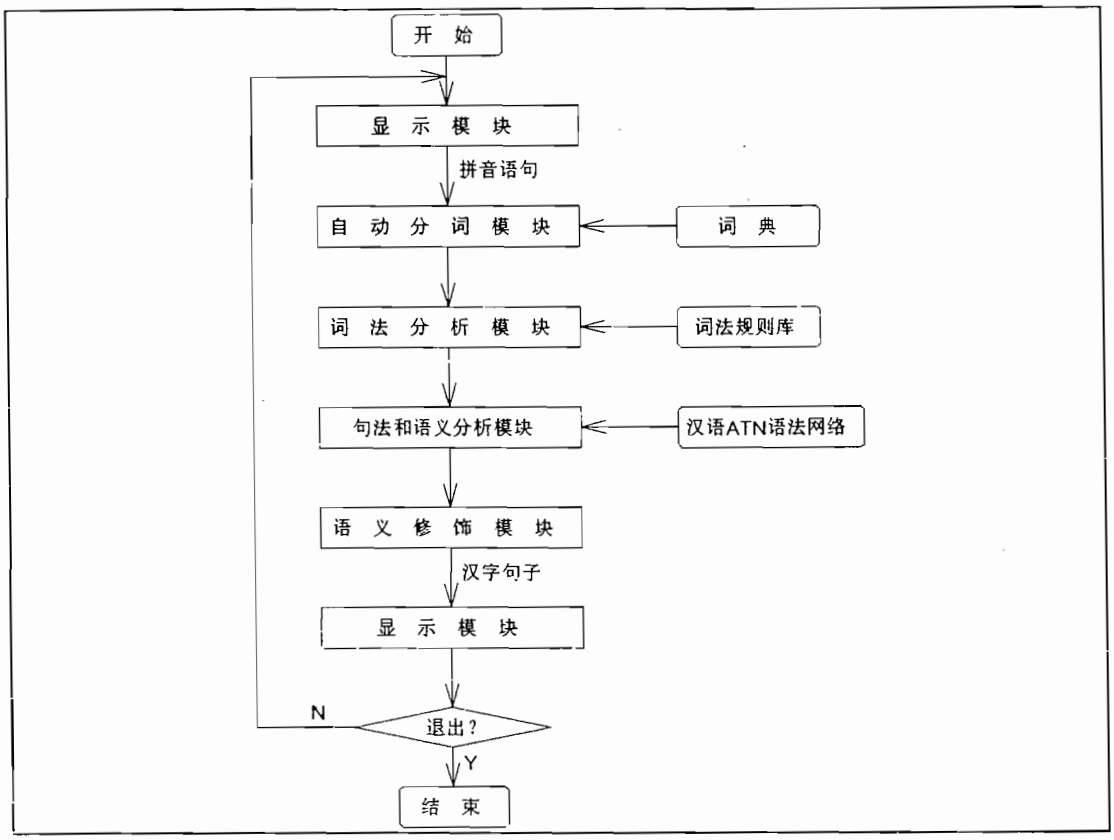


图: PCACS的总体流程示意图

2 系统的词法分析技术

汉语中词性的修饰关系不是任意的，这种关系构成了词法分析的基础．在对汉语词汇分类的基础上，我们总结了约100条词法规则，包含了可能成为相邻词的各种词类搭配．如果一个词不能与它前面相邻的任意一个同音词相搭配，并且也不能与它后面相邻的任意一个同音词相搭配，则应当把这个词删去．由于没有同音词的词其词性相对确定，每次都从这些词开始检查词类是否匹配可以最大限度地消除同音词或词类。

