基于词语消歧的分层次汉字简繁转换系统

刘汇丹 吴健*

（中国科学院软件研究所基础软件国家工程研究中心， 北京 100190）

1、 引言

近些年来，两岸四地在科技、文化、教育、新闻、出版等领域的交流与合作日益广泛和深入，造成海峡两岸信息交换的剧增。但是文字制度上几十年的差异，造成了两岸在常用语、专业术语、外国人地名翻译等各方面都存在着差异，另外繁体字和简体字数量都比较大，限于普通人的文字知识水平，难免造成交流上的困难。所以迫切需要一套汉字简繁转换系统。

本文对汉字简繁转换中涉及的众多问题进行讨论，提出基于词语消歧的分层次汉字简繁转换方案，并据此实现了一个简繁转换系统。

2、 汉字简繁转换的复杂性

简单的来看，汉字简繁转换问题主要在于汉字简化时将多个繁体字做了归并，从而造成一个简体字对应多个繁体字的情形（同时也存在一个繁体字对应多个简体字的情况）。多目标字的存在造成了简繁转换的歧义。从转换的角度来看，如果存在转换歧义，自然而然要用更大的语言单位的转换来消除这种歧义：单字转换的歧义要用词语来消除，词语转换的歧义要用短语来消除。完整的转换过程涉及到从底层的编码到上层的字、词（词汇）、语等多方面的问题，下面对这些问题做统一的讨论。

2.1 编码字符集

GB2312-80 只收录了简体中文 6763 个常用汉字和次常用汉字[1]，TCA-CNS11643-1992 收录汉字 13053 个，两个字符集不仅都没有包含所有的简体字和繁体字，前者未收录所有的简体字，后者也没有收录所有的繁体字。导致的问题就是在做简繁转换的时候必定要做编码转换，并且有可能存在 GB 编码的简体源字却不存在相应的 BIG5 编码的繁体目标字。虽然它们的衍生字符集增收了不少的汉字，使得此问题在一定程度上有很大的改观，但是限于两岸四地用户的使用习惯，强迫繁体用户使用 GB 编码或者简体用户使用 BIG5 编码都是不合适的。

国际标准编码字符集 Unicode/ISO-IEC10646（以下简称 Unicode）为世界上所有的文字进行统一的编码[2]，给每一个字符唯一的一个编码表示。Unicode 4.0 按照 CJK 认同规则共收录汉字 70205 个，其中在基本多文种平面（BMP）内收录汉字 27484 个，包含了现有规范中所有的简体字以及日常所用的繁体字，非 BMP 平面内的汉字一般用于大型工具书、古籍整理等类似的应用场合。所以，基本上 BMP 平面就可

* **作者简介**：刘汇丹(1982－)，男，硕士，助理工程师，主要研究方向是系统软件与中文信息处理；吴健(1962－)，男，研究员，主要研究方向是系统软件与中文信息处理。

以满足简繁转换的需求。

在简繁转换中采用 Unicode 字符集将有如下优点：

- Unicode 将简体字和繁体字都收录了，可以在同一个字符集内完成简繁转换；
- 在 Unicode BMP 平面内解决简繁转换问题，可以采用等长编码，方便系统实现。
- Unicode 对所有文字统一编码，在转换包含其它文种的文档时可以避免信息丢失；
- Unicode 是国际标准，两岸四地用户对其都有较高的认同感，避免了简体字用 BIG5 编码或者繁体字用 GB 编码的不习惯。

因此，在简繁转换中采用 Unicode 是比较好的选择。

2.2 单字转换

单字转换层面的主要的问题，一是简繁字范围的确定，二是多目标字转换如何消歧。

2.2.1 简繁汉字范围的确定

2.2.1.1 字形差异

表 1 关联字示例

概念	示例字	概念	示例字
正异字	嘆 vs 歎	正讹字	盜 vs 盗
繁简字	恳 vs 懇	新旧字	骨 vs 骨
中日字	價 vs 価	形近字	辨 vs 辦
古今字	燃 vs 然	通假字	蚓 vs 螾

汉字的关联字包括正异字、简繁字、中日字、古今字、正讹字、新旧字、形近字、通假字等概念。表 1 给出了一些示例。这几个概念往往难以划出严格的界限，某二字之间可能兼有繁简体与古今字之关系，新旧字也可能是正异字关系等等。由于关联字之间字形相似，导致了简繁转换单字对照关系不容易确定。例如如下的两组字：

