文章编号：

基于对照表以及语义相关性之简繁汉字转换*

庞祯军 姚天昉

（上海交通大学计算机科学与工程系，上海 200240）

摘要：目前使用的汉字有简体和繁体两大形式：中国大陆和新加坡等地使用简体字，我国港澳台地区和部分海外华人社区使用繁体字。其中大多数简体字的意义和用法与对应的繁体字是一样的，具有一一对应关系，这种情况通过查找简繁对照表就可以正确处理。然而，还有相当一部分简体字对应多个繁体字，这是简繁字转换的重点难点。基于此背景我们提出基于对照表以及语义相关性的简繁汉字转换方法。在教育部语信司及中国中文信息学会联合举办的一对多简繁转换评测中，我们的一对多简繁转换系统以 95.6%的准确率排名第一。

关键词：简体字；繁体字；简繁体字转换；一对多简繁转换

中图分类号：TP391　　**文献标识码**：A

Chinese Characters Conversion System based on Lookup Table and Statistical methods

Pang Zhen-jun, YAO Tian-fang

(Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai 200240)

Abstract: There are currently two forms of Chinese characters: Mainland China and Singapore use simplified characters; Part of Hong Kong, Macao and Taiwan regions and overseas Chinese communities use traditional characters. Most of the meaning and usage of simplified and traditional Chinese characters are the same. In this situation, the conversion between them can be processed correctly through trans-coding. However, there are a considerable simplified characters which can be transformed to many Traditional characters, which is the key and difficulty of Simplified and Traditional font conversion. Based on this background, we propose a method based on Lookup Table and statistical method. In the evaluation of conversion between simplified and traditional Chinese characters , our system ranked first at a accuracy 95.6%.

Keywords: Chinese character ;conversion between simplified and traditional Chinese characters.

1　引言

目前使用的汉字有简体和繁体两大形式：中国大陆和新加坡等地使用简体字，我国港澳台地区和部分海外华人社区使用繁体字。随着两岸三地的交流越来越频繁，简繁体字给交流带来了不便利。简繁转换技术对汉字文化圈交流起到重要作用，广泛应用于新闻出版、文化教育、古籍数字化处理等领域。

简化字总表中共收 2236 个字，其中大多数简体字的意义和用法与对应的繁体字是一样的，具有一一对应关系，这种情况通过编码转换就可以正确处理。然而，还有约 156 个简体字对应多个繁体字，例如简化字“干”对应四个不同的繁体字“幹”“干”“乾”“榦”。

*收稿日期：2013-07-07　　定稿日期：2013-07-15

作者简介：庞祯军（1987—），男，硕士生，主要研究方向为意见抽取，信息抽取，自然语言处理，pzj_636484@163.com;姚天昉（1957—），男，博士，副教授，硕导，主要研究方向为意见挖掘、信息抽取、机器学习、自然语言处理等。

一对多简体字的转换是汉字简繁转换的重点和难点。一对多简体字只有通过对文本进行语法和语义分析，利用语句甚至篇章的上下文语境才能将其正确转换为对应的繁体字[1]。因此，一对多简繁汉字转换是一个值得研究的课题，对汉字简繁转换性能起到至关重要的作用。

两岸三地现已有不少机构在进行简繁字转换的研究，如：中国科学院软件研究所，四通利方资讯有限公司，新天地公司，IBM 公司，倚天资讯股份公司及其他研发团队等。目前也有不少软件内嵌有简繁字转换功能，如 Microsoft Office, Sun 的 OpenOffice；同时在网络上也有不少的简繁体字转换工具，如谷歌翻译，快典网提供的简繁体转换功能等。但是现有的这些方法在处理一对多转换字时，准确率不高，仍需要人工来校正[2];简而言之，因为存在一对多简体字，使其在搭配不同的词或即使是同一词组在不同的语境下所对应的繁体字都不一样，例如：简体词组“晒干”的“干”字对应繁体字“乾”，而“树干”的“干”字对应繁体字“幹”；词组“下面”在表方位时“面”字对应繁体字“面”，而当其为动宾短语和面食有关时，对应繁体字的“麵”字，含义为煮面时把面放到锅里[3]。另外简繁汉字的字形差异，固定用语和地名人名等因素也会影响转换的准确率[4]。当前的简繁汉字转换领域基本还是以理论分析研究为主，而本文提供了一个切实可行且高准确率的方法及系统。

