独立于输入法的汉字输入模型及利用中文自然语言知识后处理汉字输入[1]

An input-method independent model for Chinese input and exploiting the knowledge of Chinese natural language for postprocessing of Chinese input

庞民治 (Pong Man-Chi)
香港 香港科技大学 计算机科学系
Department of Computer Science
The Hong Kong University of Science & Technology
Clear Water Bay, Kowloon, Hong Kong

张永光 (Zhang Yongguang)
美国 普度大学 计算机科学系
Department of Computer Sciences
Purdue University
West Lafayette, IN 47907, U.S.A.

摘要

本文提出一个汉字输入的模型，将汉字输入的过程分为三个阶段：预处理、输入转换、后处理。这模型的一个特点是“独立于输入法”，即它可适用于任何输入法。任何一个输入法变成是模型中的一个数据文件。本文后部重点讨论后处理阶段，并介绍两种后处理方案（词汇法及语句模板法），它们利用上下文关系信息及中文自然语言的知识，减少从输入转换阶段得来的汉字的歧义性，从而减少或消除用户选字的需要。词汇法已被实现，语句模板法将开始实现。

Summary

This paper describes a model for Chinese input, which abstracts the input process into three stages: preprocessing, input conversion, and postprocessing. A major characteristic of the model is input-method independence. The model suits any input method, which becomes a data file in the model. The latter part of the paper discusses two ways to do postprocessing: glossory method and sentence template method. They use the context information of the input characters and the knowledge of Chinese natural language to reduce the ambiguity of the Chinese characters output from the input conversion stage. Hence the need for the user to choose from candidate characters is reduced or eliminated. The glossary method has been implemented and we are going to implement the sentence template method.

1．汉字输入编码方案的模型

图 1 表示汉字输入编码方案和用户与计算机在中文输入中的关系。图左上方表示编码方案。编码方案（IC）将汉字（HZ）分成汉字组成部份（HC）。这可用以下数学关系表示

$$IC : HZ \times HC^+$$

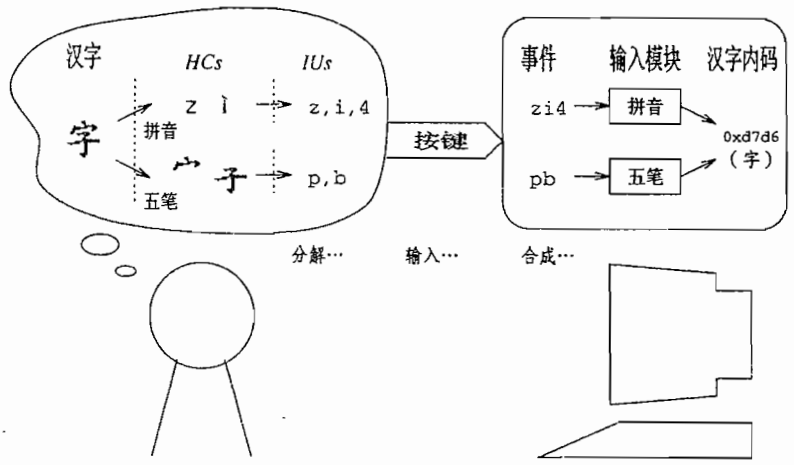
其中X^+代表X的元素出现一或多次，简称为X串。集的元素用小写英文字母串来表示。即

