古汉语通假字资源库的构建及应用研究

王兆基♠ 张诗睿♣ 张学涛♡ 胡韧奋♡,✉*
北京师范大学国际中文教育学院
♠zhaoji.wang@mail.bnu.edu.cn ♣1169882881@qq.com
♡{11112011118, irishu}@bnu.edu.cn

摘要

古籍文本中的文字通假现象较为常见，这不仅为人理解文意造成了困难，也是古汉语信息处理面临的一项重要挑战。为了服务于通假字的人工判别和机器处理，本文构建并开源了一个多维度的通假字资源库，包括语料库、知识库和评测数据集三个子库。其中，语料库收录11000余条包含通假现象详细标注的语料；知识库以汉字为节点，通假和形声关系为边，从字音、字形、字义多个角度对通假字与正字的属性进行加工，共包含4185个字节点和8350对关联信息；评测数据集面向古汉语信息处理需求，支持通假字检测和正字识别两个子任务的评测，收录评测数据19678条。在此基础上，本文搭建了通假字自动识别的系列基线模型，并结合试验结果分析了影响通假字自动识别的因素与改进方法。进一步地，本文探讨了该资源库在古籍整理、人文研究和文言文教学中的应用。

关键词： 古代汉语 ； 资源库 ； 通假字 ； 自动识别

The Construction and Application of an Ancient Chinese Language Resource on Tongjiazi

Zhaoji Wang Shirui Zhang Xuetao Zhang Renfen Hu✉
School of International Chinese Language Education, Beijing Normal University

Abstract

In ancient Chinese texts, it is common to use characters with the same sound or similar sounds instead of the original characters, that is, to use Tongjiazi. This not only creates difficulties for people to understand the meaning of the text, but also an important challenge for ancient Chinese information processing. In order to assist the manual analysis and machine processing of Tongjiazi, this paper builds a multi-dimensional language database, including three sub-databases, i.e. corpus, knowledge base and evaluation data set. Among them, the corpus contains more than 11,000 sentences with detailed annotations of Tongjia usages. The knowledge base is presented in graph data with 4185 characters as the nodes and 8350 relations between them as the edges. The attributes of the nodes and the edges are labeled from the perspectives of pronunciation, glyph and meaning. The evaluation data set is designed for automatic recognition of Tongjia usages, including training and testing data for two subtasks: Tongjiazi detection and the recognition of the original characters. Now the evaluation data covers 19678 entries. On this basis, this paper builds a series of baseline models for the automatic recognition of Tongjia usages, and analyzes the factors affecting the recognition results and the improvement methods. Further, this paper discusses the application of these resources in different fields, e.g. the collation of ancient books, humanities research and classical Chinese learning and teaching.

Keywords: ancient chinese , resource , Tongjiazi , automatic recognition

*Corresponding author.

第二十二届中国计算语言学大会论文集，第535页-第546页，哈尔滨，中国，2023年8月3日至5日。

1 引言

与现代汉语及其他语种不同的是，古籍文本中的文字通假较为常见，这为准确理解文意造成了困难。具体来说，通假指的是古人本有其字而不用，反而借用一个音同或音近字的现象，其中，被借用的字称作通假字，被代替的字称为正字或本字 (孔德明, 1993; 王宁, 2012)。例如，在“庄公寤生，惊姜氏。”（出自《左传》）中，“寤”为通假字，所通正字为“牾”，“牾生”即逆生，表示难产。