本文论述了基于对照表以及语义相关性的简繁汉字转换方法。对于繁体转简体(忽略极少部分的一个繁体字对应多个简体字的情况)以及只存在一对一关系的简体字，采用直接查对照表的方式就可以解决，因此本文的重点在于论述如何处理存在一对多关系的简体字之转换，本文所描述的一对多简繁转换方法在教育部语信司及中国中文信息学会联合举办的一对多简繁体字转换评测活动中以准确率 95.6%排名第一。本文将按照顺序描述简繁转换系统的工作流程，系统构造方法以及与其他转换系统之间的准确率对比，最后分析系统未正确转换的一对多简体字之出错原因。

2 系统描述及方法

2.1 系统工作流程

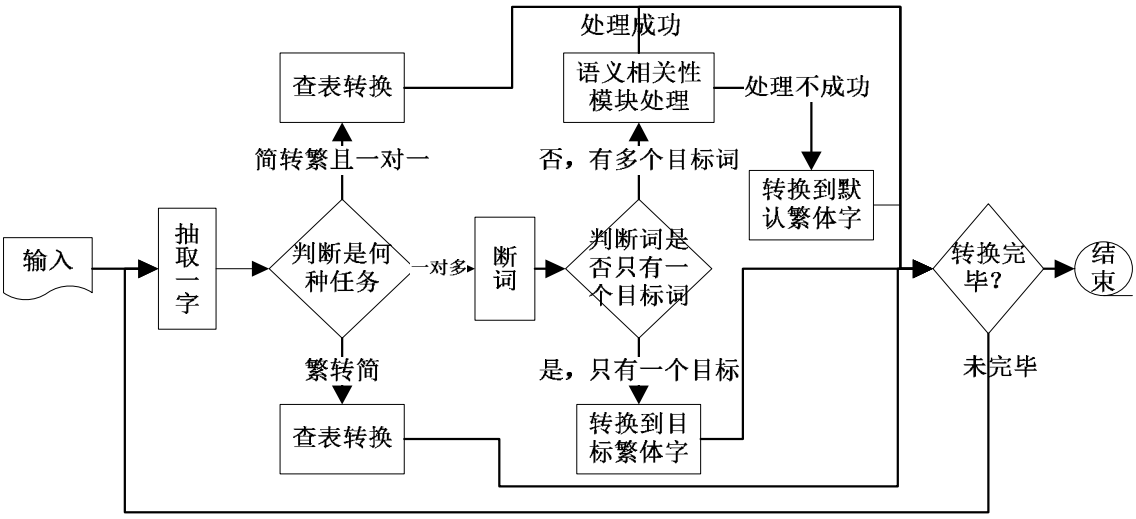
系统包括：繁体转简体对照表，一对一简体转繁体对照表，断词模块以及语义相关性模块等。对于输入系统的文本，按照阅读顺序对文本中的简体字一一转换。处理每个字的流程：

1.若为繁体字转简体字任务，则查找繁体转简体对照表直接转换；若为简体字转繁体字且该字为一对一简体字，则查一对一简体转繁体对照表直接转换。否则转 2.

2.启动断词模块进行断词，若截取的词组在任何语境下只有一个目标词组，则将该简体字转换到对应的目标繁体字；否则转 3.

3.启动语义相关性模块，根据统计信息进行简体字的转换，若根据统计信息能够正确转换简体字则输出对应的繁体字；否则转 4.

