简繁汉字转换系统的设计与实现*

辛春生, 孙玉芳

(中国科学院 软件研究所,北京 100080)

E-mail: yfsun@sonata.iscas.ac.cn

http://www.ios.ac.cn

摘要: 汉字简繁体转换对于港澳台及世界华人地区与中国大陆之间文化经济的交流极其重要.以已完成并投入使用的一个自动转换系统为基础,介绍了系统的设计与实现.给出了系统的总体结构,描述了系统的重要数据结构——词库和对照表.解释了系统处理流程,包括预处理和后处理、消除歧义、语词转换等,并对系统性能进行分析,同时,给出了测试结果.

关键词: 简繁汉字;语词转换;语词切分;歧义消除;语词库;对照表

中图法分类号: TP391 **文献标识码**: A

随着港澳的回归,海峡两岸及世界华人文化交流的日益频繁,利用电脑技术进行简繁汉字自动转换对于汉字文化以及经济交流的发展十分重要.

由于汉字简繁转换[1]在单字对应、术语应用、使用习惯及字符集标准等诸多方面存在许多复杂的和不定的因素,因此,设计和开发一种全自动、高效的简繁汉字转换系统有许多需要下力气解决的问题[2].本文以我们已完成并投入使用的一个自动转换系统为基础,介绍该系统的设计与实现工作.

本文第1节介绍系统结构.第2节描述系统的重要数据结构——语词库和对照表.第3节解释系统处理流程,包括预处理和后处理、切分算法、消除歧义、语词转换及学习与自适应算法.最后,对系统性能作一分析,并给出相应的测试结果.

1 系统结构

系统支持的内码为双字节高位置"1"和Big5,做到3个交换码集(GB 2312-80(G0),TCA-CNS 11643,ISO/IEC 10646或Unicode[3])之间两两互换.系统的结构如图1所示.图的中间为系统处理流程,左右是处理时所涉及到的数据结构,即一系列语词库和对照表.

2 系统数据结构

系统有4种语词库:通用语词库、通用术语语词库、专业术语语词库和用户定义语词库.对语词库的组织采用模块化方法,根据用户的设置参数决定语词库的优先次序,也可以设置成不用某个语词库的形式.对照表有5种:单字对照表、一对多语词对照表、专业术语对照表、通用术语对照表和用户定义对照表.对照表的组织和语词库的组织相同,也采用模块化方法.有时,对照表和语词库放在一起,如专业术语语词库和专业术语对照表就是结合在一起的.下面讨论语词库和对照表的结构.

2.1 通用语词库的结构

通用语词库目前只包含词条的词频统计信息,按照词条的长度,分别放在cw[1~7].lib这7个文件中

* 收稿日期: 1999-05-07; 修改日期: 1999-09-06

基金项目: 国家"九五"重点科技攻关资助项目(96-B08-01-03;98-779-01-02)

作者简介: 辛春生(1975—),男,江西万载人,硕士,主要研究领域为中文信息处理;孙玉芳(1947—),男,江苏张家港人,研究员,博士生导师,主要研究领域为系统软件,中文信息处理,大型数据库,网络工程.

(cw$X$.lib中的词条长度为 X 个字).