通假现象不仅常见于传世古籍，在出土文献中也有较高频率。据钱玄 (1980)统计，现存《老子》(据唐傅奕校《道德经古本篇》) 约5500余字，其中用通假字30多个，而马王堆帛书《老子》(乙本) 使用通假字320个，占全书的6%。整理古籍时，通假字识别对于准确理解文意来说十分重要，如王引之在《经义述闻·经文假借》中所述：“学者改本字读之，则怡然理顺；依借字解之，则以文害辞。”除了专业学者整理古籍时需要释读通假字，在中学文言文教学中，通假字也是一项重点和难点，掌握文言文常见通假字的用法是文言文阅读的基本功 (由明智, 2013)。值得一提的是，对于汉语史研究来说，通假字与被通假字之间的音同或音近关系可以为汉语古音和语音史研究提供宝贵的参考资料 (张儒, 1988; 党怀兴, 1998)；同时，字与字之间的通假关系亦有助于厘清词汇形式和词义演变的脉络，从而服务于词汇发展变化和词汇语义研究 (孙建伟, 2015)。可以说，无论是服务于通假字识别，还是汉语史研究，高质量的通假字资源库都必不可少。柳建钰和周晓文 (2017)从辅助校勘需求出发提出了构建通假字资源库的设想，拟基于各类通假字字典搜集整理通假字表，预计收录字头6000个左右，涵盖传世文献和出土文献的通假字。然而，目前除了字典和辞书以外，通假字相关数据资源的建设仍十分罕见。

近年来，伴随古籍整理利用的转型升级，古汉语信息处理资源建设和算法研究受到了越来越多的关注，涉及任务包括句读标点、分词和词性标注、词义消歧、命名实体识别、自动校勘、文本生成等 (邓三鸿等, 2021; 胡韧奋等, 2021; Liu 等, 2019; 苏祺等, 2021; 舒蕾等, 2022; Yang 等, 2021; Yi 等, 2018)。需要指出的是，在古汉语信息处理研究中，通假字识别是词义注释、文本校勘和文白翻译等技术的基础，目前学界尚无针对通假字自动识别的研究。当下，即使ChatGPT、GPT4等大模型具备极强的自然语言理解能力，其处理包含通假字的文本时仍会“以文害辞”，一例如表 1 所示。因此，对于古汉语信息处理来说，通假字资源库的建设及自动识别技术研究具有其现实价值和迫切性。

领域知识	输入	输出
无领域知识	请将下面的文言文翻译为现代汉语: 廉將軍雖老，尚善飯，然與臣坐，頃之，三遺矢矣。	尽管廉将军年事已高，但仍然精于弓箭。然而在与臣下一起坐着的过程中，不久，他三次失手放箭。
有领域知识	请将下面的文言文翻译为现代汉语: 廉將軍雖老，尚善飯，然與臣坐，頃之，三遺矢矣。 下面是可能的通假字用法，可结合语义判断是否使用: “矢”共有3种通假用法：①“矢”通“誓”。发誓永远要(做某事)。②“矢”通“屎”。粪便。③“矢”通“弛”。毁坏。	廉将军虽然年纪大了，但仍然能吃得下饭，然后与臣子坐下来。不久之后，他已经三次去上厕所了。 在这里，通假字“矢”应该使用第二种用法，表示“屎”。

表 1. 基于GPT-4的文白翻译实例

为了更好地辅助通假字的人工判别和机器处理，本文首先构建了一个涵盖标注语料库、知识库和评测数据集的通假字资源库。其中，标注语料库收录了11000余条包含通假字用例标注的语料，知识库以汉字为节点，通假关系为边，从字音、字形、字义多个角度对通假字与正字的属性进行加工，共包含4185个字节点、7700对通假字关联信息、650对通假字之间的形声关系信息；评测数据集分为基础版和拓展版，支持通假字检测和正字识别两个子任务的评测，收录评

第二十二届中国计算语言学大会论文集，第535页-第546页，哈尔滨，中国，2023年8月3日至5日。
(c) 2023 中国中文信息学会计算语言学专业委员会

测数据19678条。在此基础上，本文搭建基线模型开展了通假字检测和正字识别实验，并探讨了资源库在古籍整理、人文研究和文言文教学中的应用。

2 通假字资源库构建

为了让资源库更好地服务于与通假字有关的文史研究和自动识别算法研发，我们设计并构建了三个开源资源库，均以JSON格式发布[1]，包括：通假字标注语料库、通假字知识库与通假字识别评测集。

