统计分析方法相结合的小型拼音—汉字智能转换系统 PCTS

郑嵘　梁晋清　何厚存

（上海交通大学计算机科学及工程系）

摘　要：本文介绍了一个小型拼音汉字智能转换系统 PCTS（Pinyin—Chinese Translation System）的设计和实现，详述了它的设计思想和转换实例，并对它的特点和问题进行了讨论。该系统采用统计和分析相结合的方法，利用小规模的机器词典和规则库，处理输入的拼音语句，最终简洁高效地输出相应的汉字语句。

关键字：机器词典　产生式规则　词法分析

A Small Pinyin—Chinese Translation System Using Statistics and Analysis Methods

Zheng Rong, Liang Jingqing, He houcun

Department of Computer Science and Engineering,

Shanghai Jiaotong University, P. R. China

Abstract: This paper introduces a process of designing and implementing for a small Pinyin—Chinese translation System. It will describe the system's design idea and some translation examples. It will also discuss the features and problems in the system. The system deals with the input pinyin sentence using a small machine dictionary and a database of production rules, statistics and analysis methods, efficiently outputs a corresponding chinese sentence.

Key Word: machine dictionary, production rules, morphology analysis

一、引言

汉字计算机处理技术的研究已有十多年的历史。在此过程中，它体现出三个重要特征。一是从单纯的字编码发展到词处理，二是从简单的编码输入发展到智能识别，三是拼音编码又占据了主导地位。后者是前两者发展的必然结果。虽然拼音编码存在同音严重的根本弱点，但由于它易学易普及，在词处理技术的推动下，拼音编码已成为实用面最广的方法。拼音—汉字变换词处理技术的发展，集中反映了人们将解决中文输入难题的希望再次投向拼音的发展趋势。

但是，简单的拼音—汉字变换的词处理，并没有解决根本问题。例如存在单字词识别的障碍。于是，形成了以拼音为编码，以人工智能的自然语言处理为技术，以计算机自动识别同音词为目标的智能拼音—汉字变换的输入技术研究。它是汉字的键盘和语音输入研究的重要基础。同时，它在中文语音识别后处理研究中也是一个关键问题。

拼音—汉字智能转换研究一般走两条路线，一是从词的统计信息着手的统计方法，一是从词句法和语义分析着手的分析方法[1]。

统计方法识别同音词的依据是各词语在实用中的使用统计信息。基本的识别方法是：利用这些统计信息建立语言识别的统计模型，将输入的语音代码串以特定方式切分成单词的序列，然后根据某种判别原则，选择最有可能的切分和最有可能的汉词作为转换的结果。

统计方法理论上较成熟，系统开销小，易于实现。不足之处是，对上下文约束的利用不充分，只能处理相邻出现的同音字词，不能或很难实现更复杂约束条件下的识别。另外，由于是基于语料的，因而识别率受到具体应用环境的影响。

分析方法不同于单纯词处理方法，它在对汉语词系统分类（属性标注）的基础上，应用自然语言处理的词法，句法以及语义识别技术，试图达到同音词的大范围的基本唯一识别。

分析方法的识别率一般高于统计方法，但这是以大规模的规则库，复杂的分析过程和一定的延时反应换来的。汉语词法句法和语义纷繁，复杂，至今尚无统一认识，给系统的高效实现带来一定的困难。

基于一认识，本文所述系统 PCTS 利用小规模的机器词典和规则库，采用统计和分析相结合的方法，互相补充，扬长避短，对输入的拼音语句进行处理，以期简洁高效地输出相应的汉字语句。

二、系统分析

1. 拼音语句的切分

本系统采用“最小匹配法”FWM[2]，对输入的拼音语句进行初步的切分。由于后期的分析工作所需考虑的因素多，难度大，因此尽可能在初期的拼音分词阶段做较多的信息处理，有助于减少整个语言理解过程的开销。换句话说，对拼音语句分词后得到的词数越少越易于对该语句的理解。

最小分词问题可抽象为求有向图两点间最短路径问题。设输入的拼音语句字数为 n，结点数则为 n＋1，对应这些拼音词语的词典中的词均以边的形式出现，边的权都为 1。这样，求切分后的最小词数就等同于求从始点到终点的最短路径。

例如，某拼音语句有 13 个字：″jizhongbinlibaoweishouduhezhuyaoshengcheng″。系统转换的最终结果希望输出“集中兵力保卫首都和主要省城”。切分阶段该拼音语句的表现形式为 13 个字：字 1(ji)，字 2(zhong)，字 3(bin)；字 4(li)，…，字 13(cheng)。14 个结点：结点 1，结点 2，…，结点 14。