4.将简体字转换为系统设置的默认繁体字。



图一 系统工作流程图

虽然也存在一个繁体字对应多个简体字的情况，但这种情况少之又少，现有的简繁转换系统基本都将繁体字转简体字的任务处理为一对一转换。

2.2 系统构建

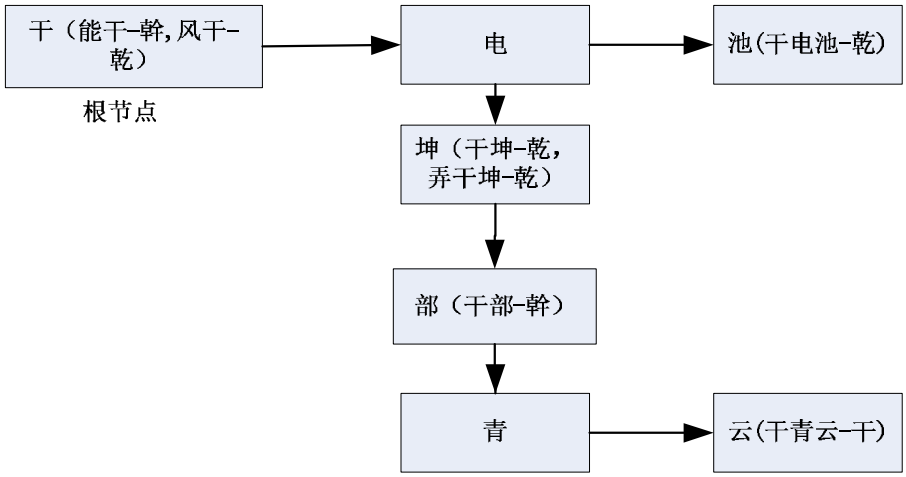
2.2.1 构建对照表

需要构建的对照表有繁体字转简体对照表和一对一简体字转繁体字对照表。此部分的繁体字与简体字的对应关系主要借鉴了北大中文系李铎博士所制作的<<简繁字对应表>>，在滤去<<简繁字对应表>>中的部分错误后构建了繁体字转简体对照表和一对一简转繁对照表。

2.2.2 构建一对多简体字断词模块

在处理一对多简体字时，根据简体字所在的词组去把简体字转换到对应的繁体字是比较有效率的处理方式。如可直接将词组“晒干”的“干”字转换到目标繁体字“乾”。此时，我们需要在系统中引入断词模块，去匹配句子中包含当前待转换字的最优目标词。如句子“这树干很粗”，在转换到“干”字时，应匹配词“树干”并将“干”字转换到繁体字“幹”。

为了提高系统断词的效率，我们需要对每个一对多简体字构造一个包含此一对多简体字的词组转换树；同时保存每个一对多简体字的最长词组长度以减少字符串比较次数，以提高匹配时的效率。以“干”字为例，所构造的树部分如下：



图二 词组转换树

树中每个节点保存两个数据：节点关键字和以根到本节点之路径中 节点关键字序列 结尾的词组转换列表。以图二所示数据为例：最左边的节点表示为树的根，以干结尾的词组有“能干”，“风干”等，其中词组“能干”的“干”字转换为目标繁体字“幹”，“风干”中的干字转换为目标繁体字“乾”。第三列的“池”这个节点，从根节点到当前节点的关键字序列为“干电池”，此节点保存以“干电池”结尾的词组之转换列表。同时在简体字的词组转换树中，每个词组需要保存简体字在词组中的位置，因为待转换的简体字可能在词组中出现多次，如一对多简体字“么”在词组“么么小丑”中出现了两次且该词组对应的繁体词组为“么麼小丑”，故应指明待转换字是词组中的位置。

同一个简体字的某个词组可能会包含另一个词组，而在这两个词组中待转换简体字的目标繁体字可能会不一致，因此在进行断词时应采用最长匹配。如词组“辟地”中的“辟”字应转换为繁体字的“闢”，而词组“其次辟地”中的“辟”字应转换为繁体字的“辟”。

表一 最长匹配示例

待转换词组	辟地	其次辟地
目标繁体词组	闢地	其次辟地

但是即使采用最长匹配也会出现问题，如待转换句子“在外面糊了纸”，在处理“面”字时可匹配词“外面”和“面糊”，且这两个词的目标繁体字不一致，见表二。

表二 匹配冲突示例

待转换词组	外面	面糊
目标繁体词组	外面	麵糊
权值	7	5

本文采用对每个词先置权值的方式来处理这种情况：对常见的词赋予一个高的权值，越常见值越高；若还是有冲突，则靠前匹配。在本例中，词“外面”比“面糊”的权值大，故应取匹配词为“外面”；假如“外面”的权值和“面糊”的权值相同，则根据靠前匹配规则还是应选取匹配词“外面”。

在断词后，若所匹配的词中的简体字只有一个目标繁体字，则直接转换到该繁体字。否则将进入语义相关性模块进行处理。

2.2.3 构造语义相关性模块

在断词模块只匹配到待转换字或者所匹配到的词存在多个目标繁体字时，需要启动语义相关性模块来转换待转换的简体字。

语义相关性模块是根据每个一对多简体字处于何种上下文环境来判定待转换字的语义，而将待转换字转换到目标繁体字。如对于待转换字“干”，如果句子中出现六十四卦中的“坤，屯，蒙，师，讼”等，则应将其转换为“乾”。使用此方式需要为一对多简体字的每个目标繁体字建立语义相关性信息，在进行转换时统计每个目标繁体字在上下文环境下所得到的加权分数，并取得分最高者作为最终的目标繁体字。例如：

待转换简体字：干

目标繁体字的部分语义相关性词条：

乾：八卦、六十四卦、坤、震、巽、坎、离、艮、兑、讼……

干：矛、盾、戈、河、江….