2.1 通假字标注语料库

目前，学界尚无专门标注通假字的文言文语料库，包含通假字的句篇信息主要见于各类辞书，其中也包括专门的通假字字典，如高亨《会典》收录了传世文献材料中的通假字用法，《简帛古书通假字大系》侧重于依据战国秦汉出土简帛文献。考虑到与通假字相关的辞书存在应用场景区别，为兼顾古汉语信息处理、文史研究与文言文教学的一般性需求，本文选择以《汉语大词典》为数据源，构建通假字标注语料库。该库可为通假字相关研究和应用提供高质量的基础性数据，亦可结合具体需求进行筛选、优化和扩充。

《汉语大词典》所收条目分为单字条目与多字条目。多字条目按“以字带词”的原则,列于单字条目之下。一个单字有两个以上字头的，在字头旁以阿拉伯数字标注序号。字头下依次标注现代音与古音，其中，现代音用汉语拼音字母标注，古音用反切标注。释义时，通假义用“通‘×’”和“‘×’的被通假字”表示。据此，我们以《汉语大词典》的标注为准，采集通假现象涉及的释义及例句，例如，在《汉语大词典》中，字头“耗3”的内容如表 2 所示，该字可通“眊1”，表示“昏乱不明”，词典收录了来自《荀子·修身》与《汉书·景帝纪》的两则包含通假字的例句。

耗3〔mào ㄇㄠˋ〕 〔《字彙》莫報切〕 通“眊1”。 昏乱不明。《荀子·修身》：“少而理曰治，多而亂曰耗。”《汉书·景帝纪》：“其令二千石各脩其職，不事官職耗亂者，丞相以聞，請其罪。”颜师古注：“耗，不明也，讀與眊同。”

表 2. 通假字标注语料库语料原文示例

经自动提取和人工校对，我们从《汉语大词典》中采集了较大规模的通假字属性及用例数据，在此基础上构建了高质量的通假字标注语料库，共包含11000余句繁体中文语料，覆盖2479个通假字。其中，用例最多的为“辨”，存在通“辯”、“變”、“班”、“般”等字的126条用例，同时，也有不少通假字的例句数量较少，例如，有833个通假字仅包含1条用例语料。如表 3 所示，语料库中的每一条语料包含11个属性：语料ID、语料文本、标注位置、通假字字头、正字[2]字头、出处、时代、释义、拼音、注音和古音。

语料ID	1
语料文本	少而理曰治，多而亂曰耗。
标注位置	10
通假字字头	耗3
正字字头	眊1
出处	《荀子·修身》
时代	战国
释义	昏乱不明。
拼音	mào
注音	ㄇㄠˋ
古音	《字彙》莫報切

表 3. 通假字标注语料库语料示例

[1]数据下载地址：https://github.com/frederick-wang/tongjiazi-resources
[2]资源库中采用“正字”标识被通假字。

2.2 通假字知识库

本文采用图数据结构设计通假字知识库，其中，汉字为节点(node)，通假关系和形声关系为边(edge)。在汉语史研究中，对通假这种字用现象的分析往往会从字音、字形、字义等多个角度展开，因此，我们在构建知识库时也分别针对音、形、义进行了属性标注。在**字音**方面，我们在通假关系边中标记了与其对应的注音和古音属性，同时参考胡韧奋等(2013)加工的形声字数据添加了字和字之间的形声关系边。在**字形**方面，我们在字节点上标记了部首、部件和结构信息，其中汉字部件信息参考了Yan et al. (2013)构建的数据集。在**字义**方面，字节点之间的通假关系考虑了义项的差别，即A字通B字时根据语境差别有多种含义，因此一对字节点之间允许有多个通假关系边相连。通假字知识库的详细规范参见附录 A。

为了更好地服务于汉语史及通假字自动识别研究，本文在《汉语大词典》的基础上，进一步从康熙字典[3]、汉典[4]、国学大师网汉语字典[5]中采集了与通假用法相关的多源异构数据，共计315万字，融合构建了通假字知识库。在融合数据时，对字而言，以字形为标准与其他来源的数据合并。对通假关系而言，以通假字、正字、释义为标准与其他来源的数据合并。最终，通假字知识库收录了4185个字节点、7700对通假关系、650对形声关系。