1. “蜈”和“蜈”；
2. “宫”和“宮”。

两组字字形都很相似，但是却是两种截然不同的关系。第 1 组中，两个字是同一个编码，只是前者采用了“宋体”显示，后者采用了“PMingLiU”显示，从而显示出不同的字形，它们其实是同一个字。第 2 组中，两个字根本就是两个编码，是两个字。在简繁转换中，第 1 组的情况是不需要考虑选字的问题的，而第 2 组的情况，就需要决定选择哪一个字了。然而由于字形相似，仅仅观察字形是很难区分是属于上述两组中哪一种情况的。虽然有《简化字总表》、《第一批异体字整理表》等比较权威的资料，但是此两表收录的字并不全面，所以在系统实现时仍然没有一个十分明确的标准可循。

因此，简繁转换系统应当包括哪些字，这还需要相关的语言学专家来做明确的限定。

2.2.1.2 生僻字

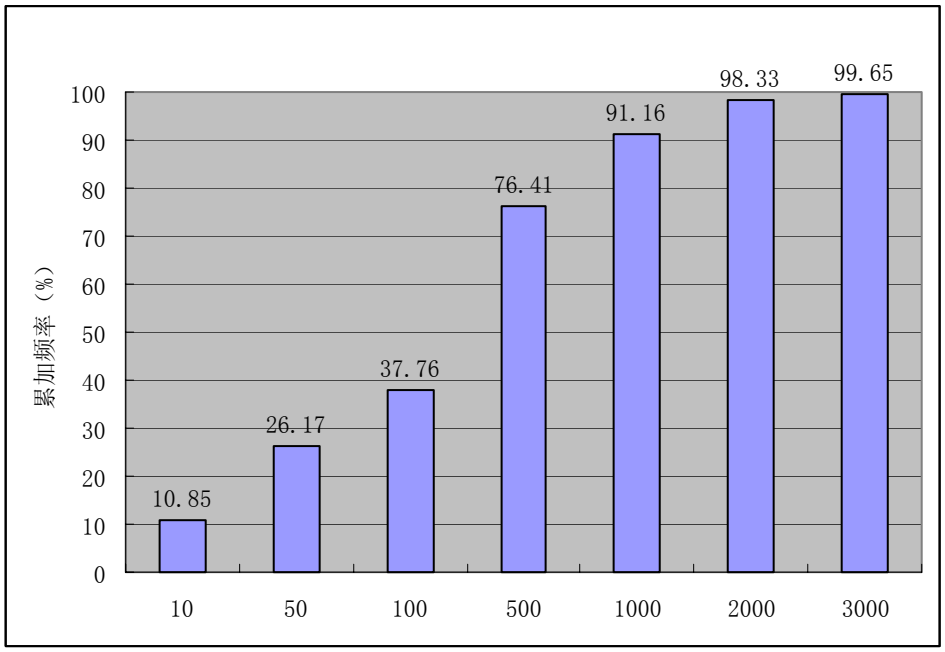

图 1 汉字高频字累加频率图

国家语言文字工作委员会对 2005 年的报纸、广播电视、网络用字的统计数据显示，按使用频率排列，前 50 个字的累加频率为 26.17%，前 2000 个字的累加频率为 98.33%[3,4]（见图 1）。而文献[5]的统计数据显示，最常用的 2000 个简体字占统计素材所有字的 97%。而简化字总表中所列的简体字有 2244 个，那么至少有 200 多个简体字是生僻字，如“佟”、“哐”、“嘧”、“怆”、“拐”、“拗”、“扩”、“栲”等[6]。

虽然生僻字出现的频率很低，似乎没有考虑的必要，但是如果不包含对生僻字的转换支持，问题总是解决得不彻底。结合具体的实现，BMP 平面内的字符在计算机里都可以统一使用两个字节表示，而 BMP 平面以外的字符需要用更多的字节表示，如果加入对 BMP 平面以外少数字符的支持，将造成存储空间的浪费，并且造成输入输出时编解码的复杂化。

所以，我们认为简繁转换至少应该包含对 BMP 平面内所有汉字的转换支持。

2.2.2 多目标字转换的消歧

表 2.和前后字都不成词的多目标字示例

例字	例句
干	昨日一天暴晒，已经干了的道路，这一夜雨又浇泞了。
后	两天后的傍晚陈真又到海滨旅馆去找周如水。
表	佩珠打算回去，她摸出表来看，快到十二点钟了。
划	她用颤抖的声音说了这句话以后，就握紧桨拚命地划。
里	目的地是四十里外的燕子崖。

