

利用拼音数字码和笔画数字码的汉字智能输入

夏莹 马少平 姜哲 朱小燕 金奕江

智能技术与系统国家实验室, 清华大学计算机系, 北京 100084

张金岭

葫芦岛亚奥计算机培训中心

摘 要: 0-9 数字键盘的汉字智能输入是一种新的输入方法. 汉字的数字编码易学易用, 并能够自动地转换数字输入码为汉字. 在手持计算机、掌上型电脑或只有十个键的小型电子产品中, 汉字输入基本上不需要人工选择候选字或词. 利用汉语的上下文关系, 数字键盘智能输入系统不仅可以完成拼音—汉字转换, 而且可以完成形码(笔画数字码)—汉字转换. 因汉字的数字编码相对来说重码率较高, 人工智能过程在汉字候选的自动选择中起了重要作用. 这篇文章介绍数字键盘的音码和形码到汉字的转换系统, 该系统适合于小型电子产品, 要求的存储空间较小, 转换的正确率对于社会科学文章达到 94%。

关键词: 汉字输入 数字码 智能输入 Markov 模型

Intelligent Inputting Chinese Characters Using Pinyin Digital Code and Stroke Digital code

Xia Ying, Ma Shaoping, Jiang Zhe, Zhu Xiaoyan, Jin Yijiang

State Key Laboratory of Intelligent Technology and System

Dept. of Computer Science, Tsinghua University, Beijing (100084), P. R. China.

Zhang Jinling

Huludao Yaao Computer Train Center

Abstract: Intelligent inputting Chinese characters using a miniature digital keyboard ("0"~"9") is new input method. The digital codes of Chinese characters are very easy to learn and to use. They can be transformed into Chinese characters automatically. Using Palm-sized PC, Handheld PC(H/PC) or miniature electric information products in which there are ten digital keys ("0"~"9"), Chinese characters can be inputted nearly without manually selection of the candidate characters and words. By means of the contextual relationship in Chinese language, the digital keyboard intelligent inputting system not only can transfer the voice codes —pinyin into Chinese characters, but also can transfer the shape code —stroke digital codes into Chinese characters. Because the "repeated codes" rate of the digital codes of Chinese characters is relatively higher, the artificial intelligent processing played an important role in automatic selection of Chinese characters. This paper introduces the automatic transformation of the pinyin digital codes and the stroke digital codes to Chinese characters. The system is suited to Palm-sized PC, Handheld PC or miniature electric information products. The correct transfer rate can reach 94% for articles on social science.

Keywords: Chinese character inputting, Digital code of Chinese character, Intelligent Inputting, Markov model

1. 引言

随着信息技术的飞速发展，计算机及其相关电子信息产品在向小型化、数字化方向发展，计算、上网与通信的结合。汉字输入不再局限于在台式计算机上，全中文媒体电话、双向寻呼机、手持计算机（H/PC）、小型信息终端等也都需要有汉字输入，即仅有 10 个数字键的小型设备也将需要输入或处理汉字。

尽管联机手写汉字识别、语音识别技术取得了很大的进步，但这两种输入方式总会有些人因识别率低而很难输入汉字，因此在小型化设备中用数字键盘输入是必不可少的。从产品的发展及用户群体的变化两个方面来看，现行的键盘输入汉字方式显然不能满足现代市场的需要，必须跟上时代的步伐，要求使用汉字数字码。汉字输入一定要简单易学、实用快捷，“上手能用”。

对于在通讯产品中使用汉字数字码，可以分为三个层次：字输入、词输入、连续句输入（即智能输入）。选用哪一种层次要根据小型通讯产品中 CPU 速度和存储量而定。

汉字智能输入应用计算机人工智能技术，使操作更加简便，几乎不需要选字，尤其对于由 0-9 数字组成的拼音编码（音码）和笔画数字码（形码），重码率相对地来说要比较高一些，智能化更加需要，依靠汉语上下文关系计算机自动地进行同码字的选择，转换为汉字，输入者基本上不用选字。

利用全拼音数字码和笔画数字码在 10 个数字键上连续输入是一项新技术，它面向最普通的用户，做到不需培训，“上手能用”，并能够自动地转换数字输入码为汉字。我们在数字音码及形码智能汉字输入方面做了有益的尝试，采用计算机人工智能技术解决了重码的选字问题，其中音码采用现行的汉语拼音，在标有拼音字母的数字键上输入，已会拼音的用户不需要再学习就可以使用。形码采用数字为代码的笔画数字码，该编码符合《现代汉语通用字笔顺规范》，并有容错能力。因此该形码易学易用，也不需要培训。

小型化的通讯产品对智能输入技术提出更高的要求，因小型化的产品存储空间小，要求汉语上下文关系库不能太大。因 CPU 速度低对智能算法提出更高要求。这方面已经达到手持通讯产品能够接受的实用水平。

