计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.009713] [CSTR: 32024.14.csa.009713]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

具有标点区分纠错能力的中英盲文转换系统①

李泽芃, 罗远新, 孙家宁, 陈　洪, 郦　铖

(重庆大学 国家卓越工程师学院, 重庆 401135)
通信作者: 李泽芃, E-mail: 2713721325@qq.com

摘　要: 盲文转换技术可以推进盲人群体的信息无障碍化进程, 有着相当的重要性. 随着信息全球化发展, 盲人不可避免地接触到包含中英双语的信息. 现有中英盲文转换系统能较好完成中英文字符到盲文的转换, 但是对标点的转换效果不佳, 表现为对一符多用情况区分效果不好、未对中英标点混用情况进行纠错两方面. 若未能恰当处理一符多用及中英标点混用情况, 很可能会对盲人阅读造成影响甚至误解. 本文详细分析以上问题, 设计并实现一种具有标点区分和纠错能力的中英盲文转换系统. 基于 BCC 语料库构建测试语料, 对该系统开展一系列测试和评估. 实验结果表明, 与其他类型转换系统相比, 本系统能够结合语言类型和上下文结构, 有效区分标点转换中的一符多用情况, 并能纠正中英标点混用问题, 对我国信息无障碍化进程起到促进作用.

关键词: 信息无障碍; 盲文转换系统; 标点转换; 一符多用; 中英标点混用

引用格式: 李泽芃,罗远新,孙家宁,陈洪,郦铖.具有标点区分纠错能力的中英盲文转换系统.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9713.html

Chinese-English Braille Conversion System with Ability of Punctuation Distinction and Error Correction

LI Ze-Peng, LUO Yuan-Xin, SUN Jia-Ning, CHEN Hong, LI Cheng

(National Elite Institute of Engineering, Chongqing University, Chongqing 401135, China)

Abstract: Braille conversion technology is crucial for advancing information accessibility for the blind. With the rapid advancement of information globalization, the blind are increasingly exposed to bilingual information in both Chinese and English. While existing braille conversion systems have successfully translated Chinese and English into braille, they fall short in accurately converting punctuation, including poor differentiation of punctuation with multiple uses and lack of error correction for the mixed use of Chinese and English punctuation. Failure to address these issues may lead to misunderstanding of text by the blind. This study delves into these problems, designing and implementing a bilingual braille conversion system capable of distinguishing multipurpose punctuation and correcting the mixed use of punctuation. The performance of the system is evaluated by using a dataset based on BLCU Chinese Corpus. The results demonstrate that the proposed system accurately distinguishes multipurpose punctuation and corrects the mixed use of Chinese and English punctuation according to language types and context, outperforming other braille conversion systems. Overall, this research has significant potential for promoting information accessibility in China.

Key words: information accessibility; braille conversion system; punctuation conversion; multipurpose character; mixed use of Chinese and English punctuation

① 收稿时间: 2024-05-16; 修改时间: 2024-06-28; 采用时间: 2024-07-11; csa 在线出版时间: 2024-10-25

信息无障碍是指利用不断发展的信息技术手段, 让所有人无障碍地获取信息资源[1]. 根据第 6 次全国人口普查我国总人口数, 及第 2 次全国残疾人抽样调查推算, 至 2010 年末, 我国残疾人总人数为 8 502 万人, 其中视力残疾 1 263 万人, 占残疾人总数的约 15%[2]. 在我国信息无障碍的进程中, 如何解决盲人群体无障碍需求是一个关键问题.

使用听觉和触觉代替视觉是解决盲人群体无障碍需求的主要思路. 基于视觉识别、文本转语音技术的盲人阅读器采取语音交互方式, 让盲人用耳朵“听”来完成阅读的功能. 基于盲文转换技术的盲人阅读器采用触觉交互方式, 将文字和符号信息转换为盲文, 让盲人通过手指触摸的方式完成阅读. 相比于听觉方式, 触觉方式在阅读备忘录、表格、科技类的资料方面有着不可比拟的优势[3]. 同时, 触觉刺激在盲文学习中也起到重要作用[4]. 因此, 发展盲文转换技术有着重要意义.

近年来, 随着英语的全球化, 中英文混用的情况愈发普遍. 为了使盲人群体能够阅读中英文信息, 现有盲文转换技术大多在汉盲转换的基础上, 加入了对英文字母、单词、标点的处理. 目前常见的处理方式可分为 3 类.

一是符号对应转换, 即通过编写编码表, 将英文字母及标点映射至盲文. 2016 年, 吕先超[5]设计并实现了一种具有 3 层构架的汉盲转换系统模型 SunBraille. 该系统通过构建数字、字母、标点到盲文 ASCII 码对照库完成非汉语符号到盲文的转换. 2022 年, 胡庆玲等[6]开发一种国家通用盲文转换系统, 该系统通过 Unicode 编码区分中文字符串与其他字符串, 通过直接查找其他字符串对应的盲文 ASCII 码表, 得到盲文点序. 2023 年, 毛扬等[7]为提高低算力环境下汉盲转换的速度, 设计并实现一种汉盲转换 SoC 系统, 可实现汉盲转换与盲文显示功能. 该系统使用选择器实现非中文-盲文转换. 陈楷闻[8]设计了一种具有中英盲文转换功能的嵌入式盲文数字化系统, 该系统对于中文句子和英文句子采用不同翻译策略. 对于中文句子, 使用函数区分中文字符与非中文字符, 并通过建立映射表的方式完成非中文字符转换.

二是使用 Liblouis 实现盲文转换. Liblouis 是一款适用于多国语言的开源盲文转换系统[8]. 该系统采用基于规则和基于词典的转换方式, 通过读取不同语言转换表中的信息将文字和标点转换为盲文. 2023 年, 陈楷闻[8]基于 Liblouis 设计了嵌入式盲文数字化系统的英文翻译部分, 实现一种独立可扩展的英文转换模块. 林栋[9]对 Liblouis 进行分析, 调用其相关接口, 实现一种基于 Liblouis 的多语言盲文转换系统, 并应用于盲文学习机. 邹成洋[10]开发一款基于 Linux 的嵌入式盲文学习机. 该学习机软件部分使用 Liblouis 与 Jieba 分词组件, 可实现盲文数字化存储与表示.