2. 统计方法的使用

由于分词阶段有可能产生若干种切分方式，所以必须对各种切分方式进行筛选，以减少后面语法分析的工作量。这时词典中的词频信息是最好的利用对象。因为，经统计而获得的词频信息代表了在普遍状况下或专业领域中词语的使用情况。当系统的输入是拼音语句而不是毫无关系的一串拼音时，便暗示了各个词语间存在一定的内在联系。

针对每种切分方式，对该切分方式中的各个拼音词语，选取词频最高的汉字词语作为初步

的转换结果。研究这些词频，筛选出最佳的切分方式。这是基于如下的考虑：如果某切分方式中各词语的词频的最大值比其他切分方式小，或者各词语的词频的最小值比其他切分方式大，那么有可能切分得不太恰当，便可考虑筛去。

这样，有效地利用统计信息，筛去不恰当的切分方式后，留下的待分析的切分方式便不多了，后面的语法分析的工作量减少，系统的延时降低，效率提高。

3. 分析方法的使用

利用规则库，对某切分方式中的汉字词语串进行词法分析和句法分析。如果分析成功，则将该汉字词语串作为最终结果输出。否则，在出错处选取次高频的汉字词语继续进行分析。

由于前面已使用统计方法对中间结果进行了精简，所以分析阶段可以比同类的其他系统相对地降低一些难度。规则库不需要包罗万象，可以构造得比较简单，缩小规模，减少分析阶段的时间和复杂度。

本系统利用 GPSG [3] 构造规则库，采用 TOMITA 算法[4]实现语法分析。

三、系统实现

1. 汉语机器词典的设计

本系统利用“现代汉语常用词词频词典”[5]，建立包含 46520 条记录的数据库文件 DIC.DBF，每条记录含有若干域：汉字词语 WORD，拼音 PINYIN，词频 FREQUENCY，范畴 CAT1，范畴 CAT2，词法属性 MORATTR (morphological attribute)，句法属性 SYNATTR (syntactic attribute)，语义属性 SEMATTR (semantical attribute)，等等。记录按拼音，词频排序。系统实际使用的词典则是由 DIC.DBF 转换而来的文本文件 DIC.TXT。

2. 语料库的设计

本系统要求的输入为拼音语句。为了便于测试和调试，建立了相应的语料库 TEST.DBF，每条记录含有汉字语句域 SENTENCE 和拼音语句域 PY。为了方便，同样也将该语料库转换为文本文件 TEST.TXT。

3. 拼音语句的切分

经最少分词后，输入的拼音语句具有若干种切分方式。例如，拼音语句“jizhongbinlibaoweishouduhezhuyaoshengcheng”经预处理后，变为“ji zhong bin li bao wei shou du he zhu yao sheng cheng”，其切分方式有：

段 1：〈段长 4〉〈词数 1〉〈起始字 1〉〈路径数 1〉

路径 1：1 5 …… jizhongbinli

段 2：〈段长 4〉〈词数 2〉〈起始字 5〉〈路径数 1〉

路径 1：5 7 9 …… baowei shoudu

段 3：〈段长 3〉〈词数 2〉〈起始字 9〉〈路径数 2〉

路径 1：9 10 12 …… he zhuyao

路径 1：9 11 12 …… hezhu yao

段 4:〈段长 2〉〈词数 1〉〈起始字 12〉〈路径数 1〉

路径 1：12 14 …… shengcheng

需分析的总路径数：2

其中，段长指该段所含的字数，路径表示该段的切分方式，路径中的数字代表结点号。

4. 多种切分方式的筛选

选取高频汉字词语后，结果如下：

段 1:〈段长 4〉〈词数 1〉〈起始字 1〉〈路径数 1〉

路径 1：1 5

集中兵力 0.00015 /* 词频 */

段 2:〈段长 4〉〈词数 2〉〈起始字 5〉〈路径数 1〉

路径 1：5 7 9

保卫 0.00606

首都 0.00417

段 3:〈段长 3〉〈词数 2〉〈起始字 9〉〈路径数 2〉

路径 1：9 10 12 /* 路径最小词频为 0.08167 */

和 1.05119

主要 0.08167

路径 2：9 11 12 /* 路径最小词频为 0.00012 */

合著 0.00012

要 0.31651

段 4:〈段长 2〉〈词数 1〉〈起始字 12〉〈路径数 1〉

路径 1：12 14

声称 0.00083

需分析的总路径数:2

对每一段，若路径数大于 1，则对每一路径，求出各词词频的最小值，作为路径最小词频；求出各词词频的最大值，作为路径最大词频。然后找到路径最小词频最大或路径最大词频最小的路径，作为该段留待分析的切分方式，其余的切分方式被筛去。

