中 文 信 息 学 报
第 17 卷 第 2 期 JOURNAL OF CHINESE INFORMATION PROCESSING Vol. 17 No. 2

文章编号:1003 - 0077(2003)02 - 0060 - 06

汉字键盘输入智能处理软件综述*

陈一凡[1],朱 亮[2]

(1. 北京信息工程学院,北京 100101;2. 广东青月亮科技开发有限公司,东莞 511700)

摘要:作为输入编码的后处理,各种类型输入软件智能化的共同目标是由软件来识别和选定上屏的重码字、词与缩短平均码长,并促使编码简单化和规范化。本文简要地论述了基于理解的智能输入、基于语用统计的智能输入、基于模板匹配的智能输入和基于上下文关联的智能输入等四种类型的汉字键盘输入智能处理软件的原理、优点和有待解决的问题,并列举了每种类型的典型作品。

关键词:计算机应用;中文信息处理;综述;自然语言理解;语用统计;模板匹配;上下文关联;后处理

中图分类号:TP391.1 **文献标识码**:A

A General Statement for the Intelligent Input Software of Chinese Characters

CHEN Yi-fan[1], ZHU Liang[2]

(1. Beijing Information Technology Institute, Bei jing;
2. QueenMoon science and technology development Ltd, Dong guan 51170, China)

Abstract: As a post processing for input code of Chinese Characters, the issues of common interest for all kinds of input software of Chinese Characters is distinguish and decide coincident code of the Chinese Character and word by the input software, deduce average code length, impel Chinese Characters code to simplify and standardize. This paper gives a general statement at the fundamentals, the merits and the demerits in four kinds of Intelligent Input software of Chinese Characters, based nature language understanding of Chinese language, based use frequency statistics of Chinese language, based matching mould of Chinese language, based context relating, and enumerate typical example of every one kinds of Intelligent Input software of Chinese characters.

Key words: computer application; Chinese information processing; overview; Nature language Understanding; Use frequency statistics of Chinese language; Matching mould of Chinese language; Context relating; Post processing.

一、前言

中国中文信息学会第一任理事长钱伟长教授于上世纪八十年代中期推出他亲自设计的“钱码”的同时坦言指出:“理想的输入方法还没有实现”。

二十多年的实践表明,单纯从汉字编码上下功夫,要得到一个易学、易用兼备的“理想的输入方法”着实艰难,就像人们很难将蒸汽机车的热效率大幅度提高,而不得不换成内燃机车、电气机车和磁悬浮列车。

还在八十年代,输入软件智能化先行者林才松先生设计了第一个智能拼音软件,尽管林氏

* 收稿日期:2002 - 08 - 28
作者简介:陈一凡(1935—),男,教授,主要从事中文信息处理领域的教学与研究.

的创举得到语委的支持和周有光教授的帮助,但 PC/ XT 和 CCDOS 却表示"爱莫能助",林先生只得无功而返。

二十年过去了,计算机软、硬件的发展速度以几何级数的形态增长,计算机系统资源的丰富为人们开发出实用的、算法各异的汉字输入智能软件创造了条件。

表 1 和表 2 反映 1984～2002 年计算机资源和汉字输入方式的开发与进步概况。

表 1　1984 年至 2002 年流行微型计算机的主要性能比较

年代	CPU 型号	时钟频率 (MHZ)	处理位数 (bit)	内存容量 (KByte)	硬盘容量 (MByte)
1984	Intel8088	4.7	8	512	10
1992	Intel486	66	16	1024	120
1994	Pentium	100	16	64×10^3	1×10^3
1996	Pentium Ⅱ	300	16	128×10^3	5×10^3
2001	Pentium Ⅲ	800	32	256×10^3	20×10^3
2002	Pentium Ⅳ	2300	32	512×10^3	100×10^3

表 2　1984 年至 2002 年系统环境与输入软件规模比较

年代	操作系统	字处理软件	字符集	输入方式	输入软件规模(KByte)
1984	CCDOS	Edline	6763	字输入	40～60
1992	CCDOS	Wardstar	6763	字、词输入	120～180
2000	Windows 9x	Word 2000	20902	智能处理字、词输入	103～55 ×103
2002	Windows XP	Word XP	27533	智能处理字、词、句输入	103～55 ×103