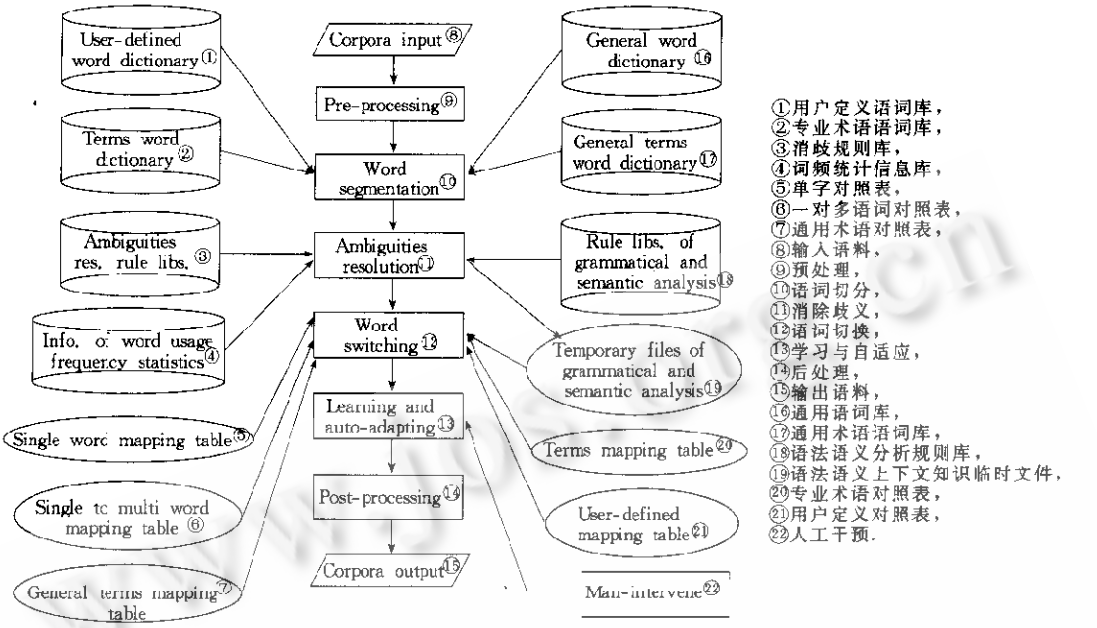


Fig. 1 The system architecture
图1 系统结构

为了减少初始化的开销,本系统把通用语词库组织成一个二进制库文件形式.初始化时,系统检测二进制库文件是否存在,若不存在,则创建一个子进程,子进程调用一个实用程序,它依次读取 cw$X$.lib 文件,进行初始化,存入结构表中,初始化完成后,把结构的所有内容写到磁盘上的库文件中.父进程在等待子进程结束后,把库文件读入结构表中,继续进行初始化.这里,似乎多了一次读盘操作,但这样做的优点有两个:(1) 用户可以直接使用实用程序更新二进制库文件.(2) 父、子进程结构清晰,各司其职,由子进程来执行一个实用程序,减少了程序的代码量.

2.2 专业术语语词库和对照表的结构

本文只针对计算机专业语料进行专业术语转换处理.但是,该算法是通用的,只要有了其他专业的专业术语语词库和对照表,就可以处理此专业的语料.

为了减少信息冗余并节省存储,计算机专业术语语词库和对照表结合在一起,按照词条长度放在 cw1[1~7].lib 这 7 个文件中.与通用语词库一样,为了节省开销,把专业术语语词库也组织成一个二进制的库文件.

2.3 其他语词库和对照表的结构

通用术语语词库和对照表、用户定义对照表和语词库都组织成一张表,表的左右两列分别是简体形式和繁体形式.一对一的单字对照表是一个一维数组,一对多单字对照表是一个二维数组.

3 系统处理流程

3.1 预处理和后处理

由于大量的电子出版物都是某种排版软件的格式文档,因此,简繁转换不能仅处理纯文本,它还应能处理各种格式文档,而格式文档中仅正文部分需要转换,因此,在转换之前对输入语料进行预处理,抽取出正文部分,待转换完成之后再嵌回原来的位置,即作后处理.目前所支持的文档格式为 Word 格式、html 格式、Postscript 格式等,今后将增加对港台常用排版格式和大陆常用排版格式的转换.

3.2 切分算法

由于语词库包含的信息少,例如,它的词条没有标出词性,因此,系统选择了一个机械切分算法,即双向扫描

法,对语料进行切分.当我们往语词库中加入词性等其他信息以后,就可以用精度更高的切分算法替换系统中现有的双向扫描切分算法.因为系统是模块化组织的,这种替换对其他部分不会产生很大的影响.

双向扫描法利用MM(the maximum matching method)方法和RMM(the reverse directional MM)方法[4]进行两次切分,得到两条词链,再对这两条词链进行切分后处理,消除歧义,从而得到最后结果.

3.3 消除歧义

消除歧义分3步完成.首先,利用消歧规则消除歧义;其次,对上一步不能处理的歧义,利用语法和语义分析得到的上下文知识来消除;最后,若前两步都得不到有效处理,则利用词频统计信息来消除.下面逐一进行讨论.