简繁转换中，对于单目标字，直接转换就可以；而对于多目标字（在简体转繁体的时候有数百个，在繁体转简体的时候约有 20 个），一般采用词语转换来消除单字转换的歧义，但是语言现象多种多样，在一些句子中，有些字和其前面的字不能成词，和其后面的字也不能成词。表 2 中列出了这类单字的例

子[7,8]，这类单字还可以再分为两类，表 2 中“干”、“后”、“里”的例子中虽然和前后字都不成词，但是仍然可以利用前后字的隐含信息来进行转换，将“干了”、“两天后”、“十里”作为词收入词语转换的对照表就可以利用词语转换进行消歧。而“表”、“划”的例子中，却无法利用前后字进行消歧，需要寻找其它的途径来解决，这是目前简繁转换的一个难点。

2.3 词语转换

词语转换包含两个方面，一方面是由于单字转换时存在一对多的对照关系，所以需要用更大的语言单位（词）来消除单字转换的歧义；另一个方面是由于简繁文字的分化已经有了几十年的历史，两岸四地在科技术语、外国人名地名音译等各方面都存在一些差异。

2.3.1 通用词语

表 3 有转换歧义的通用词语示例

源词	目标词 1	目标词 2
编制	編制	編製
制作	制作	製作
白干	白幹（白做了）	白干（酒）
下面	下面（方位）	下麵（下面条）

对于通用词语来说，如果词中不含多目标字，那么在转换时按单字转换即可，否则需要采用词语转换来消除多目标字转换的歧义。据统计，在总数 346294 条的简体词表中，含有多目标字的词语有 70906 条，占 20.48%。大部分通用词语的简体和繁体之间都是一一对应的关系，转换时不存在问题。但是也有个别的词语会有两个或者更多的目标词可选。表 3 列出了转换时有歧义的几个通用词语[9]。简繁转换中，这类词语的转换歧义在词语转换的层面仍然无法消除，这是简繁转换的另一难点。

2.3.2 成语和诗词

表 4 含有多目标字的成语和诗词示例

例字	词句
面	牛头马面、蓬头垢面、不识庐山真面目
干	枝干相持、油干灯尽、口干舌燥、精明能干
只	脚踏两只船、形单影只、片言只语
松	岁寒知松柏、衣宽带松、鹤发松姿

成语和诗词是汉语的精华，在写文章的时候，作者往往会引经据典，用一些成语和古诗词名句来点缀文章，起到了锦上添花的效果。然而有很多成语和诗词名句中含有多目标字，表 4 给出了含有多目标字的成语和诗词的示例。成语和诗词中的用字情况比普通词语更加复杂，要解决这类问题，怕是只能将这些成语和诗词名句收录到词语转换的对照表中。

2.3.3 科技名词

表 5 海峡两岸软件方面科技名词差异示例

英语	大陆	台湾	英语	大陆	台湾
Software	软件	软体	Assembly	汇编	组合

Operating system	操作系统	作业系统	Job	作业	工件
Program	程序	程式	Statement	语句	陈述
Program design	程序设计	程式设计	Dump	转储	倾印，倒出
Benchmark	基准程序	基准程式	Scheduling	调度	排程
Routine	例程	常式	Spooling	假脱机	排存

由于海峡两岸各自按照自己的理解和习惯来定名，所以在科技术语方面存在着严重的不一致（严格地讲，台湾、港澳、大陆、新加坡等各地在科技术语方面都存在一些差异，这里暂时不考虑[10]）。表 5 表列出了海峡两岸计算机软件方面科技名词的差异。

在简繁转换中，科技名词的转换会带来一些问题，主要有三个方面。

一是由于单字术语引起的。例如在计算机学科中大陆称为“宏”，而台湾称为“巨集”，大陆称为“位”，台湾称为“位元”，大陆称为“硅”，台湾称为“矽”等等，这类术语在转换的时候和单字的转换存在歧义，例如在碰到“位”时应该当作单字不转还是应该当成科技名词转换为“位元”呢？

二是由于科技名词本身存在一对多的对应关系。例如在计算机学科中，大陆称为“文件”，根据上下文的不同台湾称为“文件”或者“档案”，在做转换时很难判断是否应该将“文件”转换为“档案”。

三是由于一些科技名词本身也可以作为通用词语使用，但是在作为科技名词时有其特殊的含义，例如简体的“程序”一词，在计算机学科中指的是使用计算机语言编写的代码或者软件，对应台湾的科技名词“程式”。但是作为通用词语时指的是工作的流程，例如国际会议一般会设置“程序委员会”。在简繁转换时碰到这样的词语应该把它当作通用词语还是科技名词呢？