各种类型输入软件智能化的共同目标是由软件来识别和选定上屏的重码字、词与缩短平均码长,并促使编码简单化和规范化。

二、基于理解的智能输入软件

1. 原理

主要利用汉语语法知识来消化同音字、词,以及化解歧义分词。通常表述为计算机能够识别和处理的一系列固定搭配、公式和自定义规则。在学科分类中属于人工智能分支自然语言理解。这类软件是最早出现的也是最理想化的智能输入软件。

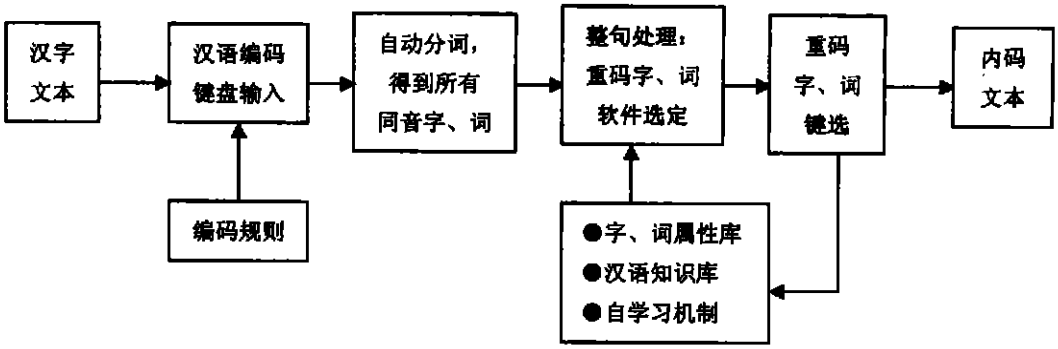

图 1　基于理解的智能输入软件结构示意图

根据自动分词得到同音字、词的候选集,查找知识库得到相关的规则,再经过归约推理,得出转换结果。利用句内编辑实时修正转换错误和批量学习可以使得系统知识不断完善和充

实,也就是自学习功能。

2. 典型作品

哈尔滨工业大学王晓龙等研制的拼音语句输入系统 InSun;

北京大学朱守涛研制的智能 ABC。

3. 优点与存在的问题

优点:⑴由于这一种自行构造的"语法体系",大体上能够包括最基本和较少歧义的汉语语法知识,因此系统的正确率比较稳定。⑵软件开销视知识库的规模可大可小,小型系统在 CPU 为 486 的机器上就能运行。

存在的问题:

⑴逐字连续拼音整句输入时,平均码长较长,采用简化拼音输入时键选率较高。⑵偏重于整句处理,当出现转换错误时,需要使用者回头去纠正,干扰了正常的思维。⑶当前,建立知识库时,汉语知识表达的困难;自动分词过程中切分歧义等因素对分词精度的影响;输入语句的语法不规范都使键选率的降低受到限制。最理想化的模型没有达到理想化的效果,因此人们不得不寻找不那么理想却比较实用的理论模型与方法。

三、基于语用统计的智能输入软件

1. 原理

主要利用语用统计的数据来消化同音字、词,以及化解歧义分词。在学科分类中属于运筹学范畴。

使用概率统计运筹决策的方案很多,文献[5]通过统计字字相关的同现概率矩阵来完成汉语语用统计库结构,这个矩阵的大小是固定不变的,只与字符集的大小有关。文献[5]作者通过搜索了 500 万字语料,给出了一个 3673 ×3673 的同现概率矩阵。