三是借助人工智能技术进行转换, 即使用经过专家校核的盲文对照语料库训练模型, 学习字符与盲文的对照关系, 实现盲文转换. 由于盲文对照语料库存在中文、英文及标点字段, 因此训练后的模型既可以进行中文翻译, 也可以实现英文转换. 2019 年, 蔡佳等[11]首次将深度学习技术引入汉盲转换领域, 提出一种基于汉盲对照语料库和深度学习的汉盲转换方法. 2021 年, 蒋琪[12]基于 NMT 技术, 提出一种端到端的盲文转换算法, 通过建设句子级对照语料库, 将文字一步转换为盲文. 2023 年, 王蕊[13]在前人工作[12]的基础上加以改进, 提出一种增加预训练模型的特征提取的改进方法, 进一步提高了盲文转换的准确率.

以上研究可以较好地转换英文字母与单词, 其中 Liblouis 盲文转换系统与人工智能技术能够实现英文二级盲文的简写翻译, 提高了盲人的阅读效率. 但是, 上述研究对标点的转换效果不佳, 表现为以下两点.

一是对一符多用情况区分效果不好. 一符多用指在中英盲文转换中, 一些标点符号根据不同使用场景转换为不同盲文表示的现象. 例如, 方括号“[ ]”在中文盲文中表示为“ ⠦ ⠴ ”, 而在英文盲文中则表示为“ ⠨⠶ ⠶⠨ ”; 英文直引号“'”作为单引号使用时表示为“ ⠄⠦/⠴⠄ ”, 而作为所有格号或省写号时, 则表示为“⠄ ”. 中英盲文中常见一符多用情况如表 1 所示. 若未能根据具体使用场景对一符多用的标点进行恰当的盲文转换, 可能会对盲人读者造成困扰和误导. 以小数点为例, 若误将其转换为英文句号对应的盲文, 可能会使读者将小数误认成两句话的结尾与开头. 目前, 仅有少数研究提出点号“.”[5,9]、直单引号“'”及直双引号“"”[9]的区分, 对于其他一符多用类型未见讨论.

二是对于中英文标点混用情况未进行纠错. 文字输入和电子排版过程中, 由于用户对计算机键盘和输入法的使用习惯不同, 经常出现标点符号的误用[14], 如将中文逗号错误使用为英文逗号. 中英文标点混用并不会妨碍明眼人阅读, 但这种混用情况会对盲人阅读

造成影响, 原因可概括如下.

首先, 单方盲文能够表示的信息有限. 一方盲文由6点组成, 其排列方式仅有64种, 不同语言之间的盲文符号不可避免地出现重复. 其次, 为节省盲文篇幅, 提高阅读效率, 盲文会进行书写简化, 如国家通用盲文基于“按声母省写”原则省略音调、英文二级盲文用一方盲文表示多个字母等. 因此, 若输入语句存在中英标点混用的情况, 混用标点对应的盲文可能与某个盲文符号重复. 而又由于省写规则, 盲人无法从结构上判断该盲文是否为混用的标点, 从而造成理解错误.

表1 中英盲文中常见一符多用情况

序号	字符	含义A	盲文符号	含义B	盲文符号
1	[ ]	中文方括号	⠒ ⠒	英文方括号	⠨⠶⠆⠄
2	.	小数点	⠄	英文句号	⠲
3	'	左直单引号	⠄⠄	右直单引号	⠄⠄
4	'	直单引号	⠄⠄/⠄⠄	所有格号/省写号	⠄
5	"	左直双引号	⠄⠄	右直双引号	⠄⠄

以“需要加药, 避免病情恶化.”这句为例. “加药”的拼音为“jia1 yao4”, 对应盲文为“jia1 yao”(⠚⠡⠄ ⠄⠄). 其中“药”由于声母省写规则省略掉其4声标调. 若该句中文逗号误用为英文逗号, 则逗号的盲文表示就从5点(⠄)变为2点(⠠), 而2点恰好与表示中文2声含义的盲文重复. “加药, ”对应盲文变为“jia1 yao2”(⠚⠡⠄ ⠄⠄⠠), 与“佳肴”(⠚⠡⠄ ⠄⠄⠠)对应盲文相同. 此句可能会被误解为“需要佳肴避免病情恶化.”. 除逗号外, 大部分标点的混用都可能给盲人阅读带来障碍, 如表2所示. 由此可见, 在盲文转换时有必要对中英标点混用情况进行纠错. 但是, 现有研究在进行盲文转换时, 无论标点使用是否正确, 都会将其直接翻译为对应盲文, 而不做额外处理. 这种方式很可能会影响盲人阅读.

表2 混用情况下可能引起阅读歧义的英文标点

序号	英文符号	符号含义	中文对应
1	,	逗号	2声
2	.	句号	ong/weng
3	?	问号	ang
4	!	感叹号	ao
5	:	冒号	wen
6	;	分号	4声
7	—	破折号	冒号
8	" "	直双引号	ang en
9	' '	直单引号	拉丁大写/ang en/3声
10	( )	括号	wang
11	[ ]	方括号	拉丁大写/wang wang/3声

综上所述, 现有中英盲文转换系统存在标点一符多用情况区分不全、中英标点混用情况处理不当的问题. 盲文转换技术是解决盲人群体无障碍需求的重要途径, 意义重大. 需进一步探索合适的盲文转换方案加以解决.

本研究中, 我们设计并实现一种具有标点区分和纠错能力的盲文转换系统. 首先, 分析一符多用与标点混用问题, 提出一种具有4层结构的盲文转换系统框架. 然后, 借助Language-Detector语言检测库[15]、Jieba分词组件[16]等第三方库, 使用Java语言在该框架基础上实现了一种能够处理标点问题, 并具有国家通用盲文、英文一级盲文转换能力的中英盲文转换系统. 随后, 设计实验研究比较了不同转换系统(本文所述系统、符号对应转换系统、Liblouis盲文转换系统、中国盲文数字平台[17]) 对一符多用及标点混用情况的处理能力. 并从准确率、转换速度方面测试了本文所述系统的盲文转换性能. 最后, 总结评述本文研究并展望未来研究方向.

1 系统设计与实现

1.1 可行性分析

本文目标是开发一种具有标点区分和纠错能力的中英盲文转换系统. 目前, 汉盲转换、英盲转换技术较为成熟. 但是对于如何解决一符多用及标点混用问题, 还没有文献参考, 需要讨论其可行性.

