<< 拼音－汉字变换输入法 >> 的性能测试

何 厚 存

上海交通大学 计算机科学与工程系

A Performance Test For < PINYIN - HANZI Conversion Input Method >

HE HOU CUN

Dept. of CS&E, Shanghai Jiao Tong University

拼音－汉字转换输入法是当前各类智能化汉字输入法的基础。这种输入法充分利用汉字、汉语的语音属性、构词规则、语法规则甚至语义属性来解决拼音－汉字转换中的音节切分、同音字（词）的分化以及词组的切分等一系列的问题。通过对拼音－汉字转换输入法的测试方法的研究，有利于探索各类智能化输入方法的测试原理、性能指标体系和具体的测试技术。

本文在分析拼音汉字转换技术的工作原理的基础上，设计了一套比较完整的测试方案，并对测试结果进行了分析，提出进一步提高转换系统的性能指标的途径。

现概述如下:

一、主要性能指标:

1、词汇覆盖率：词汇覆盖率是指在现代汉语词频表所收的词汇范围内，拼音－汉字转换系统能正确实现转换的词语的实用频率之和。

2、含多音字的汉语词语转换正确率：对含多音字的词语的转换正确率的测试将反映出转换系统对多音字的处理能力。

3、同音词转换正确率：同音词问题是影响拼音－汉字转换系统正确率的重要因素。

4、含零声母字的词语转换正确率：含零声母字的词语的转换正确率是检验转换系统中音节切分性能的重要指标之一。

5、综合转换正确率：定义为测试文本中实现正确转换的汉字数与文本总汉字数之比。

二、测试方法:

本方案生成两种类型的测试文本，一种是词语文本，主要适合于指标1、2、4的测试。另一种为连续文本，用于指标3 和5 的测试。这两种文本均为汉语拼音形式，由本测试系统的汉字－拼音转换模块自动产生。拼音－汉字转换系统以文件形式读入测试文本，并将转换结果存放到输出文件。测试系统的比较模块将输出文件与测试文本的原始汉字文本作比较，得到测试结果。

测试所用的词语来源于《现代汉语常用５０００词词表》及《现代汉语词频表》，文本为新闻、文艺类短文，共计48篇，约2万字。

三、测试结果分析

转换系统的正确性依赖于切分、解析和变换等各个环节的正确性。从本次测试结果来看，单词类的转换准确率在86－90％，而连续文本的转换准确率仅为60％左右。可见拼音串的正确切分对正确转换是至关重要的。此外，主要的错误原因尚有：词库较陈旧，缺少新词及常用短语、句法分析规则及同音词选择规则不完善以及对多音字的判断有误等。