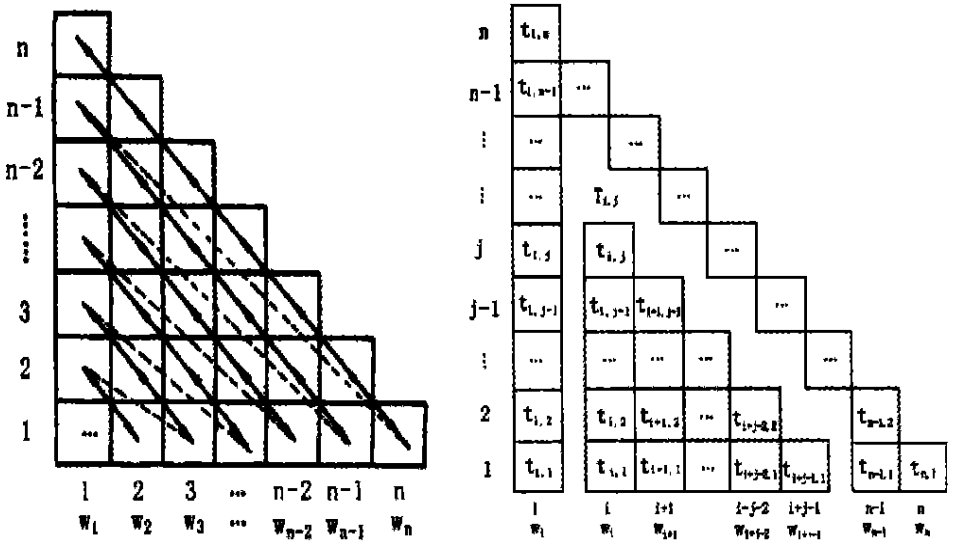

图 2 基于语用统计智能输入软件的一种算法示意图

文献[4]是基于理解和基于语用统计相结合的设计。该设计根据分词后的输入语句查找知识库,用句法、词法、语义和自定义的规则作为制约对文章进行解析推理,当存在同音词时,采用最优评价法来确定最佳选择作为转换结果。同音词的评价值,需要考虑词性、同现概率、

近期使用状况等因素。具有最优评价值的选择即为转换结果。当具有最优评价值的第一选择并非目标选择时,可选用次优选择或用手工方式进行修正,候补修正或人工修正均被记录,作为下次转换时修改计算评价值因素的依据,也就是自学习功能。

2. 典型作品

蔡榕先生设计的最优评价函数法拼音汉字转换系统;

蒋子龙先生设计的 Autoway;

清华大学人工智能实验室夏莹等研制的智能输入软件。

3. 优点与存在的问题

优点:⑴对于已经进行过语用统计或者具有相同类型的领域,系统的转换正确率比较高,或者说语用统计具有偏向性。对于每一个用户而言,在使用过程中,语用统计库将会从最初的通用型逐渐改变为符合这个用户语用习惯的专用型。⑵软件开销较小,在 CPU 为 486 的机器上就能运行。

存在的问题:⑴作为一个整体的二元同现概率矩阵,不能做到模块化、积木化。⑵偏重于整句处理,当出现转换错误时,需要使用者回头去纠正,干扰了正常的思维。⑶当前,自动分词的准确度只能达到 98 %左右,使键选率的降低受到限制。

四、基于模板匹配的智能输入软件

1. 原理

寓汉语语法知识于巨量的短语串中,进而利用这些短语串来消化同音字、词,以及化解歧义分词。这种短语串通常称之为"模板词"。

这种系统通过模板词搜索引擎来完成汉语语法体系的组织。由于需要搜索巨量的语料,获取巨量的短语串,才有可能大体上包容汉语语法知识,例如,智能狂拼搜索了 100 亿字语料,模板词库最大时需要约 540MB 存储空间。