详细分析一符多用与标点混用情况, 可以将其中具体问题分为两类. 一类问题和标点所处的语言相关, 称为语言类型问题; 另一类问题则和标点附近文字结构相关, 称为上下文结构问题. 一符多用情况的方括号和所有标点混用情况, 都属于语言类型问题. 这类问题可通过判断标点所在句的语言种类解决. 以逗号混用情况为例, “需要加药, 避免病情恶化.”这句中的逗号被误用为英文逗号, 直接转换成盲文会引起歧义. 若在进行盲文转换之前, 先检测此句的语言为中文, 即可确认该处标点有误, 并通过替换为正确的中文逗号解决. 其他一符多用情况则可归类于上下文结构问题, 可以通过设置判断规则的方式区分. 以点号为例, 小数点相邻的字符均为数字, 而英文句号前的字符是英文字母, 同时其后一般为空格. 因此, 通过编写合适的判断规则可辨别这两种使用情况.

1.2 系统总体框架

根据可行性分析, 若想将标点恰当转换为盲文. 不仅要明确该标点前后的文字结构, 还要明确标点所处文段的语言类型. 基于此, 本文设计一种由识别层、文字转换层、标点转换层及组织层构成的盲文转换系统框架, 如图 1 所示.

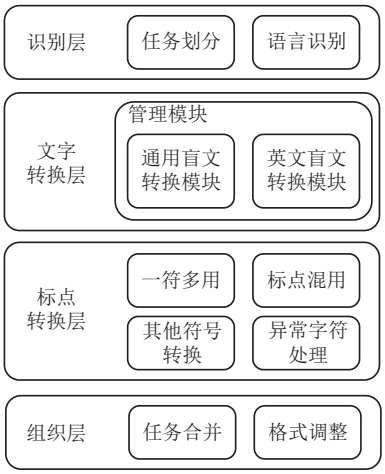


图 1　盲文转换系统框架

实际情况中, 用户输入文本可能为同时具有中英两种语言类型的混合文本. 由于不同语言中标点使用规范存在差异, 因此若没有准确识别输入内容各部分的语言类型, 则无法对标点进行区分和纠错. 针对这一问题, 本研究引入识别层对输入内容进行分析处理. 识别层的主要功能是任务划分和语言识别, 该层采用分治策略, 根据语言类型将文本切分为若干子任务. 通过在后续过程中对每个子任务进行独立处理, 实现输入内容的盲文转换.

考虑到标点转换会影响中文分词过程, 进而影响汉盲转换准确性, 本文优先进行文字转换. 文字转换层包括汉盲转换和英盲转换两个转换模块, 这两部分受管理模块控制, 协作将中英文字符转换为盲文.

鉴于转换的复杂性, 本文将标点转换功能独立设置为一层. 标点转换层一方面用于处理一符多用及标点混用情况, 根据子任务语言类型与相应判断规则, 对标点进行区分和纠错. 另一方面负责将不属于一符多用和标点混用情况的其他符号直接按规则转换成盲文. 此外, 标点转换层还负责处理系统中未收录的异常字符, 防止其影响盲文转换结果.

组织层的主要作用是整合来自不同转换模块的结果, 对格式进行调整, 并生成输出结果. 设计组织层的目的在于解决盲文转换过程中可能遇到的格式问题. 例如, 对于盲文文章, 标题应当居中书写, 而正文则应空两格进行书写. 通过设置组织层, 能够确保输出结果符合规范.

1.3 系统算法结构

基于以上框架, 本文使用 Java 语言编写程序, 并调用 Jieba 分词组件、Pinyin4j 拼音转换库和 Language-Detector 语言检测库. 实现了一种可处理标点问题的中英盲文转换系统. 程序框图如图 2 所示.

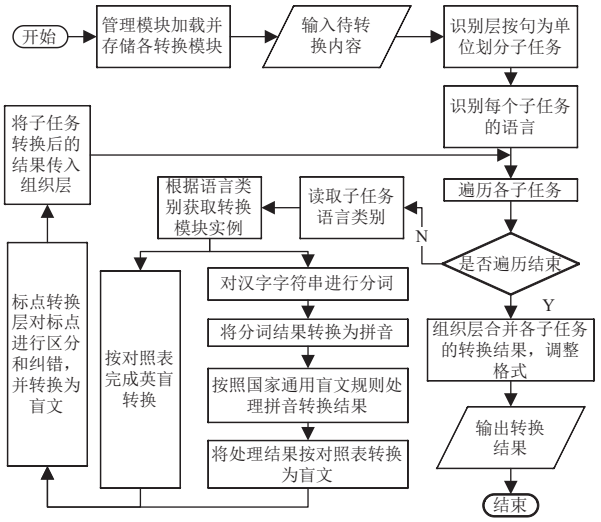


图 2　系统程序框图

系统获取待转换文本后, 由识别层将转换任务以句为单位划分成子任务, 并对每个子任务进行语言类型识别. 识别层根据识别结果将该子任务添加相应语言标记后, 将其传入文字转换层.

文字转换层的管理模块使用预先加载的方式优化盲文转换速度. 系统启动时, 管理模块使用 Java 反射技术获取并存储所有转换模块的实例对象. 在接收到子任务后, 管理模块获取中英转换模块的实例, 并调用其接口进行盲文转换. 文字转换完成后, 管理模块将结果传入标点转换层, 该层根据语言类型, 结合一符多用及标点混用情况的处理策略, 将标点和特殊符号转换为盲文.

转换后的盲文句子随即传入组织层. 组织层在合并各子任务转换结果、处理格式问题后, 将最终结果以字符串类型输出.

1.4 识别层实现

基于 Language-Detector 语言检测库, 本文实现了

识别层功能. 该层负责任务划分和语言识别, 图 3 展示了其工作流程.

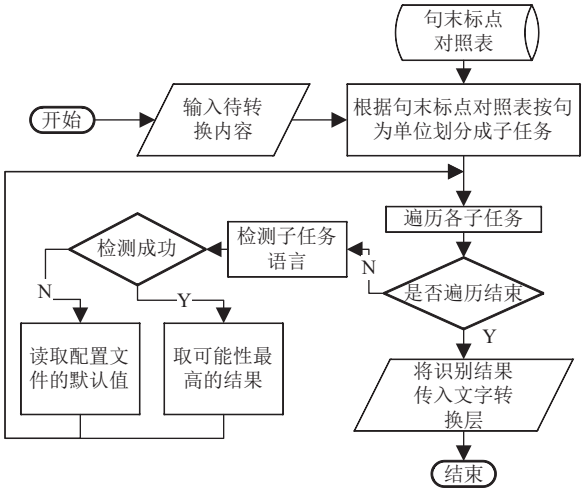


图 3 识别层工作流程

鉴于句子是语言运用的基本单位, 识别层以句为单位进行任务划分. 首先建立标点对照表, 其中包含常见的句末标点, 如句号、感叹号、问号和省略号等. 随后, 通过匹配对照表中的句末标点, 将输入文本按句子分割成多个子任务.