图 1 以“辟”为例，展示了知识库中节点和边的属性。其中，字节点属性标注在蓝框内；红色的有向边表示通假关系，通假关系的详细属性参见红框，与通假关系相关联的语料以紫框标注；绿色的有向边表示形声关系，对应的绿框为形声关系的具体属性。由图中内容可见，“辟”与“譬”之间存在3条通假关系连边，对应三种释义，同时，二者之间还包括一条形声关系连边，标识“辟”是“譬”的声符。

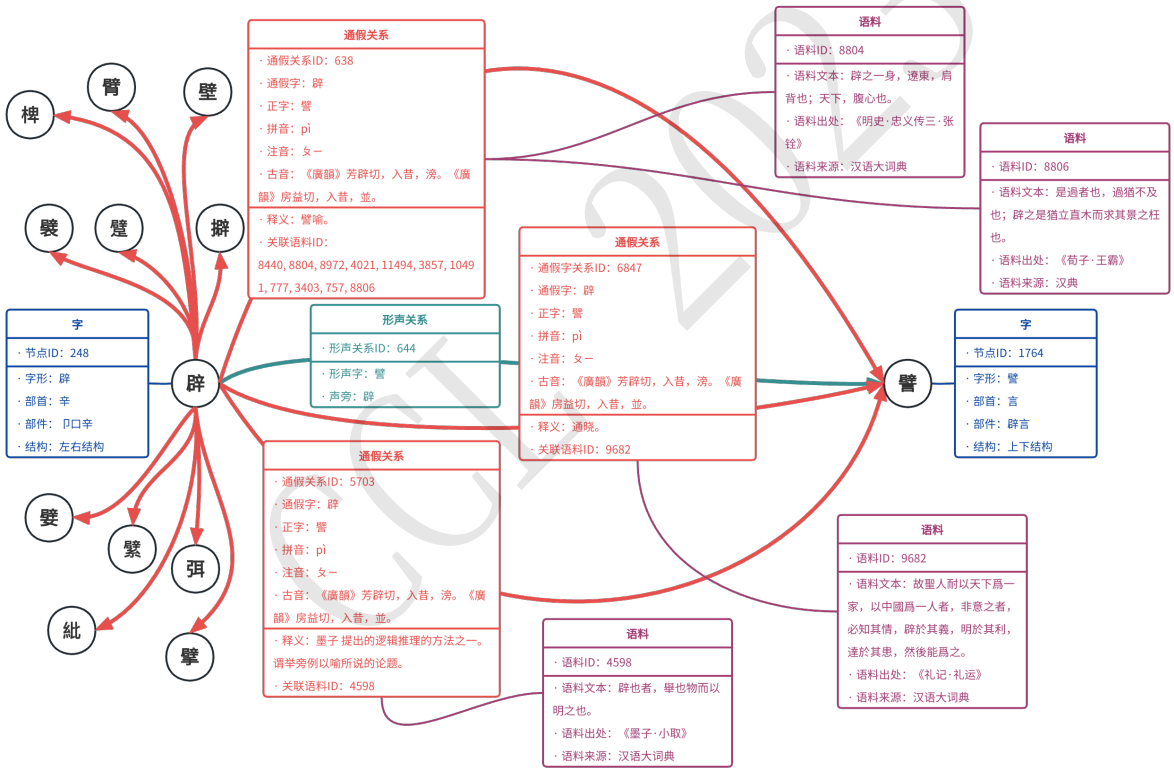

图 1. 通假字知识库示例“辟”：限于空间，图中仅展示了“辟”与“譬”之间的详细关系及部分关联语料，与“辟”有通假或形声关系的其他字仅在图中呈现节点，如“擗”、“繁”等。