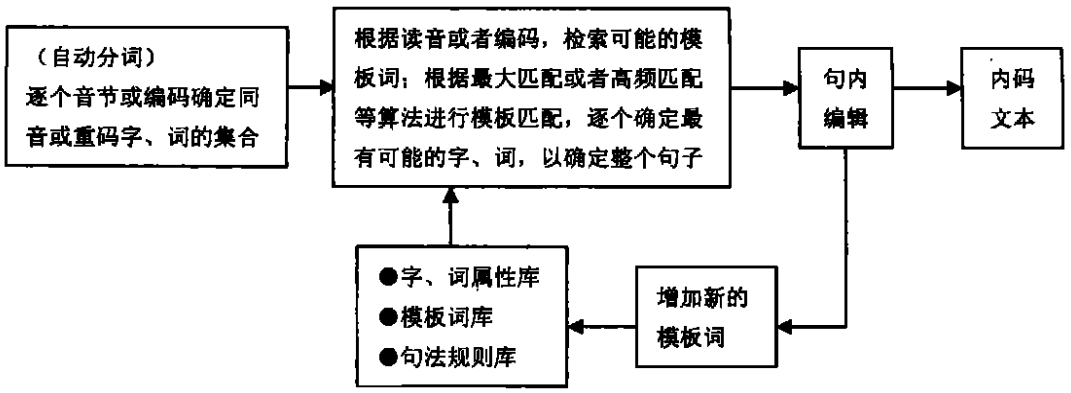

图 3　基于模板匹配的智能输入软件结构示意图

根据分词后的输入语句查找模板词库和句法规则库,然后进行匹配处理。如果匹配结果唯一,则不必再用概率推理;若存在两个以上的候选结果时,则根据句法规则或概率推断进一步判定,选出一个最有希望的可能结果作为输出。

2. 典型作品

中文之星数码科技有限公司推出的智能狂拼;

黑马电子新技术公司推出的黑马智能输入软件;

大自然软件开发有限责任公司推出的自然码 2000(句输入版) 。

3. 优点与存在的问题

优点:对于已经搜索过模板词的或者具有相同类型的领域,系统的转换正确率比较高,或者说模板词库具有偏向性。对于每一个用户而言,在使用过程中,模板词库将会从最初的通用型逐渐改变为符合这个用户语用习惯的专用型。

存在的问题:⑴由于模板词数量巨大,对电脑硬件有一定的要求,486 及以下的低档机难于使用。⑵对于拼音输入的模板匹配智能输入软件而言,通常只支持汉语拼音的 26 键位输入,注重连续和完整的音节输入,平均码长较长,采用简化拼音输入时键选率较高。⑶偏重于整句处理,当出现匹配错误时,需要使用者回头去纠正,干扰了正常的思维。⑷当前,自动分词的准确度只能达到 98 %,使键选率的降低受到限制。

五、基于上下文关联的智能输入软件

1. 原理

文献[5]采用语用统计来实现上下字关联智能输入。下面介绍的是一种基于模糊控制理论,利用上下文关联(向上关联 4 个词语,向下关联 1 个词语)的语用环境来智能选择重码字、词。在学科分类中属于自动控制分支非线性控制范畴。

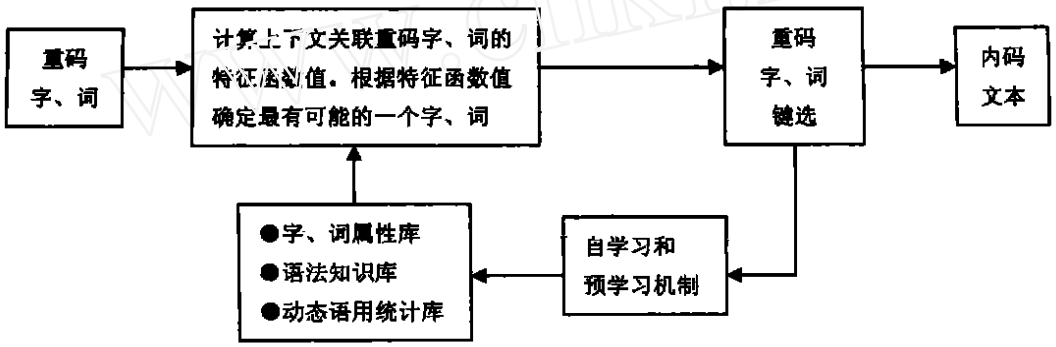

图 4　基于上下文关联的智能输入软件结构示意图

将自然语言看成是一个模糊的集合,将汉字输入系统作为一个基于非线性控制范畴的模糊控制系统来对待,预学习工具(或者转换出现错误时的手工键选信号)相当于一个传感器,算法程序、汉语知识库和动态语用统计库作为非线性调节器,使得系统的键选率和平均码长逐渐趋于最优。例如,青月亮汉字通上文关联 4 个词,下文关联 1 个词,合计上下文关联 5 个词,这一调节机制涉及到许多相互矛盾和相互牵制的受控参数,模糊集合的特征函数从[0 ,1]区间连续取值,可以较为准确地表现各种语言现象差异,获得比较好的效果。