$$IC(hz) = \{hc_1^+, hc_2^+, \ldots\}$$

[1]本研究项目得到香港科技大学“信和软件研究中心”资助。

（这里用数学关系（而不是数学函数）来表示IC，是因为一个汉字可以分解为多个不同的HC^+。下面提到的数学关系，只有说明是一一对应的才是数学函数。）

图 1： 用户与计算机在中文输入中的关系



“编码”过程（KM）是将HC分配于键盘各键作为输入单元（IU）。这可用以下数学关系表示

$$KM : HC \times IU$$

IC与KM的总和（$KM \circ IC$）可以简化称为输入法（IM），就是

$$IM : HZ \times IU^+$$

即一个汉字可以分解为一输入单元串。

每当用户击键，系统的输入模块（以下简称“系统”）接收到相应的输入事件（IE）。由于IE与IU是一一对应的，我们将只用IU代表输入单元或输入事件。

系统包含实现当前输入法的输入模块（IM'）来处理输入事件串，将它们转为汉字内码（$HCode$）。这可用以下数学关系表示

$$IM' : IU^+ \times HCode$$

由于通常一输入事件串可映射到多个汉字内码，所以IM'是关系而不是函数。

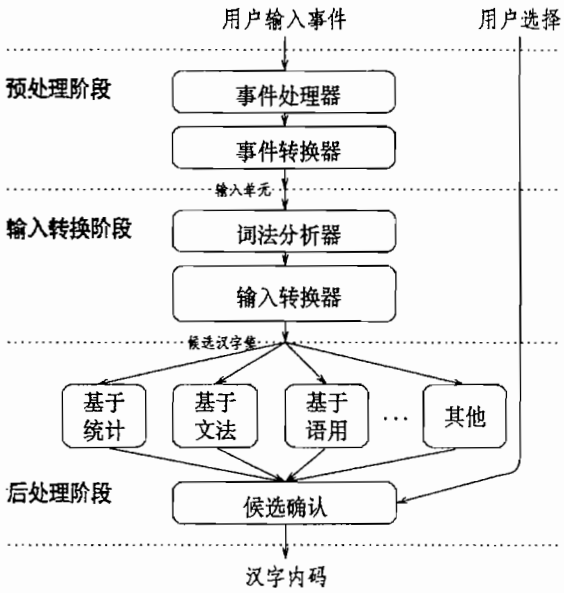
由此可见，汉字输入的过程在用户和系统方面来看是相反的关系，用户是实现关系IM，而系统是实现相反的关系IM'。

2．“独立于输入法”的输入模型

现行不少系统的实现是将整个输入过程一块处理。对不同输入法，实现上很多时需要有不同的编程。因此，从软件工程角度来看，移植和扩展这些系统都比较困难。

基于以上的分析，我们提出实现“独立于输入法”的系统模型，如图 2 所示。依此模型，汉字输入分为三阶段：预处理、输入转换、后处理。每阶段之间的接口是清础定义的。改变输入法只需更换输入转换阶段所用的有关数据文件。因为只有预处理阶段是与机器系统相关的，实现输入转换和后处理阶段的模块可不用更改而在不同机器上使用。还有，后处理阶段中可以使用不同的模块，对从输入转换阶段得来的汉字集进行歧义性处理，以提高中文输入的效率。

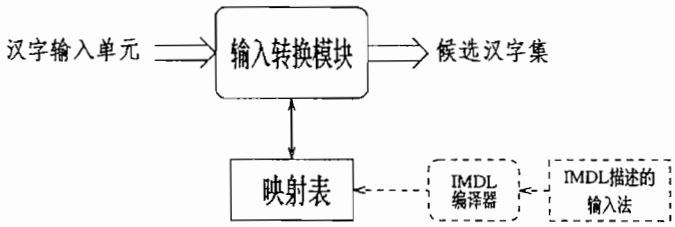
图 2： 汉字输入的处理阶段



我们提出以上模型，是因为我们观察到所有输入法都有一定共同之处，就是将字分解成汉字组成部份再编码。因此，我们设计一种机制，将各输入法的规律抽象出来，加以处理。

图 3 表示这种机制。每一种输入法用一种“输入法形式描述语言”（IMDL）描述，基本上这是描述一个IM'关系的映射表。汉字输入变换模块将从预处理阶段送出的一输入单元串，经查询代表IM'的映射表，转为“候选汉字集”。因为IM'歧异性的特性，通常候选汉字集有多过一个候选汉字。

图 3： 以输入法为导向的输入变换机制



IM'映射表的形式描述基本上的格式如下：

输入单元 1　　候选汉字集 1

输入单元 2　　候选汉字集 2

例如，不带声调拼音输入法的形式描述如下：

a　　{啊，阿，啊，……}

ai　　{埃，挨，哎，……}

…　　……

zhong　　{中，鍾，终，……}

为提高查询这映射表的效率，我们可开发一个 IMDL 编译器（见图 3 右下角），将映射表转变为便于查询的数据结构，供图 3 中的输入转换模块使用。

3．直接将词编码输入的缺点

上文提到字输入法的实现模块可以抽象为以下关系

$$IM' : IU^+ \times HCode$$

通常该模块的输出是一个有多过一个元素的集（称为候选汉字集（$hcode$））。要从中选出目标汉字，最简单的方法就是由系统显示该集，然后让用户去选。

为提高输入速度及减少用户的疲劳度，以词为输入单位是常用的手段。词输入法（PIM'）的实现模块是将一输入单元串转换为一候选汉字串。抽象为

$$PIM' : IU^+ \times HCode^+$$

用户要记得所想输入的词相应的输入单元串，才能输入该词。常用3000词可覆盖一般语料的86%[1]。但用户要记得3000条相应的输入单元串是非常困难的。

比较理想是用户只要懂得某种字输入法（最普遍的是拼音法），就无需再记忆词输入法的编码方案，系统通过后处理就可减省或甚至免掉用户选择目标汉字串的工作量。

4．后处理

依本文提出的模型，在输入转换阶段，已将输入单元串转换为候选汉字集。因此以下讨论的后处理方法是以汉字为基础的，而不是以处理输入单元为基础的。

理论上，当用户输入n串输入单元串，后处理模块会得到n个候选汉字集。假设每集的汉字数目为$m_1, m_2, \ldots, m_n$ 则有$m_1 \times m_2 \times \ldots \times m_n$ 条候选汉字串。后处理就是要减少这数目。

