采用拼音表记的语法分析型复数文节的中文句子输入系统

河野 勝也　小滝 房枝　隈井 裕之　松田 純一　([株]日立制作所 中央研究所)

摘要:本文介绍了拼音句子变换的中文输入方法。其处理过程为输入拼音流、系统利用词典查询、单词切分、语法分析、学习功能等技术实现了拼音汉字整句变换。

关键词:拼音句子输入　语法分析　共起分析　单词切分

1. 前言

历来,中文的输入方式,经常被提到的是四角号码法、五笔字形法等把汉字分解号码化以后再输入的方法[1]。

这样的输入方法,因为需要进行大量地学习,所以适合于职业操作员一面看原稿一面输入,而一面思考一面输入的时候,会对思考形成妨碍。

考虑到上面提到的原因,为了不妨碍用户的思考,认为以读音输入的方法是最适合的,所以开发了采用中文的英文符号表记(拼音)的复数文节的中文句子输入系统。本系统可在个人计算机上运行。

2. 系统概要

拼音和发音表记接近,与历来的使用号码和笔画数、部首的输入方法相比,即使初学者也容易高效率地进行输入。

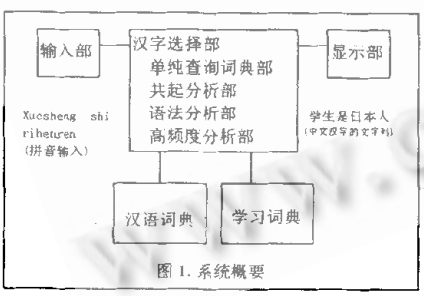

图 1. 系统概要

图 1　系统概述

象图 1 所表示的一样,本系统是把中文以句子为单位以读音输入,再变换成中文汉字的文字列的系统。汉语词典大约拥有 5 万 5 千个单词,并且拥有语法分析所需的词性[2]信息。

根据读音输入再变换成汉字的方法和日语的假名汉字变换一样,它的技术课题是如何从同音词当中选择出正确的汉字来。

为了解决这个问题,以下两种方法是行之有效的。①应用语法知识使汉字特定化,②利用汉字的使用频度情报。

3. 汉字选择方法

下面,就高准确度地选择同音词的汉字选择部的处理过程进行说明。

从键盘上根据标准拼音表记把中文的读音以句子为单位输入后,采用以下的过程进行处理。

①单纯查询词典部:进行单纯地词典查阅,如果没有同音词的话就决定这个汉字。

②共起分析部:一起出现的可能性大的读音出现了的话就决定这个汉字。

③语法分析部:根据单词及前后相邻的单词的词性的连接特性,依照语法规则进行定义,当语法规则一致的时候,就决定这个词性或者汉字。

④高频度分析部:当上述的方法不能决定的时候,就把同音词中使用频度最高的汉字作为所选汉字。

变换经过的例子如图 2 所示。

中文:学生买一本书。

拼音输入:	XUESHENG	MAI	YI	BEN	SHU.
①单纯词典查询:	学生	MAI	YI	BEN	SHU。
②共起分析:	学生	MAI	YI	本	书。
③语法分析:	学生	MAI	一	本	书。
④词频分析:	学生	买	一	本	书。

图 2. 变换过程的示范

图 2　变换过程的示范

①XUESHENG 没有同音词只要查阅一下词典就可以决定是"学生"。

②BEN 有 9 个，SHU 有 46 个同音词，但是 BEN SHU 同时出现时，用共起分析部分可以判定量词和名的组合"本" "书"一起出现的可能性大就决定这两个汉字。

③YI 有 110 个同音词，但是因为量词"本"已经决定，根据量词的前面应该是数词这个语法规则，所以决定是"一"。

④上述的方法不能决定的 MAI 有 9 个同音词，但是参阅词典里保存的频度信息，把使用频度最高的汉字做为所选汉字。

4. 语法规则

语法分析使用的语法规则是以中文的语法知识为基础的，设定出了以下的规则。

(1) 文头、文尾的分析规则。用文头、文尾的信息来分析。规则的例子如图 3 的 R1 到 R3 所示。

(2) 决定汉字的两旁分析规则。注视紧接前后的单词的词性和特定的读音等来决定汉字。举例来说，助动词"DE"的同音汉字"的"、"得"、"地"的分类选择规则如图 3 的 R4 到 R7 所示。

R1:文章的开头是"，"号的时候，文章的开头是感叹词或者是助词。
R2:文章的开头不是"，"号的时候，文章的开头的词性里有代词的话就是代词，不是代词有介词的话就是介词。
R3:文章的最后的词性里有助词的话就是助词。

R4:读音是 DE 前面是名词或者是代词的时候就决定是"的"。
R5:读音是 DE 后面是名词或者是"。"的时候就决定是"的"。
R6:读音是 DE 后面是动词前面是形容词的时候就决定是"地"。
R7:读音是 DE 后面是动词前面不是形容词的时候就决定是"得"。

R8:前面如果是量词的话，后面是名词。
R9:后面如果是量词的话，前面是数词。
R10:后面如果是动词的话，前面是副词、形容词。
R11:前面如果是助动词，后面是动词、形容词。
R12:前面如果是介词的话，后面是名词、代词、数词。