对于每个子任务, 识别层使用 Language-Detector 语言检测库确定其语言类型. 该检测库基于 N-gram 算法, 以概率方式表示该句语言的判断结果. 本文取可能性最高的语言类型作为本句的识别结果. 若未能识别出语言类型, 则识别层会读取配置文件中的默认值, 并将其指定为该句的语言类型.

1.5 中英盲文转换模块实现

本文使用多步法实现国家通用盲文转换功能, 程序流程如图 4 所示, 包括分词、拼音转换、通用盲文规则处理及盲文对照 4 个步骤. 首先, 使用 Jieba 分词组件对待转换句子进行中文分词, 并以列表形式存储分词结果. 随后, 调用 Pinyin4J 拼音转换库将分词结果转换为拼音. 由于盲文转换所用拼音与汉语拼音略有不同, 因此需对拼音转换结果进行调整, 例如将“y”修正为“i”, 将“w”修正为“u”. 然后, 对调整后的拼音按通用盲文标调规则及简写规则进行处理, 得到中文的通用盲文拼音. 最后, 将通用盲文拼音按照拼音-盲文对照表对应转换, 即得到通用盲文转换结果.

由于在英文一级盲文中, 英文字符与英文盲文符号为一一对应关系, 因此本文通过编写字母-盲文对照表的方式实现英文一级盲文转换. 根据英文盲文规范建立盲文对照表, 以键-值对的形式存储英文字母及其盲文符号. 在进行盲文转换时, 遍历待转换内容的各字符. 若该字符为字母, 则从对照表中获取其对应的盲文符号, 并替换该字符. 遍历结束后, 即得到英文一级盲文转换结果, 转换流程如图 5 所示.

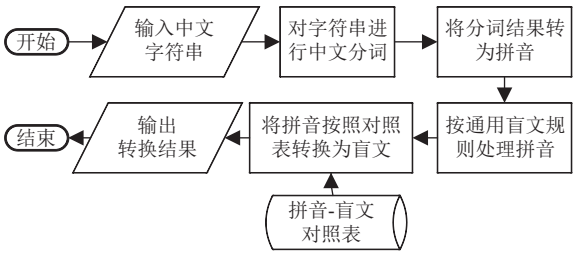


图 4 国家通用盲文转换流程

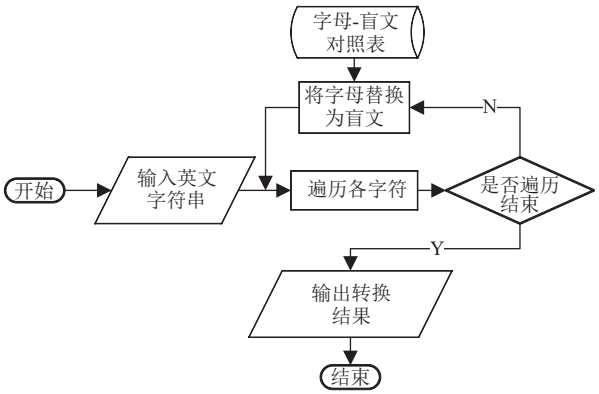


图 5 英文一级盲文转换流程

1.6 标点转换层实现

前文所述, 一符多用与混用情况可进一步分为语言类型问题和上下文结构问题. 对于语言类型问题, 可通过判断该标点所在文段的语言解决; 对于上下文结构问题, 可通过制定恰当转换规则解决. 根据以上思路, 本文设计标点转换流程, 如图 6 所示.

首先, 标点转换层读取待转换任务的语言类型, 并加载该类型对应的标点对照表. 标点对照表以键-值对形式存储了某种语言各标点符号及其对应的盲文表示, 同时也记录了易混用的标点内容. 例如, 在中文标点对照表中, 不仅记录了各中文标点和对应盲文, 还列出了可能与之混淆的英文标点, 如图 7 所示.

此外, 为了区分一符多用中的上下文结构类问题, 标点转换层还应加载标点规则库. 标点规则库通过一系列规则判定标点的具体用法, 图 8 是规则库中针对上下文结构问题的判断流程.

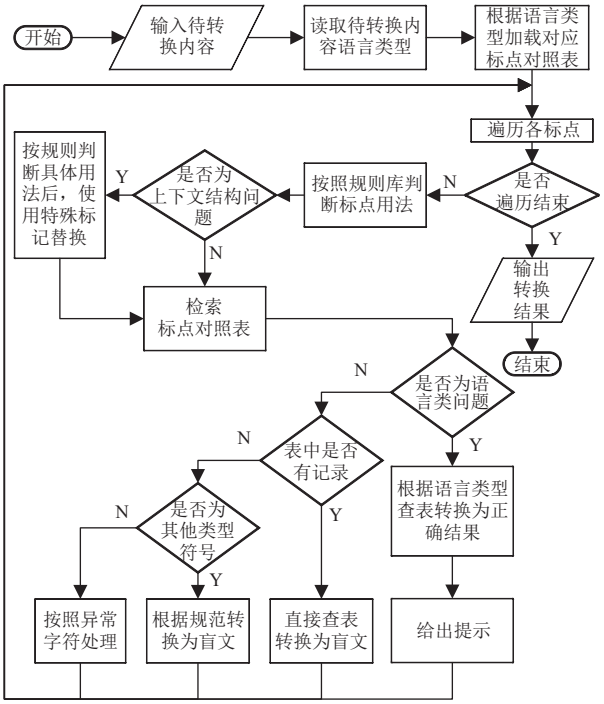


图 6 标点转换流程

随后, 标点转换层遍历待转换任务中的各标点. 首先根据规则库判断标点用法, 并使用特定标记替换具有上下文结构问题的标点. 例如使用“dp” (decimal point) 指代小数点, 采用“lq” (left quotation) 替换左直单引号等. 由于标点转换的优先级在文字转换之后, 因此采用字母标记的方式并不会影响后续转换过程. 完成判断后, 转换模块会根据句子语言类型在相应标点对照表中查找该标点 (或特殊标记). 若该标点存在于标点对照表中且不属于语言类型问题, 则直接根据对照表将该标点转换为对应盲文; 若该标点属于语言类型问题, 则将其转换为正确盲文结果, 同时给出提示信息.