科技名词在简繁转换时存在一个词语对应多个候选目标词的问题，目前还没有切实可行的解决办法。

然而，科技名词是否转换，涉及的只是适应用户习惯的问题，转或不转，不是对与错的问题，我们倾向于交给简繁转换系统的用户决定是否转换科技名词。

2.3.4 国内命名实体

命名实体主要包括人名、地名、机构名、时间词语等，在简繁转换中对于人名、地名、机构名需要区别对待。

2.3.4.1 人名

目前所发现的问题有两种。

第一种是人名中的姓氏问题。简繁转换中多目标字的目标字中可能有一个或多个可以作为姓氏，如简体的“范”字的两个繁体目标字为“范”、“範”，其中“范”作为姓氏是一大姓，“範”也是一个姓氏，但是姓此姓的人极少[11]。而更严重的问题是，像简体的“于”，对应繁体的“于”和“於”，两个目标字都是姓氏，在简体转换为繁体时即使识别出来“于”是一个姓氏，仍然不能确定它应该转换为哪一个目标字[12]。这样的情况怕是手工转换也难以决定该选择哪一个目标字。

第二种是人名中的名。例如“周润发”的“发”转换为繁体之后应该是“發”而不应该是“髮”[11]，但是文献[13]中提到中国现代文学史上有位象征派诗人叫李金发，繁体为“李金髮”。另外，如果人名用字中包含了单字术语，也会出问题，在我们对微软 Office 2003 的测试时无意中发现将人名“X 宏 X”转

换成了“X 巨集 X”。

总的来说，如果加入了对人名的识别，总还是能解决一部分问题。

2.3.4.2 地名

表 6《简化字总表》公布时部分被简化的地名

江西省	贵州省	四川省	新疆维吾尔自治区
雩都县改于都县	婺川县改务川县	酆都县改丰都县	和阗专区改和田专区
大庾县改大余县	鳚水县改习水县	石砫县改石柱县	和阗县改和田县
虔南县改全南县		越嶲县改越西县	于阗县改于田县
新淦县改新干县		呷洛县改甘洛县	婼羌县改若羌县

在《简化字总表》发布时，有一批地名用字因为生僻难认，经国务院批准进行了更改，表 6 列出了此次更改中涉及到的部分地名[6]。在进行简繁转换时，碰到这些地名应该进行相应的转换，这只需要将相应的地名对照关系添加到对照表中即可，没有什么困难的。

2.3.4.3 组织机构名

组织机构名的转换出现的问题是有些名称中包含了科技名词，例如“中国科学院软件研究所”、“中华人民共和国信息产业部”，分别包含了科技名词“软件”和“信息”，简繁转换中如果包含科技名词的转换，那么这两个机构名称将被转换为“中國科學院軟體研究所”和“中華人民共和國資訊產業部”。对于这样的结果，有些人认为是合理的，另一些人却认为组织机构名称在简繁转换时应该保持原貌，应该逐字进行转换，包含的科技术语不应该转换。这样的话，简繁转换中还应该包含对组织机构名的识别，以避免将组织机构名中的科技术语也转换掉。

2.3.5 外国命名实体

表 7 海峡两岸对外国命名实体音译的差异示例

大陆	台湾	大陆	台湾
厄立特里亚	厄利垂亞	布什	布希
卢旺达	盧安達	克林顿	柯林頓
冈比亚	甘比亞	萨达姆	海珊
利比里亚	賴比瑞亞	戴安娜	黛安娜

外国人名、地名、机构名等在翻译为汉字时一般是采用音译的方式，同科技名词的情况类似，海峡两岸对外国命名实体的音译也是各自独立进行的，造成了对同一个事物有不同的称呼，而且这些简繁两种叫法不存在单字的一一对应关系。表 7 中列出了海峡两岸在外国地名、人名音译方面的部分示例。在做简繁转换的时候应该考虑这些差异，只要将这些对照关系加入到词语转换的词表中即可。

有意思的是国外的一些机构名，例如“佐治亚软件研究所”，不仅包含了音译还包含了科技名词，如果是海峡两岸分别翻译的话，肯定是一个翻译为“佐治亚软件研究所”，另一个翻译为“喬治亞軟體研究所”，那么简繁转换时就应该考虑将外国的机构名中的科技名词做转换[9]。