待转换句子：“八卦为：干、坤、震、巽、坎、离、艮、兑。”

加权结果：若每个被匹配的词分数为 1，则目标 繁体字“乾”总得分为 8，其他目标繁体字得分均为 0，则应将“干”字转换为繁体字“乾”。

编写语义相关性词条，我们可以参考繁体字的相关资料，作者参考了<<臺灣國語詞典>>，这样不需要对繁体字有很深的造诣也能够编写出质量不错的语义相关性词条。同时，应选择那些具有大区分度的词作为候选词组，这样才能保证转换的准确度。

3 实验测试之简繁转换评测

在系统构建成功后，共处理简体字 2775 个，繁体字 3358 个；一对多简体字 156 个，一对多词组转换树包含词组 34207 组，词组的获取方式主要有两种：从<<臺灣國語詞典>>获取和维基百科的简繁转换一对多对应表获取，但维基百科的词组有不少错误需要人工滤过处理。

3.1 一对多简繁转换评测

实现了本文方法的一对多简繁转换系统参加了教育部语信司和中国中文信息学会联合举办的简繁转换评测活动，以一对多简繁转换准确率 95.6%在评测中排名第一。

评测活动(任务一) 针对一对多简繁汉字的转换，对给定数据集（以 utf-8 方式编码，全部为包含有一对多简体字的句子），要求参加评测的系统给出句子中指定简体字的目标繁体字。本次评测活动共包含一对多简体字 135 个，测试数据 76540 条；测试数据包括文言文，诗歌，现代小说等各种形式的文字。且给定测试数据格式为：

句子 1：他对这件事毫无<待转换字>干<待转换字>劲

句子 2：是用<待转换字>干<待转换字>冰和金属录制的

评价结果指标：

一对多简化字转换准确率 = 评价语料中转换正确的字数目 / 评价语料中待转换的字数目

本次共有 11 个单位或个人参加了一对多简繁字的转换评测，性能指标结果如表三（系统 G 为我们的系统）：

表三 系统性能

系统代号	性能（转换准确率）
系统 C、***系统 G***	0.956
系统 I	0.926
系统 H	0.916
系统 J	0.893
系统 E	0.880
系统 A	0.874
系统 F	0.845
系统 D	0.833
系统 K	0.808
系统 B	0.805

通过本次评测活动表明，我们的方法是行之有效的。实现了本文方法的简繁转换系统也是处于相对较好层次的。

3.2 结果分析

根据评测组委会发布的简繁一对多转换的参考答案，我们查看了转换出错的测试数据，统计出出错主要有以下几类：

1.古人名，固定用语转换出错：

表四 固定用语出错示例

待转换字	数据	正确目标	实际结果
干	曾运***于***正读：“罪重者比于上刑，罪轻者比于下刑也	乾	幹
台	元薩都剌《钓***台***夜兴》诗：“仙茶旋煮桐江水，坐客遥分石壁灯	臺	台
台	《明史·刘***台***传》：“瀚（张瀚）生平无善状……官缺必请命居正，所指授者，非楚人亲戚知识，则亲戚所援引也；非宦楚受恩私故，则恩故之党助也	臺	台
周	《语文·***周***语上》：“昔昭王娶于房，曰房后	周	週
范	***范***望注：“哜哜，忧悲	范	範
钟	***钟***广言注：“喇子，又名红宝石，色红，透明	鍾	鐘

此类问题解决可通过增加词库词条来解决一部分，同时在语义相关性中为每个词增加合适的相关性词汇以期能够解决更多的问题。但因存在很少一部分繁体字可以对应多个简体字的情况，如此表中的第一条数据中的“曾运干”，因其是人名而又存在简体字“乾”，故此测试数据之正确性值得商榷。

2.有多个目标繁体词组的情况处理出错。

表五 词组多目标转换出错示例

待转换字	待转换词组	可选目标	目标词含义	组合示例
干	不干	不干	不愿意，不罢休	我不干，你凭什么这样

		不乾	有水的，湿的	桌面不干，有水
面	挂面	掛麵	面的一种	煮挂面来吃
		掛面	与这人不熟悉但见过	我和他挂面认识
面	白面	白麵	面粉，用来吃的	你磨的白面真好！
		白面	白净，多形容读书人	你真是白面书生！

此类词组是少数，通过完善语义相关性模块中的词条即可达到正确的处理效果。如简体词组“复姓”的语义相关性词条可设置为“复姓-2-復、原来$複、诸葛、司马、司徒、欧阳、令狐”。其中繁体词组“復姓”表示恢复原来的姓氏，繁体词组“複姓”指由两个及以上（以两个为主）汉字的组成的姓氏。