3.3.1 消歧规则库的设计和组织

语词库可供利用的信息决定了现阶段的消歧规则不会很复杂,消歧规则库中主要是语词搭配规则,它利用词之间的依赖关系消除歧义.这些规则对方位词、连词、副词的处理比较有效.在扩充了语词库的信息之后,就可以加入比较复杂的规则,增强消除歧义的能力.本系统的规则库分为两类:

(1) 系统规则库:包含系统预定义的规则,在系统设计时预先放入系统中.

(2) 用户规则库:包含用户定义的规则,用户可随时对这个规则库进行增删.

3.3.2 统计方法的运用

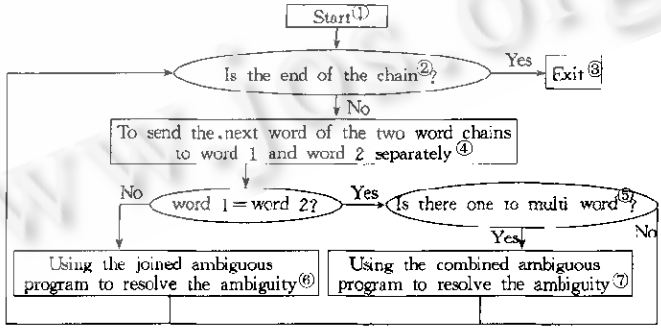
歧义字段从构成形式上可分为两类[4]:在交集型歧义字段中,最常见的是两个二字词交叉形成的切分歧义,如果没有相应的语词搭配规则,则利用词 AJ 和 JB 的词频统计来决定切分结果.

在组合型歧义字段中,可以选取几个加权因子,以使词 AB 的词频和词 A、词 B 的词频可以比较公平地比较.因为根据大词切分规则,词 AB 是正确切分的可能性很大,但由于词 A 和词 B 的词长比词 AB 的小,它们的使用频率将比词 AB 的大得多,有时甚至要大好几个数量级,因此必须给它们乘以一个加权因子,使它们具有可比性.加权因子依赖于词的总体使用频率,例如,二字词的总体使用频率.

3.3.3 其他问题

当利用规则不能消除切分歧义时,可以先对语料进行语法和语义分析,得到句子的上下文语法结构信息,再据此消除歧义.由于对语料进行语法和语义分析是一个很耗时的过程,因此,用户可以选择是否进行语法和语义分析.此模块可以放在动态库中,不会增加系统目标码的大小.另外,统计方法中的加权因子不应该是一成不变的,在系统实现时,应该用相当的语料进行统计,参考词的总体使用频率选择一个较佳的值,随着系统的不断学习,应该在原值上增加偏移,这样可以针对不同的应用增强适应能力而赋予系统一定的智能.

3.3.4 消除歧义的程序框架(如图2所示)



①开始,②到词链尾吗,③退出,④把两条词链的下一个词分别送入word 1,word 2,
⑤有一对多的字吗,⑥用交集型歧义处理程序消除歧义,⑦用组合型歧义处理程序消除歧义.

Fig. 2 Ambiguities resoluting
图2 消除歧义

当由MM算法和RMM算法切分出两条词链后,系统就对其进行处理,以便消除歧义.由于本系统旨在成功地转换,因此,对于没有一对多汉字的句子的处理与有一对多汉字的句子的处理是不一样的.若句子无一对多汉字,则不管两条词链是否相同,均以RMM方法得出的词链为准.若句子有一对多的汉字,则无论两条词链是

否相同,均需进行处理;并且如果词链相同,则可能是组合型切分歧义,如果词链不同,则可能是交集型切分歧义.