2. 典型作品

青月亮科技开发有限公司推出的青月亮汉字通智能输入软件平台 GM3. 1;

二笔软件有限公司推出的二笔智能输入软件(26 键位和 10 键位);

字原科技有限公司推出的 101 智能输入软件 TZ8. 2/ 9. 1/ 2000。

3. 优点与存在的问题

优点:⑴对于已经预学习过或者具有相同类型的语料,键选率比较低,或者说汉语知识库具有偏向性。对于每一个用户而言,在使用过程中,汉语知识库将会从最初的通用型逐渐改变为符合这个用户语用习惯的专用型。⑵青月亮汉字通作为一种音码、形码和笔画码三位一体

的通用智能处理平台,支持 26 键位、10 键位、8 键位和 5 键位规模的键元集,支持 GB 18030 的 27533 超大字符集,为各种输入法增加上下文关联智能输入的后处理支持,让这些输入法变得更易学、更好用。⑶采用字段输入,不使用语句级输入,使语法规则简约化,易于知识表达。此举不但降低了键选率,还大大缓解了输入过程中"回头看"的问题,基本上贴近了人们的使用习惯。⑷程序开销积木化,在 CPU 为 486 的机器上就能运行。⑸遵从一系列信息处理用的国家标准和规范,特别是与汉字输入密切相关的两个标准,《GB/ T18031 - 2000 信息技术数字键盘汉字输入通用要求》和《GB(待定) 信息技术通用键盘汉字输入通用要求》。⑹在拼音输入时,采用人工分词,在形式上与英文接轨,既可以避免 3 %的歧义分词错误,也可以兼容简拼输入,大幅度降低键选率和平均码长。青月亮汉字通在保证支持完整拼音输入的同时,尤其提倡使用简拼输入(一个音节要么只取音节的首字母,要么就取完整音节)。

存在的问题:

⑴字段输入还未能完全根治输入过程中"回头看"的问题,当终选字词有错时,仍然需要近距离的即时修改。⑵对于"上下文关联"机制的人机界面,用户需要一个熟悉的过程,因此青月亮汉字通也允许使用者关闭"上下文关联"智能输入,仅保留"上文关联"机制和恢复逐字、逐词上屏的输入方式。

六、结束语

上世纪八十年代我国学者提出的"从开发人脑到开发电脑"、"字为基础、词为主导、智能处理"只是指明了汉字输入技术的发展方向。时至今日,已有近十种输入法采用了智能化处理重码的技术。这些智能输入软件中有的采用人工智能理论,有的依据概率统计方法,有的依靠自动控制技术,多数软件设计两种方法并用。效果较好的青月亮软件更是在模糊控制的框架内使用了语法规则和动态语用统计。如果像当年"万码奔腾"一样,智能化软件也形成"百花齐放,百家争鸣"的局面,则钱伟长教授要求的易学、易用的理想的输入方法的实现为期已经不远了。

感谢:在本文撰写过程中,周志农先生参加了讨论会并提出有益的建议。

参 考 文 献

[1] 俞士汶. 中文输入中语法分析技术的应用. 中文信息学报,1988(3).
[2] 王晓龙等. 语句级汉字输入技术. 中文信息学报,1996(12).
[3] 章森等. 语句拼音——汉字转换的智能处理机制分析. 中文信息学报,1998(2).
[4] 蔡榕. 最优拼音汉字一次输入变换法及拼音汉字转换系统的实现. 第三届中文信息处理国际会议论文集,1992. 10.
[5] 夏莹等. 利用上下文相关信息的汉字文本识别. 中文信息学报,1996(2).
[6] 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,1993(1).
[7] 仲兴国. 多词组一次性拼音汉字变换. 中文信息学报,1990(6).
[8] 张普. 智能化——汉字键盘输入法的最重要发展方向. 中国标准出版社,1997(2).
[9] 陈一凡,汉字编码输入技术的发展趋势. 计算机世界,1987. 11(8).
[10] 陈一凡、胡宣华. 汉字键盘输入技术与理论基础. 清华大学出版社,1994(6).
[11] 黄昌宁. 中文信息处理的主流技术是什么? 计算机世界报第 24 期,2002. 7.