2.4 其它问题

2.4.1 汉语分词

词是最小的能够独立活动的有意义的语言成分，然而，汉语文本中词与词之间却没有明确的分隔标记，而是连续的汉字串。显而易见，自动识别词边界，将汉字串切分为正确的词串的汉语分词问题无疑是实现中文信息处理的各项任务的首要问题。目前比较优秀的分词系统的召回率都能达到 90%以上，但是对于未登录词的识别效果还不尽如人意[14]。

汉字简繁转换一般是需要使用更大的语言单位来消除较小的语言单位转换时的歧义，多目标字需要通过词语转换来消除转换时的歧义，而词语转换也需要更长的词语转换来消除歧义。然而规范的汉语分词将句子中的词语分裂开，导致简繁转换时不能充分利用上下文的信息来消除转换歧义。例如句子“上演了一出好戏”的分词的结果是“上演/了/一/出/好/戏”，这里将“出”字被分成了独立的单字词，不能确定该转换为哪一个目标字。如果不分词，便可以利用前面的数词“一”的信息来正确地转换为“齣”。然而是不是碰到“一出”就一定转换为“一齣”呢？也不尽然，例如“太阳一出来”中的“一出”显然不能转换为“一齣”！所以词表里肯定还要有“一出来”这个更长的词，这样才能消除“一出”这个词转换时的歧义[7,8]。

规范的汉语分词除了割裂了词语之间的隐含信息外[14]，其对切分歧义的处理结果往往会出错，而简繁转换中容易出错的地方也是往往切分歧义的地方。

所以，规范的汉语分词可能导致简繁转换正确率的下降！

2.4.2 单字转换与词语转换的协作

对于单字转换的多目标字，一般是根据使用频率来选择一个目标汉字，在引入了词语转换之后，应该考虑在去除了可以使用词语转换消歧的所有情况以外，各个目标字的使用频率的高低。

有部分多目标字，其两个目标字中的一个，只在个别的几个词语中出现，那么在词语转换中加入了这几个词语以后，就可以把这个字当成只有一个目标字的汉字，表 8 列出了部分示例，所列简体字在转换为繁体字的时候都可以保留原字不转（因为另一个目标字和原字是同一个字）。这样的情况可以推而广之。

表 8 有一个目标字只在个别词语出现的汉字示例

例字（简体）	目标字（繁体）	说明
据	据、據	“据”只用于“拮据”一词
家	家、傢	“傢”只用于“傢俱”、“傢俬”等少数词
卜	卜、蔔	“蔔”只用于“萝蔔”一词
板	板、闆	“闆”只用于“老闆”一词

单字转换与词语转换的协作关系会影响到单字转换中多目标字的范围，一个多目标字在单字转换的时候是否需要按多目标字对待，还需要考虑它的构词情况，这需要有更多的语言学知识做基础。

3、 系统设计

3.1 系统主要技术特点

根据前述的分析，我们将在目标系统中采用的主要技术包括：分层次转换、词语消歧、面向简繁转换的分词和词典查找、命名实体识别、转换正确性评估等，下面分别说明。

3.1.1 分层次转换

分层次转换将整个转换过程分成术语转换、普通词语转换、单字转换三个层次，三个层次的优先级顺序是术语转换优先级最高，普通词语转换其次，单字转换的优先级最低。分层次的目的是用高优先级的转换来消除低优先级转换可能出现的歧义或者错误。术语转换和普通词语转换都可能查不到要转的词语，单字转换就是整个转换过程最后一道保障线，因为每个字它总是有目标字的，目标字可能是别的字，也可能就是这个字本身。

3.1.2 词语消歧

词语消歧就是利用词语转换来消除单字转换的歧义，利用长词转换消除短词转换的歧义或者错误。分层次转换也体现了词语消歧的特点，在术语转换和普通词语转换两个层次，也都要利用长词来消除短词转换的歧义。例如在术语转换的层次，“信息”在繁体里面有两个目标词“信息”、“資訊”，但是“信息产业”对应“資訊產業”，需要“信息产业”这个长词来消除歧义。在普通词语转换层次，“一出”对应“一齣”，“一出来”对应“一出来”，这导致“一出”这个词有转换歧义，所以，需要用“一出来”这个长词来消除歧义。像“一出来”这样的例子，其本质上是一个多目标字，在一些句子中和前后字都能成词，但是对应的目标字却不相同，从而导致了转换时的歧义，这种情况就需要用长词来消除歧义。