R13:同样的读音重复的时候，或是动词、或是形容词、或是名词。
R14:"一"或"了"夹在中间，并且读音相同的时候，是动词或者是形容词。

图 3 语法规则

图 3

(3) 决定词性的两旁分析规则。注视紧接前后的单词的词性和特定的读音，或者已经决定了的汉字等来决定词性。如果决定了的词性中有多个汉字时，最终用高频度分析法来决定汉字。如图 3 的 R8 到 R12 所示。

(4) 根据读音特性决定词性的两旁分析规则。在中文当中，重复使用同一个汉字以增加文章意思的方法经常被使用。

如果前后的汉字决定了，被重复使用的汉字的词性也就可以决定。如图 3 的 R13 到 R14 所示。

5. 连续输入拼音的对策

一般输入拼音时，应以单词为单位进行输入。具体方法是于单词与单词之间输入空格。在本系统输入复数的单词短语时，在拼音间输入空格作为单词与单词之间的分隔符。

但是，从日语的假名汉字变换输入方法，可联想到很多用户也希望能将复数文节的中文拼音连串地输入计算机，所以本系统实现了对连串输入的拼音也能变换成复数文节的中文的句子输入方法。

将连串输入的拼音变换成汉字的时候，单词有多种切分的可能性，即多种切分方法能产生多种不同的意思，与用户直接用空格键进行切分的方式相比，变换性较低。为了得到正确的变换结果，必须选择正确的切分位置。

Woshiribenren

卧室 /日本人
我/是 /日本/人
握/十

图 4 不同的切分法引起的歧义

图 4 不同的切分法引起的歧义

连续输入拼音时，因不同的单词的切分而产生不同意思的例子如图 4。

日语的假名汉字变换可根据最长一致法，最小文节数法等算法，一般都能把连续输入的假名正确切分成多个单词。这是基于文节越长构成日语单词的可能性越大的搜索规则。

但是，中文切分单词时把较长的拼音进行切分不如把较短的单词进行切分来的正确性高。这是因为中文单词特别是基本单词中，单文字的单词比较多。

再者，日文存在附属语，并且有音读与训读，单词的切分比较容易判断。而中文中没有特别的表示间隔的拼

音，而且，中文比较简短 。

为了消除单词切分而产生的歧义，可采用所有可能产生的组合都进行分析的方法，但这样可能使变换时间过长，缺乏实用性。

本系统对拼音的性质进行研究的结果，关于连续输入的拼音的高速单语切分采取了以下方法。

把中文词典中的单词进行常用单词与非常用单词的分类，然后，追加如下的以拼音的开头来决定单词的切分位置的单词切分处理方法如图 5。

图 5 进行单词切分时位置的决定方法

图 5 进行单词切分时位置的决定方法

①将输入的拼音，检索中文词典，生成所有的单词切分的组合。

②对于单词的切分组合，使用共起部中的规则，确定单词的切分。

③如果②中无适用的规则，并且存在常用词与非常用词时，以常用语的长拼音优先，如果只存在常用词或非常用词时，以长拼音的优先来决定切分位置。

④根据第 3 节［汉字选择方法］来选择同音词。

6. 评价

开发了使用以上的分析处理方法的中文输入系统。表 1 为连续输入拼音时，以各领域 19 篇文章(16826 个文节)为对象的变换率评价结果。变换正确的文节数和总文节数之比称为正确变换率。全部文章的平均正确变换率为 75.5% 。而未按文节的单词切分处理方法进行变换时，正确变换率为 58%。

再者，做为以单词为单位先切分再进行拼音输入时的变换率为：用文法规则时的正确变换率为 92%，不用时为 87%。

通过以上结果可判定本系统的文法规则与单词切分方式的有效性。本系统可预期达到具有较高实用性的变换率。

表1 评价结果

分类	编号	文节数	正确变换数	正确变换率
语法书	1	2888	2262	78.3%
	2	5132	4185	81.5%
文学·思想	1	828	645	77.9%
	2	1264	908	71.8%
电脑方面	1	678	540	79.6%
政治·经济	1	920	651	70.8%
	2	349	261	74.8%
	3	838	599	71.5%
理工学	1	355	254	71.5%
教育与运动	1	73	46	63.0%
	2	465	338	72.7%
	3	277	208	75.1%
	4	648	464	71.6%
其他	1	345	242	70.1%
	2	493	216	43.8%
	3	731	473	64.7%
	4	247	188	76.1%
	5	125	82	65.6%
	6	170	136	80.0%
合计		16826	12698	75.5%

7. 今后的课题

通过评价结果的分析，并且验证规则的适用条件的有效性，进一步提高变换精度是今后的课题。

参考文献

[1] 陈他：中国语の汉字入力の一方法；情学会第 35 回全国大会

[2] 香坂：现代中国语辞典；光生馆

[3] 三野：中国语文法の基础；三修社

[4] 朱、杉村他訳：文法讲义；白帝社

[5] 王他、林訳：中国语动词活用辞典；方店

[6] 高桥他：中国语虚词类义语用例辞典；白帝社

(来稿时间：1998 年 10 月)