另外，对具有相同属性并且出现概率很大的词，如结构助词“的、得、地”，我们使用一些特殊的规则来处理．

3 系统的句法和语义分析技术

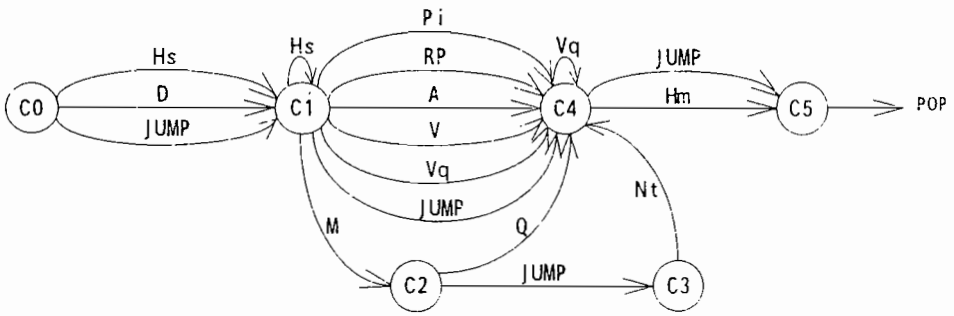
句法分析是拼音汉字转换的关键．目前流行的句法分析方法，如FUG、LFG文法等多是在词典中增加有关词的约束信息以减小语法分析的复杂度．但由于拼音汉字转换中某个拼音所对应的汉语词汇是不确定的，无法进行相关词的限制，且不便于进行回溯，所以我们选择了扩充转移网络模型[6]．它的优点表现在以下几个方面:

1. 扩充转移网络在递归转移网络的基础上添加了“条件”、“动作”和“成分寄存器”，可以方便地进行相关词汇的匹配和句型的限制。更重要的是，ATN模型突破了语境自由语法的局限，使语法信息和语义信息在句子分析过程中可以很好地结合起来，比较适合音字转换的需要。

2. ATN是基于图论数学概念的应用和语法研究的有限状态机，它既可以看作是一种语法的形式化，又可看作是一种机器，易于转换成计算机程序。

3. 利用ATN语法分析句子的过程与人们理解自然语言的过程非常相似。用于句法分析的汉语ATN网络给人一种一目了然的感觉。

基于以上几点考虑，我们选择了扩充转移网络模型，并根据汉语句子中意群的语法功能，在系统中建立了一个汉语句子的ATN网络和九个汉语ATN子网络，它们是：方位词分析子网络(TP)、名词代词分析子网络(PP)、数量词分析子网络(MP)、形容词副词分析子网络(AP)、动词分析子网络(VP)、名词短语分析子网络(NP)、介词短语分析子网络(RP)、状语分析子网络(YP)、补语分析子网络(CP)。下面给出了系统中的一个ATN模型—CP。



如何把ATN的分析能力与音字转换的特点相结合，以求得较高的汉语句法分析效果，是我们研究的主要目标。为此，我们在句法分析过程中使用了以下技术：

● 高频词优先

汉字和词汇的一个重要的统计特性就是字频、词频集中，为此我们在词典中把使用频度高的词放在其同音词序列的前面。另外，有些字虽然常用，但很少单独使用，则在字典中把这些字排在其同音字序列的后部，而把出现频率高且常单独使用的字放在前面。在进行句法分析时，我们没有采用以往常用的规则库的方法，而是取同音词序列中最前面的一个词，以其词性为依据选择匹配的弧路进行分析。如果可以配匹，则保留与第一个词词性相同的词，把其余的词放到回溯栈中。当分析失败发生回溯时，从回溯栈中取出最上面的词继续分析过程。由于每次都是取可能性最大的一个词的词性来匹配，就把回溯的可能性减到了最小，而且也避免了对规则库中的规则进行排序的问题，这样系统的分析速度自然也就提高了。

● 增加测试条件

为了尽可能减少回溯，我们在有可能走错的弧路上设置了测试条件。如动词短语的主要语法功能是作谓语，但它也可和“的”构成“的”字短语作主语或宾语，那么在分析到动词短语的时候，需要向前查看几个词，如果不构成“的”字短语，再把它按句子

的谓语来处理。这种测试对于减少回溯，提高句法分析的正确性以及改善分析效率都是很有效的。

● 句法和语义分析并行

我们认为语法和语义是不可分割的，句法分析既需要语法知识，也需要语义知识。实际上，语义可以看作是对语法的细化和进一步的限制，这样两种知识就可以应用到一个分析过程中去了。