3.断词错误

表四 断词错误示例

待转换字	数据	正确断词	实际断词
干	元高文秀《黑旋风》第一折：“我和你待摆手去横行，管教他抹着我的无***干***净	***干***净	无***干***
干	唐杜甫《茅屋为秋风所破歌》：“床头屋漏无***干***处，雨脚如麻未断绝	***干***处	无***干***
面	八角或六角形的灯，每***面***糊绢或镶玻璃，并画有彩色图画，下面悬挂流苏	每***面***	***面***糊

第一个例子中的“无干”和“干净”均有保存，但应采取的是权值比较故系统选择匹配词“无干”而导致错误；第二个例子中的词组“干处”未保存在词库中故出错；第三个例子中的词组“每面”未保存在词库中故出错；

通过分析出错原因可知，若碰到词组匹配冲突可采取权值比较及语义相关性分析综合的方式来解决此类问题。即在出现断词冲突时，可通过语义相关性模块对待转换字的各个目标繁体字进行评分，再给综合断词结果所对应目标繁体字进行一定的加分，最后通过最终的评分来选择目标繁体字。

4.语义相关性处理不够好

语义相关性词条的添加是一项繁琐，需要仔细思考的工作，需要我们要有一定的繁体字功底；关键是选取足够的具有强区分性的词组。这项工作若有相关专业人士的帮助将会事半功倍。

从维基百科收集的部分数据存在错误，通过这一次评测可以将相关词组纠正过来；在后续系统的运行过程中会发现其错误并纠正绝大部分问题。此类问题的存在是因为当前互联网上简繁转换的资料良莠不齐，只有从正规机构得到的资料才会有高的可信性。

4 总结与展望

虽然在过去的十余年里，很多研发机构和公司都在努力研发实用化的简繁转换工具，但是目前还没有一个真正的精密转换系统被研发出来，同样的内容，不同的工具转换出不同的结果来，导致结果的不可信。究其原因，解决问题的方法没有找准，长期以来都是企图依靠纯技术解决问题，而缺乏对文字学的研究导致部分问题迟迟无法解决。

本文主要研究改善传统的简繁转换只考虑一对一直接转换，而对一对多简体字转繁体字考虑不全面导致一对多简体字转繁体字无法有效地执行。因此我们提出了基于对照表和语义相关性的简繁字转换方法，通过评测组委会给出的性能评价结果表明我们的方法是行之有效的。在具体实现系统的过程中还需要对一对多词库的完善和语义相关性规则的完善，这是实现我们方法的重点难点。

语义相关性的完善需要我们对繁体字有一个较为深入的研究，选取足够的具有强区分性的词组。需要对 156 个一对多简体字建立完整的语义相关性资料，这本身就是一个具有挑战性的工作。语义相关性是简繁转换中计算机技术和文字学研究的良好结合点。同时本文描述的方法是一个开放性的方法，在对一对多词库和语义相关性词条进行逐步完善的过程中，将会越来越圆满地解决简繁转换中的瓶颈问题。

参考文献：

[1]教育部语信司，中国中文信息学会. 简繁汉字智能转换评测大纲

[2]王曉明, 魏林梅, "談簡繁轉換的幾個關鍵問題", 5TH CDF 研討會數位社群雙效（CD2E）, 2008年12月24.

[3]李樹德, "Word“中文簡繁轉換”存在的問題與解決對策", http://www.yywzw.com/show.aspx?id=1570&cid=142.

[4]刘汇丹 吴健, "基于词语消歧的分层次汉字简繁转换系统", 5TH CDF 研讨会数位社群双效(CD2E), 2008年12月24日.

[5] <<臺灣國語詞典>>.

[6]谷歌翻译. http://translate.google.com.hk/#zh-CN/en/.

[7]李民祥，楊秉哲.基於對照表以及語言模型之簡繁字體轉換[J].台湾:朝陽科技大學資訊工程系，2011.

[8]快典网. http://ft.kdd.cc/.

[9]王寧, 王曉明, "兩岸四地漢字的轉換與溝通", 第三屆兩岸四地中文數位化合作論壇, 2005年10月.

[10] Martin Hepp, Katharina Siorpaes, Daniel Bachlechner, "Harvesting Wiki Consensus: Using Wikipedia Entries as Vocabulary for Knowledge Management," IEEE Internet Computing, vol. 11, no. 5, pp. 54-65, Sep./Oct. 2007.

[11]李铎. 简繁字对应表.北京.北京大学语言文学系.