3.3.5 交集型歧义字段的处理

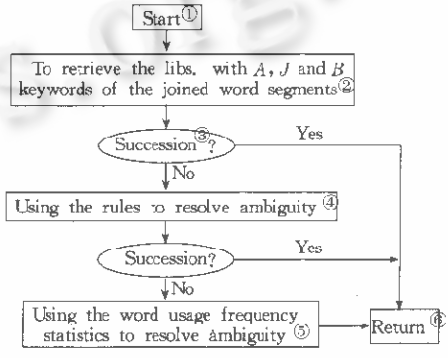
对交集型歧义字段的处理,首先利用用户规则库和系统规则库,若能达到消除歧义的目的,则退出.若利用规则不足以消歧,则利用词频统计消歧.使用词频统计方法,要注意区分专业术语和通用语词.因为专业术语没有词频统计,即使有,也不应该在专业术语的词频和通用语词的词频之间进行比较.因此,本系统设立了一个标志以区分专业术语和通用语词.若一个词是通用语词,则它的词频大于或等于0.词频为0的情况是极少见的,只有那些不在切分语词库中的单字词其词频才可能为0.这些不在切分语词库中出现的单字词一般不能单独成词,万一碰到这种情况,则记它的词频为0.对于专业术语,则记它的词频为-1,以便识别.如果歧义字段中有专业术语,系统则认为按专业术语切分的词链是正确的.处理程序的描述如图3所示.

交集型歧义字段一般由两个词交叉组合形成,但也有可能由多个词交叉组合形成.由于这类交集型歧义比较少见,所以只需用词频统计对其进行处理.我们把由MM方法切分得到的组成这类交集型歧义字段的词记为词链A,相应地,把由RMM方法切分得到的词记为词链B.根据词链中有无专业术语,可分为以下4种情况:

(1) 词链A中含专业术语、词链B中不含专业术语,系统认为词链A切分正确.

(2) 词链A中不含专业术语、词链B中含专业术语,系统认为词链B切分正确.

(3) 词链A和B均含专业术语,以词数最少的词链或RMM算法的切分结果为准.



①开始,②以交集型字段的A,J,B为关键字查找规则库,③成功,④利用规则消除歧义,⑤利用词频统计消除歧义,⑥返回.

Fig. 3 Processing of the joined ambiguous word segments
图3 交集型歧义字段的处理

(4) 词链A和B均不含专业术语,利用词频统计进行处理:词链A的词乘以加权因子,累加后送入$freqA$;词链B的词乘以加权因子,累加后送入$freqB$.然后比较$freqA$和$freqB$,得出切分结果.

3.3.6 组合型歧义字段的处理

对于专业术语,不进行组合型歧义处理,这主要有两个原因:首先,专业术语没有词频统计,可供利用的信息几乎没有;其次,对于专业术语来说,如果它有组合型歧义,又可以分两种情况:(1) 词A或词B是通用语词,这时,应该切分为AB,即这个专业术语不应该被拆开;(2) 词A和词B都是专业术语,这种情况也应该大词优先,即切分为AB.对于通用语词,如果它有组合型歧义,则还是利用规则辅以词频统计来处理.首先,查找系统规则库和用户规则库.如果查找成功,则对其进行处理,处理成功则返回.如果不成功,则利用3个词的词频消除歧义,由于词长不同,它的总体使用频率相差是很大的,这就需要乘以一个加权因子以达到可比性.这在前面的小节中已有描述.处理程序的描述如图4所示.

3.4 语词转换

消除歧义后,系统进行语词级转换.根据当前待转换语词的类型,可分为以下几种情况:(1) 此语词是用户定义语词,利用用户定义对照表加以转换.(2) 此语词是一对多的语词,利用语词搭配规则或语法语义分析得到的上下文知识或词频统计信息加以转换.此语词可能是普通语词,也可能是专业术语或通用术语,这时,可先查找专业术语对照表或通用术语对照表.(3) 此语词是一对一的语词,但含有一对多的单字,查找一对一语词对照表,加以转换.此语词可能是普通语词,也可能是专业术语或通用术语,这时,可先查找专业术语对照表或通用术语对照表.(4) 此语词既不是一对多语词,也不含一对多单字,则利用单字对照表加以转换.若是术语,也应该先查相应的术语对照表.

为了加快查找一对多字,系统以这些字为关键字建立一个HASH表,数据项中有指向对照表的指针.为了

节省存储,没有建立一对一的普通语词库,而使用了一种折衷方案,即建立单字对照表和含有一对多字的语词对照表.这样,一对多语词的转换就是通过查找上述两个对照表而进行单字转换,从而使一对一语词中的一对多单字可以得到正确转换.