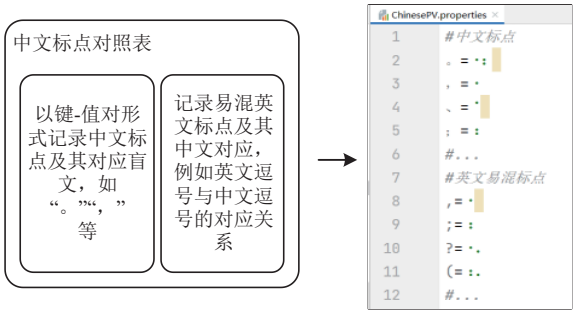


图 7 中文标点对照表示意图

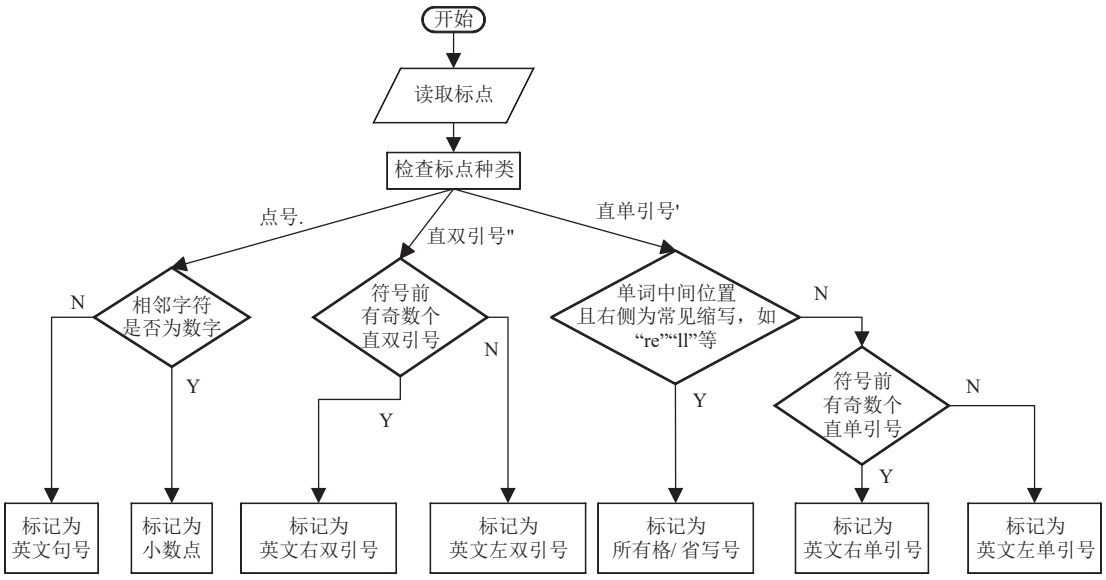


图 8 标点规则库判断具有上下文结构问题标点的流程

最后, 标点转换层将其他类型的符号转换为盲文, 并处理可能出现的异常符号情况. 标点对照表中主要记录标点符号, 但实际使用中, 可能会出现如阿拉伯数字等其他种类符号. 这些符号通常具有明确的书写规范, 在不同语言的盲文转换规则中也有相同盲文表示形式. 因此本文通过制作对照表的方式进行转换. 此外, 盲文转换过程中可能会由于输入内容乱码或对照表内容不全等问题, 出现无法转换的异常字符. 标点转换层会记录这些异常字符, 用盲文空方填充其位置, 并给出相应提示信息.

本文以“[残疾人创业]增多, 超 7.0%。”字符串为例, 说明系统执行过程. 此字符串中, 含有标点混用 (英

文逗号)、一符多用 (方括号、小数点)、中文标点 (句号) 以及其他种类符号 (数字). 其中, 汉语部分按照通用盲文转换规则处理将中文转换为盲文, 标点部分则经过标点转换层处理得到正确的盲文点序, 如图 9 所示.

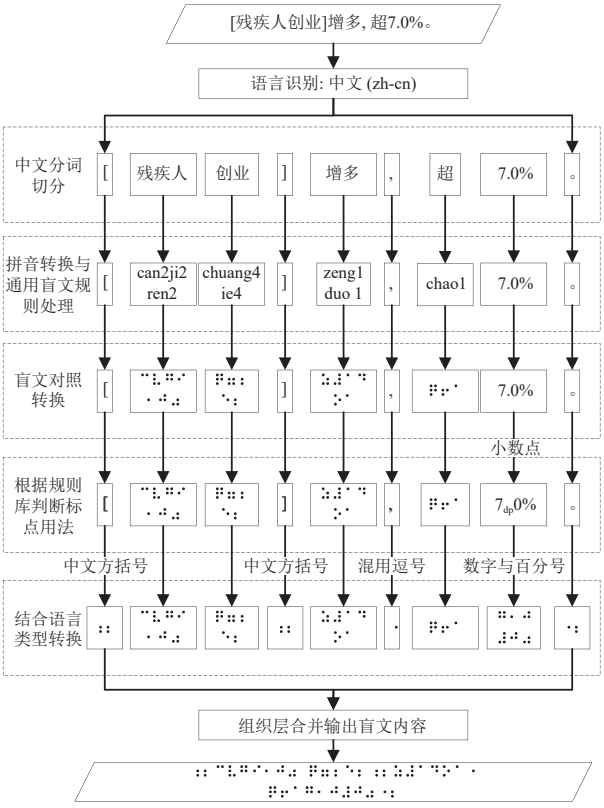


图 9 系统各阶段转换流程示意图

由系统各阶段运行结果可见, 该系统正确进行了语言识别与通用盲文转换, 同时具有标点区分和纠错能力, 能够按照语言规范将各标点转换为盲文. 根据此结果可初步判断该系统达到预期目标, 能够可靠地进行盲文转换.

2 系统测试与分析

本部分对前文所述盲文转换系统设计实验并进行测试. 从标点转换效果、准确率和转换速度等角度出发对系统进行测试, 从多方面检测系统性能.