3.1.3 面向简繁转换的分词

在术语转换和普通词语转换两个层次，都需要先分词，然后查找目标词进行转换。如果切分出来的词语在单字转换的层次就能够正确转换，那切分这个词就是在做无用功，所以分词采用的词表应该和词语转换所用的词表一致。由于分词也包含查词的过程，如果能够将分词与查找目标词的过程合而为一那就更好了。我们的系统利用双数组 Trie 树这种数据结构及搜索算法来实现分词和查找目标词的统一。关于双数组 Trie 树的相关资料请参考文献[15-19]。

3.1.4 命名实体识别

这里的命名实体识别用于避免将命名实体中的术语按术语转换。对于涉及命名实体的其它问题，根据不同情况采用如下方式处理：

- 将特殊的人名地名加入到普通词语转换的词库中，实现正确转换，普通人名中的多目标字在单字转换层次按使用频率选目标字转换，不作特殊处理；
- 将外来音译命名实体词语加入到普通词语转换的词库中，实现正确转换；
- 将国外组织机构名称中的科技名词按等同国内命名实体对待。

3.1.5 转换正确性评估

根据前面的分析，简繁转换中容易出现错误的情况有如下几种：

- 有多目标字的单字转换：单字转换没有利用上下文相关的信息，而只是预先根据使用频率从多个目标字中选出一个作为实际转换的目标字，是很容易出现转换错误的；
- 术语转换：术语转换中有些科技名词本身在不同的上下文中就有不同的目标词，另外有些术语也可以作为通用词语使用，这也导致转换结果的不正确；
- 单字术语转换：虽然在术语转换中加入了命名实体识别的机制，但是命名实体识别并不能保证100%的正确率，而且除了命名实体以外，可能还有其他的场合不该把这类字当成术语来转换；
- 有多个目标词的通用词语转换：通用词语转换中有些词根据不同的上下文也有不同的目标词，所以转换结果也可能出现错误；
- 由于系统实现因素引起不准确性：这个主要是因为命名实体识别的不准确性导致的简繁转换可能出现错误。

在我们的系统中，转换正确性评估根据上面的分析，对每次字词转换进行评估，将转换结果分为如下三级：

- 一级：包括普通词语转换和单目标字转换，这部分转换一般不会出错；
- 二级：包括普通的术语转换，根据上下文的不同，术语转换可能会出现错误；
- 三级：上述可能出错的情况中不包含在一级、二级的其他情况，这部分出错的可能性比较大。

转换正确性的三级，分别以不同格式在转换结果中做出标记，从而将转换正确性评估的结果反映在系统界面上，提示用户在校对时特别注意。

3.2 简繁转换流程

图 2 给出简体转换为繁体的流程图，繁体转换为简体的流程与此类似。

下面我们以“刘汇丹在中国科学院软件研究所食堂吃了一碗面”这个句子为例，详细介绍将简体转为繁体的整个转换过程。假设系统被设置为要转换科技名词，但是不转换命名实体中的科技名词。每一步转换都要进行转换正确性评估，各步不再重复说明。

1. 首先，当前位置在句子首字“刘”，经术语转换和普通词语转换都没有发现当前字可以和后面的字成词，然后发现这个字是单目标字，转换为“劉”，当前位置移至“汇”字；
2. 经术语转换和普通词语转换都没有发现当前字可以和后面的字成词，然后发现这个字为多目标字，最终在单字转换时选择了目标字“匯”，当前位置移至“丹”字；
3. 从“丹”到“院”几个字都是单目标字，按单字转换。当前位置移至“软”字；
4. 经术语转换发现“软件”是科技名词库中的词语，触发命名实体识别，识别出“中国科学院软件研究所”是一个机构名称，所以不进行术语转换。经词语转换发现“软”和后面的字不成词*，最终“软”被当作普通单字转换为“軟”；