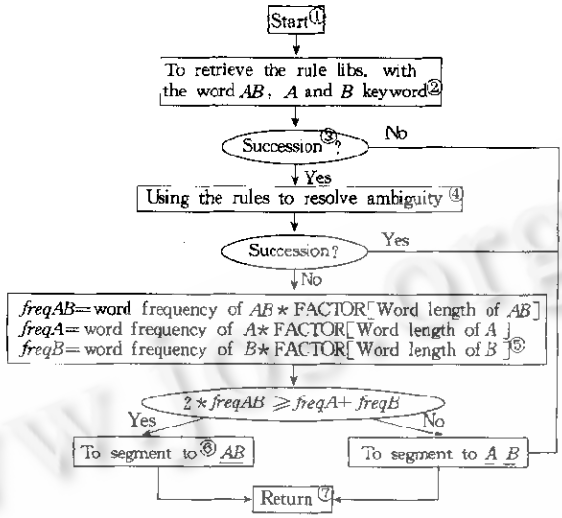


Fig. 4　Processing of the combined ambiguous word segments
图4　组合型岐义字段的处理

3.5　学习与自适应

大多数用户一般都局限于一些特定的应用,用户的干预往往是基于这些特定应用的专业知识,而这正是系统所缺乏的.因此,系统必须把正确的转换记忆下来,修改系统内部的相关信息,以便在下次碰到时增大转换正确的可能性.另外,专业领域的知识更新很快,大量的新词、新术语会不断涌现,系统也把这些词加入自己的语词库中.为了增强系统的学习能力和方便用户,应把人工干预的界面设计得尽可能友好,把可能的错误转换用彩色或反显标出.用户用鼠标点击后则弹出所有可能的转换,可任意选取一个,也可以直接输入正确的词.用户甚至可以用鼠标选定整个句子,点击后弹出此条语句的所有可能的切分结果,通过用户的干预,将信息反馈到切分规则库中.

4　小　结

本文在自动转换简繁体汉字方面做了一些探索.系统已在试运行,源代码约5 000行,分为4个源文件和5个前导文件(不包括词库和对照表).

4.1　系统性能分析

对一个简繁汉字自动转换系统来说,其性能可以分为以下3个方面:

(1) 系统的可维护性

本系统除了可以让用户方便地往规则库中增删规则之外,还提供了一些实用程序,以增强系统的可维护性.

(2) 转换正确率

对于普通语料,一对一的汉字不会出现转换错误,转换出错只能发生在一对多的汉字上.产生错误的原因主要有:① 切分错误导致转换出错;② 切分正确但转换时出现了错误.通常来说,第2种情况的可能性比较小,且只要把一对多汉字相关的词收入足够多,转换出错的概率就可以降低到最小.切分精度的提高主要是看切分算法对歧义字段的处理能力.

根据对两万个语料的统计,本系统采用的双向扫描法能检查出大约90%的交集型歧义字段.而采用语词搭配规则和词频统计方法,可以处理约80%的被检查出的交集型歧义字段.这样,所有交集型歧义字段的72%可

以得到正确处理.对于组合型歧义字段,采用的是加权词频统计方法来消除歧义,可以消除约50%的组合型歧义.

(3) 系统的转换速度

系统的转换速度主要取决于语词切分的速度、歧义处理的速度和输出转换的速度.语词切分是系统最耗时的,因为它需要查找庞大的语词库.而歧义处理和输出转换相对来说要快得多.

4.2 测试结果

我们从《人民日报》随机选取了19 540个字的普通语料,在主频100MHz的PC机和SCO OpenSever 5.0上测得汉字由简到繁转换的一些数据,见表1.另外,我们也随机地从《中国计算机报》选取了16 051个字的计算机语料进行汉字由简到繁的转换,测得的数据见表2.

Table 1 The testing result of general simplified to traditional Chinese corpara conversion

表1 普通语料汉字由简到繁的转换测试数据

Conversion time (sec.)①	0.93
Conversion speed (words/sec.)②	21 010
The occurring numbers of the joined ambiguous word segments③	150
The joined ambiguous word segment numbers to be tested④	136
The joined ambiguous word segment numbers to be segmented correctly⑤	109
The occurring numbers of the combined ambiguous word segments⑥	29
The combined ambiguous word segment numbers to be segmented correctly⑦	15
The Chinese word numbers to be converted incorrectly⑧	14
The conversion correction⑨(%)	99.93

①转换时间(秒),②转换速度(字/秒),③交集型歧义字段出现次数,④检查出的交集型歧义字段总数,⑤得到正确切分的交集型歧义字段总数,⑥组合型歧义字段出现次数,⑦得到正确切分的组合型歧义字段总数,⑧错误转换的汉字总数,⑨转换正确率.