在实现中，系统反应时间是很重要的，否则用户就会不耐烦。因此，一些需时良久的自然语言理解技术，例如，动态规划[2]，就不适合于后处理中应用。下面将讨论两种后处理方法。

4．1．词汇后处理法（词汇法）

词汇法是利用一个只含词组的词汇（不含代表词组的输入单元串）。首先系统用当前使用的字输入法，将用户输入的n串输入单元串（例如以'/'分隔），同时映射到n个候选汉字集，然后将每个候选汉字集内的汉字匹配词汇内n字词组的相应的每个字。

抽象来说，词汇法可用关系$GM(g, im, [iu^+]^n)$ 表示。其中g是一个词汇，im 是某一当前输入法，$[iu^+]^n$表示一串由n个IU^+的元素组成的串。设 $iu_1^+, \ldots, iu_n^+$ 是$[iu^+]^n$的元素，$im(iu_j^+)$分别为在输入法im下输入串iu_j^+ 所产生的候选汉字集，则

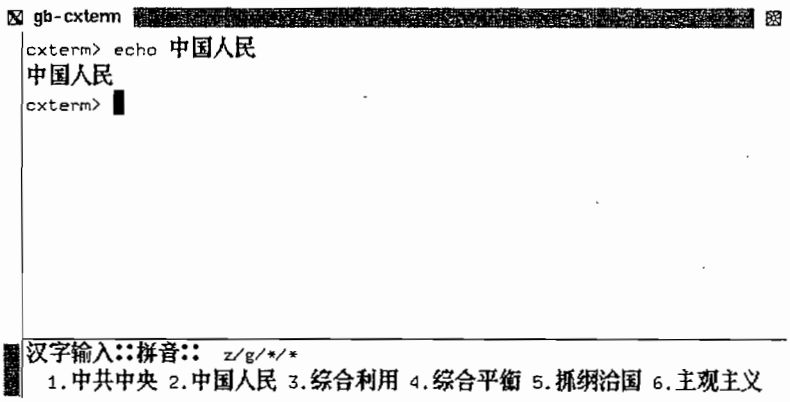
$$GM(g, im, [iu^+]^n) = (im(iu_1^+) \times \cdots \times im(iu_n^+)) \cap g$$

如果经过词汇法处理后仍有多个选择，则用户仍需从中选择所需目标词组，但词汇可以是按词频排列，系统每匹配一定数量的词组（总字长度 $\geq$ 屏幕阔度），就可显示让用户选择。如用户按特殊键（如'>'）表示仍未有所需目标词组，要求下一屏幕的选择，系统就再选择下一批词组。因为常用词组先出现，通常用户很快就能找到所需词组。只有在所需词组是不常用或不存于词汇时，才会匹配整个词汇。作为最后手段，用户仍可用所选输入法逐字输入。

在X窗口系统中开发的汉字仿真终端程序 cxterm [3] 已实现了本文提出的输入模型及词汇法后处理模块[4]。图 4 显示在 cxterm 中使用的词汇法。

Cxterm 中的词汇法模块还容许用户输入不完整的单元串，即用万能匹配字符'?'和'*'代表部份输入。例如，在拼音输入法中，用户可能不清楚某汉字的读音，则输入 "zh?ng" 将匹配相应于 "zhang", "zheng", "zhong" 的汉字。甚至'*'也可作为输入单元，它将匹配任何字。输入 "zh*/*/ren/min" 将会匹配 "中国人民" 等词组。再例如，用五笔输入法，输入 "k/w/w/n/a/t/l" 将会匹配 "中华人民共和国" 等词组。

图 4： cxterm 中使用的词汇法



总的来说，如果有一个足够大及富代表性的词汇，用户可以利用词汇法比较快地输入所需词组。与词输入法相比，词汇法可配合任何一个字输入法共用，不需用户记忆大量代表词组的输入单元串。

4．2．语句“模板”后处理法（模板法）

利用中文自然语言的知识和汉字在一词或一句上下文相关的信息，可将输入的歧义性减少，直至得出唯一的选择，那时用户就无需作出选择，因而可提高输入速度。（注意，提高输入速度是与用户友好程度有关，而非只是简单地优化击键次数。）