---

* “软件”这个词中不包含多目标字，未将其收入词语转换对照表中，所以系统不将其作为一个词处理。

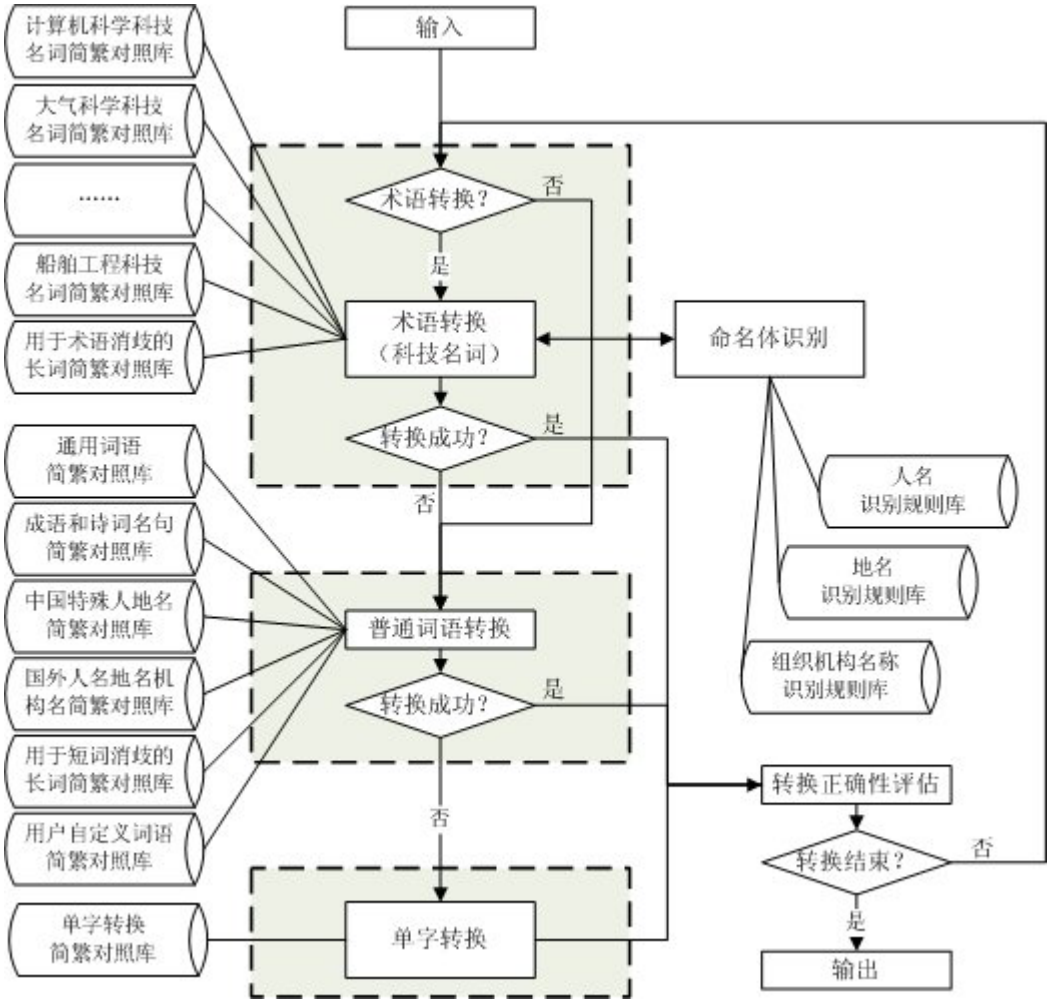

图 2.汉字简繁转换流程图

5. 从“件”到“碗”都是只有一个目标字的单字，按普通单字进行转换。当前位置移至“面”；
6. 经术语转换和词语转换匹配不成功，发现“面”有多个目标字，尝试进行基于搭配关系的单字转换，匹配成功（向前搜索到了“吃”字），转换为“麵”。已经到达结尾，转换结束；
7. 输出全句转换结果“劉匯丹在中國科學院軟件研究所食堂吃了一碗麵”。

4、 系统实现

根据以上的分析与设计，我们实现了此系统。目前可用于简繁转换的权威的语言资源还不是很多，我们从北京师范大学王宁老师处获得了一份词表，还从网络上搜集了一些资料，经过处理之后用在了系统中。但是由于种种原因，图 2 中所列各个语言资源库还未搜集到，另外有些资源需要在软件实际应用过程中慢慢积累。