2. 拼音数字码智能输入

小型电子通讯产品或掌上型电脑一般只有 10 个数字键（或只用 8 个键），在这类数字键盘上也可使用智能全拼音输入，下面介绍在数字键上利用汉语拼音输入汉字的方法。

1. 将汉语拼音字母按次序设定在 10 个数字键上，每键二至三个：

0	1	2	3	4	5	6	7	8	9
ABC	DEF	GHI	JK	LM	NP	OQR	STU	VWX	YZ

2. 安排在英文标准键盘副键盘的数字键上，如下图。（若键位不同可以改变该图。）

7 STU	8 VWX	9 YZ	+
4 LM	5 NP	6 OQR	
1 DEF	2 GHI	3 JK	Enter
0 ABC		. Del	

3. 规则：按拼音击相应字母数字键，声母如是黑体字母重复该键一次

4. 输入实例

李 LI 击 42 米 MI 击 442（声母 M 为表中黑体字母重复一次）

张 Zhang 击 92052 北京 Beijing 击 00123252（B 重复一次）

5. 重码情况

GB—2312 汉字不含音调时有拼音 411 组。按以上形式分配到 10 个数字键，能编出 406 组数字码，增加重码情况如下：

002	bi (52 字)	cai (11 字)	重 63 字
006	bo (21 字)	cao (9 字)	重 30 字
02	ai (22 字)	ci (17 字)	重 39 字
772	sui (19 字)	ti (24 字)	重 43 字
7705	tan (26 字)	suan (4 字)	重 30 字

以上五组重码为拼音数字码在拼音基础上增加的重码。也就是说，汉字拼音总音节有 411 组，用数字表示后，增加了 5 组重码，有 406 组拼音数字码。在新增重码中最大重码数为 63 字，最小为 30 字。

由于该方案拼音字母是按次序排列在数字键上，使用起来非常方便，它解决了输入汉字先要记住无序的 26 个字母位置及双手操作不协调的问题。不通过任何培训就能让初学者和小学生在数字键上输入汉字。同时也满足了对输入汉字速度要求不高、工作量不大的用户需要。

3. 笔画数字码智能输入

音码有“音”的长处也有它的弱点，形码正补充了“音”的不足。我们的笔画数字码——形码编码规则非常简单，实现了“不需要培训、人人上手能用”的目标。符合国家语委颁布的《现代通用字笔顺规范》，并有容错能力。

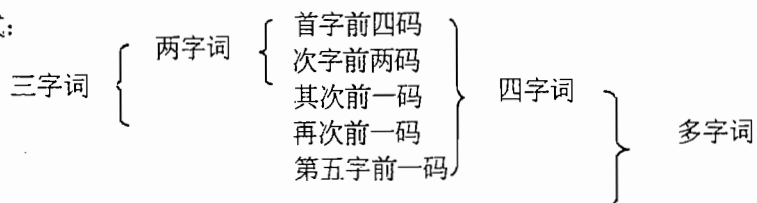
笔画数字码是以汉字的五种基本笔画为基础，利用笔画之间有无交叉笔的特性，将五种基本笔画扩展为 10 个数字而形成的数字编码法，码长最大为 5。

1. 笔画数字码编码方法

笔画代码规则:

笔画	横	竖	撇	点	折
	—		丿	丶	乙
无交叉	1	2	3	4	5
有交叉	6	7	8	9	0

- 笔画归并: 提归一, 竖钩归|, 捺归丿, 各种折笔均归乙
- 例字: 工-121 土-671 父-3489 田-25671
- 五笔以上汉字, 取前四笔和后一笔编码: 价-32342 设-45359 秃-36735
- 词组取码方式:



2. 输入实例

清 华 大 学

44161 32807 684 44346

笔画数字码是按汉字的笔顺次序取码, 这与汉字的书写习惯保持了一致。采用容易判断的笔画有无交叉的特征, 又大大地拓展了编码范围。加上计算机人工智能技术的运用, 使得汉字平均笔画数虽在 10 画以上, 每个汉字最多取五个笔画就足让您高准确率地连续输入汉字了。

4. 小型化数字键盘汉字输入的人工智能处理技术

数字音码及形码智能输入系统的智能处理, 主要体现在自动处理重码技术, 不是靠人工选择同码的候选字, 而是用马尔柯夫模型作为输入码到汉字的转换模型。

$S = \langle S_1 S_2 \dots S_n \rangle$ 为一句输入串, S_i 为一个词或者字的编码;

$T = \langle T_1 T_2 \dots T_n \rangle$ 为一可能的汉字串, T_i 为输入码 S_i 的一个词或字 (即候选词或字);