2.1 标点转换对比实验

本实验对比分析不同盲文转换系统, 即本文所述系统、符号对应转换系统、Liblouis 转换系统及中国盲文数字平台对盲文转换过程中的一符多用及标点混用情况的转换效果.

上述 4 种盲文转换系统中, Liblouis 转换系统是一款开源盲文转换系统, 可直接使用; 中国盲文数字平台则对用户开放了网络接口服务, 可通过登录其网站进行在线盲文转换. 而符号对应转换方式目前尚未有现成的可用系统, 需自行搭建.

本文参考各符号对应转换方式的实现流程, 使用建立对照表的方式完成系统开发, 程序流程如图 10 所示. 在标点-盲文对照表中收录了常用中英文标点和对应的盲文符号. 同时, 为了控制第三方库对实验结果的影响, 符号对应转换系统仍沿用前文提及的 Jieba 分词组件和 Pinyin4J 拼音转换库进行分词和拼音转换.

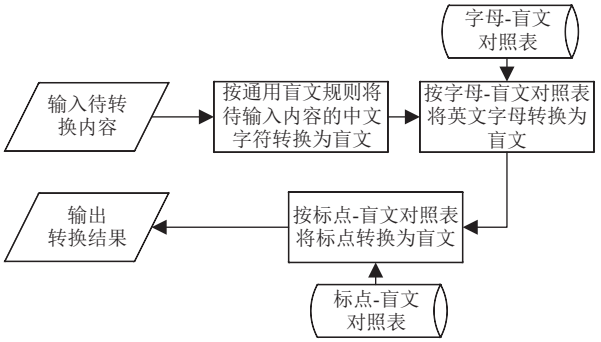


图 10 符号对应方式系统流程图

标点转换问题中的中英标点混用情况实质上是一种书写错误, 具有随机性、偶发性的特点. 若直接在各类发行刊物及网络资讯中搜索这类错误, 难度较大且效率不高. 因此, 本文使用 BCC 语料库[18]并辅以《基础盲文》[19]书中的示例, 构建了标点转换实验的测试语料, 具体内容和来源如表 3 所示. 测试语料包括多种一符多用及标点混用情况, 涵盖报刊、文章及网络对话等不同来源, 能较为全面检测系统标点转换性能.

分别使用本文转换系统、符号对应转换系统、Liblouis 盲文转换系统及中国盲文数字平台对测试语料进行转换, 得到各系统的转换结果, 如表 4 所示. 其中标注下划线的盲文符号为测试标点对应的盲文转换结果. 同时考虑到篇幅所限, 部分文字转换结果由省略号代替.

接下来, 本文从标点转换准确性、一符多用处理情况以及中英混合标点识别情况 3 方面, 对上述 4 种转换系统的性能进行了评估和比较, 结果如表 5 所示.

表 3　标点转换实验测试语料

类型	例句	类型	来源	语料库
英文逗号	样样都有, 伙计	文学	海明威/丧钟为谁而鸣	BCC
	吃天下不能吃之物,	报刊	人民日报海外版2005年07月02日	BCC
	多谢老细,	对话	网络	BCC
英文问号	发生了奇迹吗?	文学	阿来/尘埃落定	BCC
	日本专家的信任?	报刊	人民日报海外版2016年12月22日	BCC
英文冒号	他加了一句:	文学	阿瑟·高顿/死灵魂	BCC
	案例六: 违规	报刊	人民日报海外版2017年06月03日	BCC
	于小兔的内心:	对话	网络	BCC
英文分号	来发表意见;	文学	拜伦/唐璜	BCC
	求精、求珍、求罕;	报刊	人民日报2016年04月05日	BCC
英文破折号	喂, 那里—	文学	岩井俊二/华莱士人鱼	BCC
直双引号	笑道: "嗳哟	文学	张爱玲/半生缘	BCC
	实质性突破. "	报刊	人民日报2002年08月01日	BCC
直单引号	应似飞鸿踏雪泥. '	文学	王火/战争和人	BCC
	中国饮食不仅'香'	报刊	人民日报海外版2005年11月18日	BCC
	亲'我真心想睡觉'	对话	网络	BCC
英文感叹号	抗美援朝胜利万岁!	文学	魏巍/东方	BCC
	建成农民富裕村!	报刊	人民日报海外版2006年03月25日	BCC
	还有那个春卷!	对话	网络	BCC
方括号	商[商旧误作调]	古汉语	燕乐考原 艺藏\音乐	BCC
	进去了吗? [远处]	文学	扬·马特尔/少年Pi的奇幻漂流	BCC
	the color [red]	例句	基础盲文	基础盲文
点号	the worst of times.	例句	基础盲文	基础盲文
	规模提高了10.3倍	报刊	文汇报 2005年9月18日	BCC

表 4　各系统转换结果

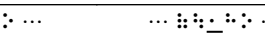
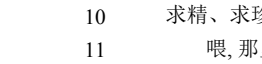
序号	例句	中国盲文数字平台	符号对应转换	Liblouis转换	本文转换系统
1	样样都有, 伙计				
2	吃天下不能吃之物,				
3	多谢老细,				
4	发生了奇迹吗?				
5	日本专家的信任?				
6	他加了一句:				
7	案例六: 违规				
8	于小兔的内心:				
9	来发表意见;				
10	求精、求珍、求罕;				
11	喂, 那里—				
12	笑道: "嗳哟				
13	实质性突破. "				
14	应似飞鸿踏雪泥. '				
15	中国饮食不仅'香'				
16	亲'我真心想睡觉'				
17	抗美援朝胜利万岁!				
18	建成农民富裕村!				
19	还有那个春卷!				
20	商[商旧误作调]				
21	进去了吗? [远处]				
22	the color [red]				
23	the worst of times.				
24	规模提高了10.3倍				

表 5　4 种转换系统性能评估概况

评估指标	中国盲文数字平台	字符对应转换系统	Liblouis盲文转换系统	本文转换系统
准确性	破折号、左方括号转换有误	基本准确	破折号、英文双引号、方括号转换有误	基本准确
一符多用	能够识别部分一符多用情况	无法区分一符多用情况	无法区分一符多用情况	能够根据实际用法区分一符多用标点
标点混用	无法对混用情况进行纠错	无法对混用情况进行纠错	无法对混用情况进行纠错	能够识别和纠正中英标点混用情况

参照《基础盲文》书中“国家通用盲文规范”与“英文盲文规范”部分关于标点符号的描述, 发现部分系统在标点的盲文表示上存在错误. 在中国盲文数字平台的转换结果中, 方括号及英文长破折号“—”的转换有误. 其中左方括号被转换为“⠠⠶”, 与中文写法“⠶”及英文写法“⠲⠶”均不同; 英文长破折号被转换为“⠤⠤”, 该写法实际上是英文连字号“-”的盲文表示, 正确的英文长破折号写法应为“⠤⠤⠤”. Liblouis 对英文长破折号的转换存在相同问题. 此外, Liblouis 对英文双引号和方括号的转换也存在错误. 英文双引号的正确盲文表示应为“⠦ ⠴”, 而 Liblouis 转换后的结果为“⠲⠶⠶⠲”; 方括号的正确表示应为“⠲⠶ ⠶⠲”或“⠶ ⠶”, 但是 Liblouis 将其转换为“⠶⠦ ⠴⠶”, 这种盲文表示不属于任何标点符号.