为此，我们对词汇成分作了细分类，把分类的结果作为语义信息，即进一步的句法限制条件，及时消除违背这些限制条件的同音词，同时也提高了句型匹配的正确性。如动词可分为不及物动词、单宾及物动词和双宾及物动词，如果一个句子中的谓语由不及物动词充当，则消除其后有名词或代词词性的同音词；如果一个句子有宾语，则应去掉动词同音词表中具有不及物动词类别的词。

另外，系统对一些常用的汉语关键词汇的连接。如："宁可…为好"、"被…给"等句式结构进行特殊处理。因为这些固定搭配中的句法结构常常是一定的，通过关键词将各部分的句法结构连接起来就可以形成整个句子的结构了。

4 交互式界面的设计

为了使系统接近实用化，我们在界面设计过程中尽量方便用户的使用。系统的用户界面主要由输入窗口和输出窗口两部分组成，提供三种拼音语句输入方式，即单句输入、调入已存在的拼音文件、现场编辑拼音文件。对于不正确的转换结果，用户只要在相应的汉字或词处按鼠标的中键或键盘上的指定键，系统就会显示一个包含其同音词列表的菜单，用户可移动光标选择正确的结果，它会自动取代原来的内容，从而得到正确的转换结果。

三、结　论

PCACS是一个限定句型的拼音汉字自动转换系统，它利用汉语词法、句法和语义知识完成汉语拼音语句到汉字语句的自动转换。该系统即可以作为汉语语音识别的后处理过程，又可以作为智能化键盘汉字输入系统的核心部分，具有较高的实用价值。

PCACS中现已建有近40000现代汉语常用词。系统软件采用C语言编写，在SUN工作站上运行。通过对《实用现代汉语语法》[7]和《汉语五百句》[8]中的两百个句子进行测试的结果表明：系统能处理约一百种汉语常见句式(包括句子成分的各种变化)，并能对比较复杂的汉语句子中的同音词进行分析，正确转换率达97%。转换基本实时。例如系统可以处理这样的句型：

● zhe4liang4zi4xing2che1yu3qi2zhe4me5fan3fu4de5xiu1li3，bu4ru2huan4yi1liang4xin1de5.
这辆自行车与其这么反复地修理，不如换一辆新的。　(复句)

● tong2xue2men5shang4wan2ke4zou3chu1jiao4shi4lai2dao4cao1chang3shang4.
同学们上完课走出教室来到操场上。　(连谓式)

● ta1jiang3de5gu4shi5hen3neng2xi1yin3zhu4hai2zi5men5de5zhu4yi4li4.

他讲的故事很能吸引住孩子们的注意力。　　(复杂结构作主语)

另外，对于样本集外的日常用语，系统也能得到较好的转换结果。

由于不同属性同音词的存在，使得6%的句子分析产生了歧义。这种歧义可以通过在词法分析部分附加更严格的限制条件或在句法分析之前增加一个长距离词性约束分析过程来减少。目前，系统对一些同音词还无法区分，如“形势”和“形式”、“作”和“做”等。

参 考 文 献

[1] 杨长生，何志均，汉语同音词汇辨析，计算机研究与发展，1987.1

[2] 汤建华，徐近霈，利用句法语义循环递归网络实现汉语拼音—汉字转换，中文信息学报，Vol.3, No.4, 1989

[3] 王晓龙，音字流切分及其相互转换的理论研究与系统实现，哈尔滨工业大学博士论文，1989.1

[4] 万建成，FPY中的同音词智能识别方法，中文信息学报，Vol.7, No.2, 1993

[5] 刘源，梁南元，汉语言处理的基础工程—现代汉语词频统计，中文信息学报，1986.1

[6] Woods, W.A.,Transition Network for Language Analysis, Communications of the ACM, Vol.13, No.10, 1970

[7] 刘月华，潘文娱等，实用现代汉语语法，外语教学与研究出版社，1983

[8] 林杏光，汉语五百句，陕西人民出版社，1980