$P(T|S)$ 表示当输入是 S 时, 输出为 T 的概率。

当 $P(O|S) = \text{MAX}\{P(T|S)\}$ 时, O 即为最佳结果。由 Bayes 公式得:

$$P(O|S) = \text{MAX}\{P(T|S)\} = \text{MAX}\{P(T) * P(S|T) / P(S)\}$$

在上式中, 由于输入 S 已定, 故 $P(S)$ 项不影响选择, 可以不加考虑。当 T 在候选集之内时, $P(S|T)$ 项可以用 1 来代替。因此

$$P(O|S) = \text{MAX}\{P(T)\} \quad (1)$$

我们把汉语语句看作是一个 Markov 源 (即某状态的发生概率仅与其以前的状态有关), 那么:

$$P(T) = P(T_1)P(T_2|T_1) \dots P(T_n|T_1 \dots T_{n-1}) \quad (2)$$

如果我们认为第*i*个字的出现仅与前面很少的*n-1*个字有关，则问题就会大大简化。这样的模型叫做*N*元语法。如果采用二元语法模型 (bi-gram)，即取*n=2*，也就是说，在确定第*i*个字时只考虑前面一个字的出现情况，则可得下式：

$$P(T_2) = P(T_1)P(T_2|T_1) \quad (3)$$

对于以上情况，其参数项 $P(T_2|T_1)$ 称为二元同现概率，它们可以从对话料文本的统计计算中获得。我们可以用最大似然估计法 (MLE) 来计算上述二元同现概率，其计算公式如下：

$$P(T_2|T_1) = N(T_1T_2) / N(T_1) \quad (4)$$

在上式中， $N(s)$ 是字符串*s*在语料库中出现的次数。

采用最大似然估计法 (MLE) 方法来计算模型的转移概率，在训练语料不足或参数空间庞大的情况下，会遇到数据稀疏 (Data Sparsness) 问题：即有许多合法的在未来的文本中要遇到的标记同现现象在统计语料中从未出现过，因而在遇到这种情况时，会出现零概率情况。对于合理地平滑处理数据稀疏的估值算法，目前有很多。一种较简单的解决此问题的方法是使用二元和单字频率的加权平均，该方法是Markov模型的数据平滑方法，其基本思想就是：若统计数据不充分确切地说不可信时，我们宁可回到*n-1*元组来计算。在实际应用中，是用它们的线形组合来实现的。在我们的模型中即使用了此方法来计算同现概率。同现概率反映中文文本中汉字的相邻关系。为了获得汉语词词和字字的二元同现概率我们对大规模语料进行了统计。并用动态规划法求最佳路径得到汉字串。

与台式计算机不同，对于小型化通讯设备或手持计算机来说，CPU速度低，存储量小，要求同现概率库不能太大，对于统计得到的二元同现概率，要进行筛选、整理、压缩，现在已经可以作到二元同现概率库小于 1.4MB，并根据用户的实际需要求最佳路径算法进行改进，已经达到手持通讯产品能够接受的实用水平。

5. 测试结果

用人民日报、中国青年报及参考消息等社会科学方面的文章进行测试，自动转换汉字的正确率如下：

智能输入自动转换汉字的正确率：

全拼音 (字母键)	94.2%
(数字键)	93.9%
笔画数字码 (数字键)	95.6%

例句 1: 工 作 时 间 不 能 踢 足 球
全拼音

数字键输入: 22652 976 722 3205 .007 5152 772 97 627
字母键输入: GONG ZUO SHI JIAN BU NENG TI ZU QIU
笔画数字码
字输入: 121 32311 25114 42521 1324 54255 25123 25124 16714
字词输入: 12132 251142 132454 25123 251216

例句 2: 实 现 经 济 发 展 和 社 会 进 步
全拼音

数字键输入: 722 8205 3252 32 110 9205 21 721 272 325 007
字母键输入: SHI XIAN JING JI FA ZHAN HE SHE HUI JIN BU
笔画数字码
字输入: 44544 16715 55151 44142 08094 51364 36731 45241 34114 66874 21213
字词输入: 445416 551544 080951 36731 452434 668721

6. 结论

从对数字码汉字输入分析不难看出它的优良特性是:

1. 拼音数字码和笔画数字码都实现了“不需要培训、人人上手能用”的目标。
2. 在手持通讯产品中使用 0-9 编码的汉字数字码, 包括音码和形码。可以分为三个层次: 字输入、词输入、连续句输入(即智能输入)。简单易学, 通讯用户都能方便地驾驭汉字输入。
3. 对于装有数字码智能输入的设备, 由于应用了中文上下文关系的智能技术, 输入更加方便, 连续输入基本上不需要选字。

参考文献:

- [1] 马少平, 夏莹, 朱小燕, 基于词词同现概率的拼音汉字自动转换方法, 电子计算机与外部设备, Vol. 21, No. 3, P16-19, 1997. 3
- [2] Xia Ying, Ma Shaoping, etc. The statistical co-occurrence probability of Chinese characters and its application. The 3rd National Joint Conference on Computing Linguistics, 1995, Shanghai
- [3] Xia Ying, Ma Shaoping, etc. Automatic post-processing of off-line handwritten Chinese Text recognition, ICCV'96. 1996. 6, Singapore
- [4] Jin Yijiang, Xia Ying, et., Using Contextual Information to Guide Chinese Text Recognition. ICCPOL'95, 1995, Haraii