在一符多用情况区分方面, Liblouis 转换系统与符号对应转换系统只能按照预先设定的对照表进行转换, 无法根据实际情况动态处理. 例如, 在转换点号“.”时, 两种转换系统都将其视为英文句号进行处理. 这就导致在将句子 24 中的数字“10.3”转换为盲文时, 系统将其转换为“⠼⠁⠚⠨⠲ ⠼”, 而正确结果应该是将其视为小数点进行转换, 即“⠼⠁⠚⠨⠼”. 相比之下, 中国盲文数字平台能正确识别点号等部分一符多用情况, 但是无法区分中英方括号、所有格号与直单引号情况. 在句子 14、15、16 中, 该平台将所有直单引号都错误转换成了所有格号, 即将“⠲⠦”误转换为“⠲ ”; 在句子 20、21、22 中, 将所有方括号均采用同一种方式书写, 没有区分中英文的差异. 本文所述转换系统能够从语言和规则两方面识别大部分一符多用情况, 如中英方括号、点号以及英文直引号等.

Liblouis 转换系统、中国盲文数字平台及符号对应转换系统不具备中英标点混用的纠错能力, 在转换混用标点时, 均直接对其进行转换, 不能识别和纠正这种混用情况. 例如将句子 17、18、19 的英文感叹号“!”直接转换为“⠖”, 将句子 9、10 的英文分号“;”直接转换为“⠆ ”. 本文转换系统通过对待转换内容进行语言识别, 能够有效纠正中英标点混用情况. 从转换结果中可以看出, 本系统能够识别出测试语料中全部误用的英文标点, 并将其更正为正确的中文标点用法.

2.2　盲文转换准确率实验

本文考虑盲人群体学习需求、信息准确率以及信息广度等因素, 选择中学语文课本作为语料来源. 选取其中 20 篇文章构建汉盲对照语料库, 涵盖小说、散文、新闻、科普文章、诗歌等多种类型. 语料库共计 4.2 万字, 包括中文字符、英文字符、数字、中文标点等, 各部分组成如表 6 所示.

表 6　汉盲对照语料库各类符号数量

中文字符	英文字符	数字	中文标点
37316	70	47	5209

本文通过对比转换结果和标准结果判断转换准确率. 本文采用中国盲文数字平台获取标准结果. 首先选择未翻译的语料文件, 利用中国盲文数字平台将其转换成通用盲文, 以盲文字符形式存储. 随后, 针对语料中的多音字等特殊情况, 采用人工校对的方式逐一查看, 保证标准结果的准确性.

在准确率判定方面, 本文采用文本对比软件 Beyond Compare 进行文档分析, 采用 16 进制模式分析转换结果与标准结果差异. 比较结果包括相同字节、左边独有字节、右边独有字节和差异字节 4 部分, 如图 11 所示.

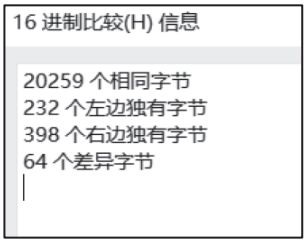


图 11　Beyond Compare 文档对比结果示例

比较分析中, 本文将左边文件设定为转换结果文档, 将右边文件设定为标准结果文档. 相同字节指转换结果和标准结果一致的字节, 这部分字节翻译正确. 左边独有字节指转换结果相比标准结果多出的字节, 这部分字节出现的原因可能包括分词粒度过细、标调冗

余或者拼音转换错误等. 右边独有字节指标准结果比转换结果多出的字节, 其原因可能在于转换结果分词粒度过粗、不恰当的省写规则等. 而差异字节则代表对同一内容, 转换结果和标准结果间的差别, 可能由错误的拼音或标调造成.

文章准确率的计算公式如下:

$$A = (T - M) \div T \times 100\% \quad (1)$$

其中, A 代表准确率, T 代表标准结果总方数, M 表示转换结果与标准结果对比的错误个数. 若分别用 C、L、R、D 代表两文档比对结果的相同字节、左边独有字节、右边独有字节和差异字节. 则标准结果总方数 T 可表示为 (C+R+D), 即相同字节、右边独有字节及差异字节之和. 错误个数 M 可表示为左边独有字节、右边独有字节及差异字节三者之和, 即 (L+R+D). 因此可得使用 Beyond Compare 软件对比的准确率计算公式为:

$$A = (C - L) \div (C + R + D) \times 100\% \quad (2)$$

使用本文所构建的中英盲文转换系统将中文语料转换为国家通用盲文, 对结果进行对比分析并计算准确率, 得到表 7 所示结果.

表 7 盲文转换准确率测试结果

相同字节	左边独有字节	右边独有字节	差异字节	原文件字数
287 730	3 895	5 264	749	42 642

根据准确率公式 (2) 及表 7 内容, 可计算出本文转换系统对通用盲文转换的准确率为 96.62%. 本系统与其他文献盲文转换系统准确率对比如表 8 所示. 由于数据集的不同与盲文转换标准的不同, 表 8 中的准确率高低并不能完全代表转换系统的性能高低. 不过可以说明本文所述系统有较高的准确率, 可以满足盲人的日常阅读需求.

表 8 盲文转换准确率对比 (%)

方法	准确率
文献[6]	97.42
文献[11]	85.11
本文转换系统	96.62

翻译错误大多出在分词及拼音转换部分. 分词的粒度过细或过粗, 均会影响转换结果和标准结果的差异, 进而影响其准确率. 拼音转换部分的错误大多由多音字引起, 典型错误包括对“着”“地”等字读音的错误识别以及轻声转换错误等.