Table 2 The testing result of the special purpose simplified to traditional Chinese corpara conversion

表2 专用语料汉字由简到繁的转换测试数据

Conversion time (sec.)①	0.8
Conversion speed (words/sec.)②	20 000
The occurring numbers of technical terms③	2 372
The technical terms numbers to be converted incorrectly④	8
The conversion correction of the technical terms⑤(%)	99.7
The Chinese word numbers to be converted incorrectly⑥	24
The conversion correction⑦(%)	99.85

①转换时间(秒),②转换速度(字/秒),③专业术语出现次数,④错误转换的专业术语数,⑤专业术语转换正确率,⑥错误转换的汉字总数,⑦转换正确率.

这里的转换速率包括了系统初始化的时间.对于专业术语的转换,错误主要发生在一些容易和通用语词混淆的词上,而对于普通语料的转换,错误发生在单字词上的可能性比较大,这个问题在完成语法分析和语义分析的工作之后将会得到比较好的解决.

References:

[1] Liu, Shing-huan. An automatic translator between traditional Chinese and simplified Chinese in unicode. In: Proceedings of the 7th International Unicode Conference. San Jose, CA: Unicode Consortium, 1995.

[2] Du, Lin, Wu, Jian, Sun, Yu-fang. The analysis and implementation of the intelligent conversion system JFC for simplified and traditional Chinese languages. In: Proceedings of the 7th Annual Chinese Coding Special Subcommittee of Chinese Information Association of China and the 5th Annual Chinese Information Technology Special Subcommittee of China Computer Society. Suzhou: Chinese Information Processing Society, 1996 (in Chinese).

[3] Wu, Jian, Sun, Yu-fang. ISO10646 character set and its internal coding schemes. In: Proceedings of'96 Hong Kong/

Chengdu Joint International Computer Conference. Chengdu: Southwest Jiaotong University Press, 1996 (in Chinese).

[4] Liu, Yuan, Tan, Qiang, Shen, Xu-kun. Modern Chinese Word Segmentation Specification and Automatic Segmentation Methods for Information Processing. Beijing: Qinghua University Press; Nanning: Guangxi Science and Technology Press, 1994 (in Chinese).

附中文参考文献:

[2] 杜林,吴健,孙玉芳.智能简繁汉语转换系统JFC的分析与设计.见:中国中文信息学会汉字编码专业委员会第7届年会暨中国计算机学会中文信息技术专业委员会第5届年会论文集.苏州:中文信息学会,1996.

[3] 吴健,孙玉芳.ISO10646字符集及其内码体系.见:'96香港、成都国际计算机会议论文集.成都:西南交通大学出版社,1996.

[4] 刘源,谭强,沈旭昆.信息处理用现代汉语分词规范及自动分词方法.北京:清华大学出版社;南宁:广西科学技术出版社,1994.

Design and Implementation of a Simplified-Unsimplified Chinese Character Conversion System

XIN Chun-sheng, SUN Yu-fang

(*Institute of Software, The Chinese Academy of Sciences, Beijing* 100080, *China*)

E-mail: yfsun@sonata.iscas.ac.cn;

http://www.ios.ac.cn

Received May 7 1999; accepted September 6, 1999

Abstract: Simplified and unsimplified (traditional) Chinese character conversion is very important for the cultural and economic exchange between the mainland of China and overseas Chinese speaking areas. In this paper, the design and implementation of an automatic conversion system is presented, which has been put into use. The system architecture is shown firstly, and some important data structures, the word dictionaries and mapping tables are presented. Then, the system processing flow is outlined, which includes preprocessing and post processing, ambiguity resolution, word conversion, etc. Finally, the system performance analysis and the test result are listed.

Key words: simplified and unsimplified (traditional) Chinese character; word conversion; word segmentation; ambiguity resolution; word dictionary; mapping table