下面是此系统的界面以及转换正确性评估的效果。这里考虑到打印时的效果，我们将一级结果正常显示，二级结果显示为粗体，三级结果显示为粗斜体，如下图所示。

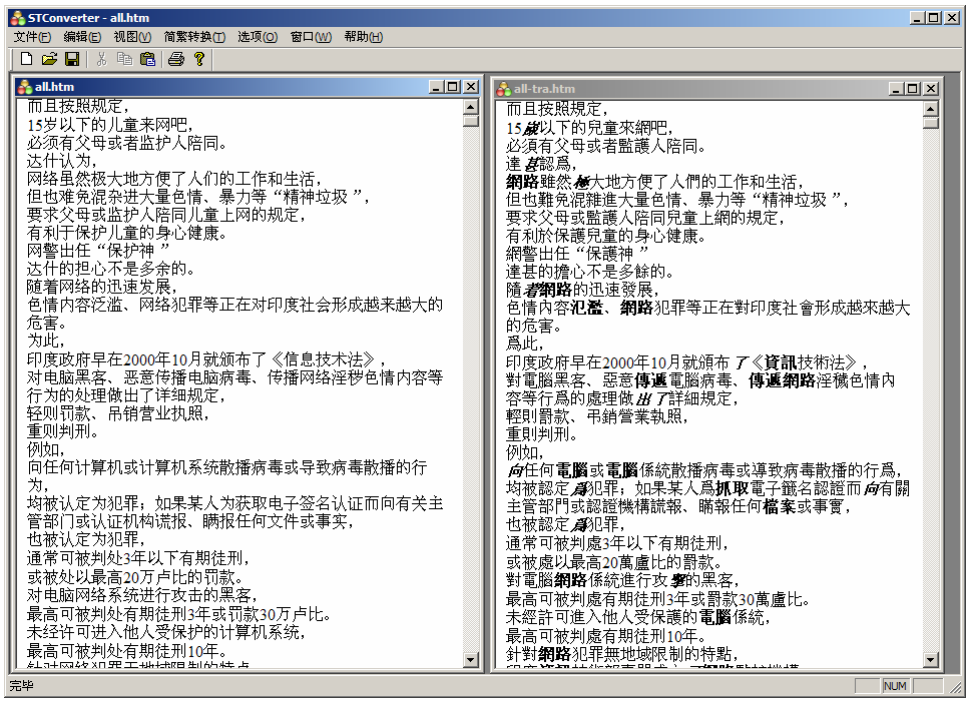

图 3 简繁转换分级显示效果图

5、 结束语

本文从编码字符集、单字转换、词语转换等各个方面分析了汉字简繁转换的复杂性，提出了一个基于词语消歧的分层次汉字简繁转换系统的涉及方案，并据此实现了一个汉字简繁转换系统。在与本文相关的工作中，我们深刻地认识到，要做出一个好的汉字简繁转换系统，一方面要有足够的语言学知识的积累，另一方面也需要在软件实现中考虑后续校对工作的方便。就目前的情况来看，还需要相关的专家学者以及计算机软件技术人员在相关领域做大量的工作。

参 考 文 献

[1] GB2312-80.信息交换用汉字编码字符集 基本集[S].1980.

[2] The Unicode Consortium. The Unicode Standard 4.0[S].2004.

[3] 国家语言资源监测与研究中心.中国语言生活状况报告 2005(上编)[M].北京：商务印书馆，2006.

[4] 国家语言资源监测与研究中心.中国语言生活状况报告 2005(下编)[M].北京：商务印书馆，2006.

[5] 现代汉语频率词典[M].北京：北京语言学院出版社,1986.

[6] 中国文字改革委员会.简化字总表[M].北京:语文出版社,1986.

[7] 王寧,王曉明.兩岸四地漢字的轉換與溝通[A].第三届两岸四地中文数字化合作论坛[C].2005.

[8] 王宁.基于简繁汉字转换的平行词语库建设原则[A].第四届两岸四地中文数字化合作论坛[C].2007.

[9] Jack Halpern,Jouni Kerman.The Pitfalls and Complexities of Chinese to Chinese Conversion[A].Fourteenth International Unicode Conference in Boston. 1999.

[10] 程祥徽.港澳用语用字问题[J].语文建设,2001,(05).

[11] 张书岩.简繁、正异字辨析(四)[J].语文建设,1995,(04).

[12] 龙城顺.有的“於”不能简化作“于”[J].语文建设,2001,(04).

[13] 苏培成.“发”字的尴尬[J].语文建设,2001,(12).

[14] 刘开瑛.中文文本自动分词和标注[M].北京：商务印书馆,2000.

[15] Aoe,J.An Efficient Digital Search Algorithm by Using a Double—Array Structure[J].IEEE Transactions on Software Engineering.1989,(09).

[16] Jun-Ichi Aoe,Katsushi Morimoto,Takashi Sato.An Eficient Implementation of Trie Structures[J]. Software-Practice and Experience,1992,22(09).

[17] KATSUSHI MORIMOTO,HIROKAZU IRIGUCHI,JUN-ICHI AOE.A method of compressing Trie structures[J]. Software-Practice and Experience.1994,24(03).

[18] 李江波,周强,陈祖舜.汉语词典快速查询算法研究[J].中文信息学报,2006,20(05).

[19] 王思力,张华平,王斌.双数组 Trie 树算法优化及其应用研究[J].中文信息学报,2006,20(05).