于是，上例经筛选后结果为：

段 3:〈段长 3〉〈词数 2〉〈起始字 9〉〈路径数 1〉

路径 1：9 10 12 /* 段词频为 0.08167 */

和 1.05119

主要 0.08167

5. 规则库的设计

GPSG 中统称词类，短语和句子为范畴。

参考[7][8]等的约定，本系统范畴如下：

范畴={名词 n，动词 v，形容词 a，数词 m，量词 q、代词 r，副词 d，介词 p，连词 c，助词 u，叹词 i，语气词，象声词，前缀 h，后缀 t，　　/*词语*/
NP，VP，PP，AdjP，AdvP，NumP，TimeP，LocalP，　　/*词组*/
S，　　/*句子*/
…　　/*其他*/
}

次范畴={名词 n：有生命 n1，无生命 n2，抽象 n3、时处 n4，方位 n5，称呼 n6，姓氏 n7，专有 n8，…
动词 v：能愿动词，趋向动词、…
数词 m：基数 m1，序数 m2，分数 m3，倍数 m4，概数 m5，系数 m6，位数 m7，…
…
}

规则构造举例如下：

词法规则：

〈人〉—〉〈姓氏〉〈称呼〉
〈人〉—〉〈老|大|小〉〈称呼〉
〈人〉—〉〈老|大|小〉〈姓氏〉
〈人〉—〉〈老|大|小〉〈姓氏〉〈称呼〉
〈人〉—〉〈姓氏〉〈老|大|小〉〈称呼〉
〈人〉—〉〈姓氏〉〈序数〉
〈人〉—〉〈姓氏〉〈序数〉〈称呼〉
〈人〉—〉〈姓氏〉〈名字〉
〈系数〉—〉一|二|三|四|五|六|七|八|九|零
〈位数〉—〉个|十|百|千|万|亿|兆
〈基数〉—〉〈系数〉
〈基数〉—〉〈系数〉〈位数〉
〈基数〉—〉〈系数〉〈位数〉〈基数〉
…

句法规则：

S—〉NP，VP
S—〉LocalP，VP
S—〉TimeP，VP
…
NP—〉VP/NP，助[的]
NP—〉S，助[的]，NP
NP—〉名，NP
NP—〉名
…

6. 语法分析的实现

举一个简单的例子，上例中，我们可以清楚地看到："shengcheng"一词，系统的首选结果为"声称"，而不是所希望的"省城"。这是由于在词典中，"声称"的词频比"省城"要高。但"声称"是动词，"省城"是名词，系统通过使用 TOMITA 算法（广义 LR）进行分析后，会作出正确的识别，重新进行选择。

7. 系统的回溯

若留待分析的各种切分方式都不能通过语法分析，则回到拼音语句分词阶段，重新进行切分，求次最长路径，继续分析。例如，对于拼音语句"woshizhongxuesheng"，系统求出最短路径切分方式"woshi zhongxuesheng"，转换成汉字语句"卧室中学生"，语法分析出错。回溯后，将"woshi"拆开，成为"wo"及"shi"，经各种尝试和分析，转换成"我"及"是"二字，输出希望的结果"我是中学生"。

四、总　　结

整个拼音—汉字智能转换处理是一个十分复杂，涉及面很广的交叉学科研究方向。其难度远远高于各种编码技术的研究，需要进行系统的分析和深入细致的工作，才有可能从根本上使中文输入技术再有一个大的突破。

本系统有效地利用了统计方法，相对地降低分析方法的难度。因此，整个系统规模不大，效率却比较高。通过对数千字的普通新闻类拼音文本的测试，发现在仅利用词频信息而未进行词法分析的情况下，识别正确率已可达到 70%以上。

但是，由于对于汉语的一些基础理论，语法界缺乏统一的标准和规范的描述，所以本系统对部分语法现象还未找到合适的解决方案，有待于以后的进一步探索。

参考文献：

[1] 万建成，"语音代码—汉字智能转换研究"，《中文信息学报》，Vol. 8，No. 2

[2] 王晓龙，等，最少分词问题及解法，《科学通报》，1989

[3] G. Gazdar，G. Pullum，"Generalized Phrase Structure Grammar"，Harvard University Press，1985

[4] Masaru Tomita，"An Efficient Augmented—Context—Free Parsing Algorithm"

[5] 刘源，等，现代汉语常用词词频词典（音序部分），宇航出版社，1990. 6

[6] 现代汉语词典，商务印书馆，1980. 6

[7] 信息处理用现代汉语分词规范，GB13715

[8] 张松林，现代汉语语法表解，四川科学技术出版社，1986. 5

[9] 仲兴国，多词组一次性拼音汉字变换，《中文信息学报》，1990

[10] 郭进，统计语言模型及汉语音字转换的一些新结果，《中文信息学报》，Vol 7，No. 1，1993