我们可建立一个有关中文语句的“匹配规则”库，称为模板库。以下是一些模板的例子（用BNF语法定义式表示）：

（1）<数词>“只”［［<副词>］<形容词>“的”］<动物>

（2）<数词>“支”［［<副词>］<形容词>“的”］<长型物体>

用户可连续输入n串输入单元串，输入转换阶段将它们变为n个候选汉字集。然后系统采用模板法，首先利用词频，可以较肯定地消除由一些候选汉字集组合的词和字的歧义性。再以此等字或它们的组合词中最高频的作为“关键词”去搜索模板。

例如，用户输入拼音“yi1 zhi1 bi3”，候选汉字集会包括

{一，依，医，．．．}，

{之，知，只，指，支，．．．}，

{比，笔，彼，．．．}。

（每集内的字依字频顺序排列。）

再依词频得出关键词“一只”和“一支”等。再用这些关键词去搜索模板库中含有此等关键词的模板，然后匹配模板中其他的项。

匹配上面的模板（2）得出“一支笔”和“一支匕”等。依词频确定此时的输入为“一支笔”。

当用户再继续输入时（即有更多信息时），已经确定的汉字是有可能被改变的。

例如，“只”和“支”是同音字。当拼音输入“yi1 zhi1 ke3 ai4 de1”，系统依使用频率显示“一只可爱的”。但当跟著的输入为“bi3”时，系统这时利用模板（2）会将显示变为“一支可爱的笔”。

因为在“一只……动物”和“一支……笔”的例中，量词“只”、“支”与名词“动物”、“笔”之间可以被一定数量的其他汉字分隔，单纯利用词与词间匹配（如 Markov 链模型），不一定能分辨两

种情况。利用模板，是较容易应付被分开但是有关连的词。

有时，用最高词频初选的关键词搜索得出的模板中的其他各项并不能匹配在其他候选汉字集的字。这时，系统可采用其他次高词频的关键词或重新组合候选汉字集中的字，从而得出其他关键词以搜索其他模板。最后还可以依词频和字频来解决不能分析清楚的候选汉字集的字。如果有足够代表性的模板库，能成功匹配的机会是很高的。

台湾的“国音智慧型输入系统”[5] 亦有采用类似方案。它将匹配规则称为“样板”（template）。它存有对汉语常用句子句型分析得来的模板约三万条，在对一般白话文辨别同音字的测试中，据称平均正确率达98%。并且系统响应时间能追上用户击键速度。因此，匹配模板的实时速度是可以接受的。该系统只支持用注音符号或汉语拼音符号输入。

我们现正开始建立模板库及参考一些比较新的语法理论来建立模板，希望可以减少模板的数目。我们的目标是将模板法加进 cxterm 作为其中一个后处理模块。

5．结语

本文提出一个“独立于输入法”的汉字输入模型，将输入过程分为三个阶段：预处理、输入转换、后处理。预处理阶段对付依赖于系统特性的用户输入单元串；其余阶段独立于系统特性及输入法。输入转换阶段将输入单元串变为候选汉字集。当前用户使用的输入法只作为一个数据文件被此阶段使用。转换输入法只是更换相关的数据文件，而不改变此阶段所用的算法。后处理阶段对候选汉字集同时进行处理，以减少甚至消除输入的歧义性。本文并讨论了两种后处理方法，分别利用词汇和语句模板库。词汇法已被实现及应用于 cxterm 汉字仿真终端程序[4]。我们现在正建立模板库，亦准备加进 cxterm 中，以提高中文输入的效率。

参考文献

[1] 刘英林，宋绍周，“汉语常用词的统计与分级”，《中国语文》，1992，第 3 期，174–181页。

[2] Hung-yan Gu, Chiu-yu Tseng, and Lin-shan Lee, “Markov modeling of Mandarin Chinese for decoding the phonetic sequence into Chinese characters,” *Computer Speech and Language*, Vol 5, No. 4, 1991, pp. 363-377.

[3] Man-Chi Pong and Yongguang Zhang, “cxterm: a Chinese language terminal emulator for the X window system,” *Software - Practice and Experience*, Vol. 22, No. 10, Oct. 1992, pp. 809-826.

[4] Man-Chi Pong, Yongguang Zhang, and Chung-Kei Wong, “Extension of Chinese phrase input using a glossary of phrases,” *Proc. 3rd Int'l Conf. Chinese Information Processing*, Beijing, Oct. 1992, pp. 201-206.

[5] 许闻廉，陈克健，“国音智慧型输入系统”，未发表文章，台湾中央研究院资讯科学研究所，台北，1993。