2.3 盲文转换速度实验

盲文转换速度是影响用户阅读的重要因素, 本文就转换速度进行了相关测试.

本文测试环境为 Windows 11 系统、Intel Core i7 2.60 GHz CPU、16 GB 内存.

系统以第 2.2 节所搭建的中文语料库作为输入, 测量翻译每篇文章所用时间并记录, 结果如表 9 所示.

表 9 盲文转换速度测试结果

原文件字数	转换时间 (ms)
42 642	538

由表 9 可计算得, 本文转换系统平均转换速度约为 79 260 字/s. 该速度远大于盲人摸读速度 (小于 300 字/分), 可以满足盲人即时阅读的需要.

3 总结与展望

盲文转换技术在推进盲人群体的信息无障碍化进程中具有重要意义. 现有盲文转换系统能较好转换英文字母及单词, 但是对标点符号的转换效果不佳, 表现在一符多用及中英标点混用方面. 本文详细分析以上问题, 设计并实现一种具有标点区分和纠错能力的中英盲文转换系统. 通过构建测试语料, 对该系统开展一系列测试和评估. 实验结果表明, 与其他类型转换系统相比, 本系统能够结合语言类型和上下文结构, 有效区分标点转换中的一符多用情况, 并能纠正标点混用问题. 此外, 在文字转换性能方面, 本系统的盲文转换正确率达 96.62%, 准确性较高. 本系统还具有较高的转换速度, 能够快速处理大量语料.

下一步工作, 可在标点转换及实验设计上进行进一步研究.

不管是判断一符多用情况还是标点混用情况, 本系统实质上是采用基于规则的判断方式. 这种方式能够解决大部分标点转换问题, 但是仍有部分使用情况无法区分. 比如英文所有格的一种使用情况是“复数名词的所有格直接在单词后加‘’”, 例如“Students'”. 这种写法和右直单引号的写法一致, 仅依靠规则难以分辨, 需要结合语境表述进一步区分. 在下一步研究中, 可以考虑引入自然语言处理技术, 进一步提升标点识别能力.

本文考虑标点符号的常见用法, 从 BCC 语料库中选择部分例句, 构成了标点转换实验的测试语料. 但是

实际情况中, 标点使用情况往往会更为复杂. 因此, 为了更加全面的评估和提升系统性能, 有必要从标点运用的角度出发, 构建涵盖使用范围更广、用法更全的标点语料库, 并进一步测试系统性能.

未来工作中, 我们将逐步完善本盲文转换系统在以上方面存在的不准确之处, 不断提高系统整体性能.

参考文献

1 何川. 国内信息无障碍的现状及展望. 现代电信科技, 2007, 37(3): 4–8. [doi: 10.3969/j.issn.1002-5316.2007.03.002]

2 赵燕潮. 中国残联发布我国最新残疾人口数据. 残疾人研究, 2012(1): 11.

3 钟经华. 盲文应用的现实困境与思考. 现代特殊教育, 2016(13): 27–28. [doi: 10.3969/j.issn.1004-8014.2016.07.011]

4 陈楷闻, 林栋, 钟泽栋, 等. 视听触同步刺激的数字化盲文学习方法. 计算机系统应用, 2021, 30(9): 262–270. [doi: 10.15888/j.cnki.csa.008100]

5 吕先超. 视障汉语转换软件 SunBraille 的设计实现 [硕士学位论文]. 兰州: 兰州大学, 2016.

6 胡庆玲, 林栋, 陈楷闻, 等. 基于国家通用盲文标调规则的汉盲转换系统. 计算机系统应用, 2022, 31(12): 59–68. [doi: 10.15888/j.cnki.csa.008877]

7 毛扬, 梁宏博, 邹成洋, 等. 基于 Cortex-M3 的汉盲翻译 SoC 设计. 计算机系统应用, 2023, 32(10): 132–139. [doi: 10.15888/j.cnki.csa.009256]

8 陈楷闻. 嵌入式盲文数字化系统研究 [硕士学位论文]. 杭州: 浙江理工大学, 2022. [doi: 10.27786/d.cnki.gzjlg.2022.001132]

9 林栋. 面向国家通用盲文方案的盲文转换方法研究 [硕士学位论文]. 杭州: 浙江理工大学, 2022. [doi: 10.27786/d.cnki.gzjlg.2022.000779]

10 邹成洋. 基于 Linux 系统的盲文学习机系统设计 [硕士学位论文]. 杭州: 浙江理工大学, 2023. [doi: 10.27786/d.cnki.gzjlg.2023.000825]

11 蔡佳, 王向东, 唐李真, 等. 基于汉盲对照语料库和深度学习的汉盲自动转换. 中文信息学报, 2019, 33(4): 60–67. [doi: 10.3969/j.issn.1003-0077.2019.04.007]

12 蒋琪. 基于 NMT 的端到端汉盲转换方法研究 [硕士学位论文]. 兰州: 兰州大学, 2021. [doi: 10.27204/d.cnki.glzhu.2021.001436]

13 王蕊. 基于预训练模型的汉盲转换方法研究 [硕士学位论文]. 兰州: 兰州大学, 2023. [doi: 10.27204/d.cnki.glzhu.2023.003612]

14 许书道. 科技出版物中、英文标点符号使用的问题. 编辑学报, 2013, 25(6): 547–548. [doi: 10.16811/j.cnki.1001-4314.2013.06.015]

15 丁波. 基于 N-Gram 向量特征的社交媒体短文本语种识别方法研究 [硕士学位论文]. 北京: 北京邮电大学, 2020. [doi: 10.26969/d.cnki.gbydu.2020.002040]

16 史国举. 基于 Python 的中文分词技术探究. 无线互联科技, 2021, 18(23): 110–111. [doi: 10.3969/j.issn.1672-6944.2021.23.052]

17 苏伟, 许存禄, 林和, 等. 中国盲文数字平台建设研究. 现代特殊教育, 2021(14): 68–73. [doi: 10.3969/j.issn.1004-8014.2021.14.010]

18 荀恩东, 饶高琦, 肖晓悦, 等. 大数据背景下 BCC 语料库的研制. 语料库语言学, 2016, 3(1): 93–109, 118.

19 琚四化. 基础盲文. 南京: 南京大学出版社, 2020.

(校对责编: 张重毅)