通假字知识库对通假字与正字之间的关系进行了详细加工，相关信息可以从数据和特征两个维度为通假字的自动识别提供支持（参见本文第 3 节）。同时，通假字与正字之间的音、形、义关联也可为汉语史领域的相关研究提供参考。首先，它可以帮助相关研究者更高效地开展传统研究，常见的应用场景包括：

[3]使用“汉典”版的《康熙字典》数据。
[4]汉典：https://www.zdic.net/
[5]国学大师网汉语字典：http://www.guoxuedashi.net/zidian/

1. **字词考证**：通假字知识库可以帮助我们迅速辨别出通假字，识别出这是常用的通假还是在特定语境中出现的借字。例如，在图 1 中，“辟”字与“譬”字之间的通假关系，可以帮助我们了解到“辟”字在某些语境下可以作为“譬”的通假字使用。

2. **词汇语义研究**：通假字知识库可以帮助我们将某些和本义无关的假借义从词义引申中剔除，例如图 1 中“辟”通“譬”对应的三种释义。此外，通假字关联网络还能帮助系联同义、近义或词义相关的词，从而辅助词汇语义研究。

3. **形声字研究**：知识库中的字节点之间除了通假关系边，还有形声关系边，例如，在图 1 中，“譬”字是一个形声字，其声旁为“辟”。通假与形声关联数据可以辅助我们进一步研究形声字及语音的发展规律。

值得一提的是，通假字知识库能够提供传统辞书**无法呈现**的大规模通假字关联网络信息，这也为汉语史研究提供了新的视角，潜在的应用场景包括：

1. **量化通假强度**：在传统研究中，字与字之间的通假关系仅分为“有”和“无”，但这种粗粒度的判断方式并不精确。事实上，有些通假关系应用广泛，而有些仅为辞书中的孤例。基于通假字图知识库，我们可以通过字与字之间不同义项的通假关系数量（边数）以及相关联的语料数量来量化“通假关系的强度”，为后续研究提供更多可能性。

2. **利用子图探究通假规律**：传统研究范式下，由于人的时间和精力有限，研究通常仅针对一个字的通假关系及其相关的几个被通假字进行，相当于仅能研究图中的几个节点及其边。借助图数据库，我们可以根据分割条件迅速将所有数据划分为多个子图，研究子图中所有通假字节点与通假关系边的内在规律，并探讨子图间的联系。这将有助于我们发现更多的通假规律。例如，研究一个通假字的所有通假变化轨迹，实际上就是寻找该节点所在的子图并获得一个子图的生成树。

3. **辅助古汉语语音演变研究**：通假关系存在的前提是字之间的音同或音近，而不少汉字的读音在历史上经历了较大变化。利用通假字图知识库，我们可以为相关语音研究提供支持。例如，我们可以根据通假关系边关联语料的“出处”数据，获取不同时期的字与字之间的通假关系并生成关联子图，进而量化估计在某一特定时代，两个字的发音可能相同；而在另一时代，这两个字的发音可能不同。如此一来，我们便能从历时角度利用图知识库为语音演变研究提供支持。

2.3 通假字识别评测集

为了推动通假字自动识别算法研究，我们基于高质量的通假字标注语料库构建了通假字识别评测集。评测集分为两个子任务：通假字检测与正字识别。为了更好地评估模型的泛化能力，每个子任务均分为基础版与拓展版，其中，基础版任务的训练集与测试集覆盖的目标字范围一致，而拓展版任务的测试集中则包含训练集未出现的通假用法，其自动探测和识别的难度更高。接下来，本节将介绍两个评测子任务的形式及评测数据集的构建方法。

2.3.1 评测任务设计

表 4 给出了两个子任务的示例，其中，通假字检测任务旨在识别古汉语文本中的通假字位置，即给定一段输入文本，需输出文本中所有通假字的位置（从0开始计数）。如果该文本中没有通假字，则输出`[]`。计算精确率和召回率时，使用（句子，位置）二元组作为计算单位。

正字识别任务的目标是识别出古汉语文本中通假字所对应的正字，输入一段文本和通假字位置，需输出该位置的通假字所对应的正字。计算精确率和召回率时，使用（句子，位置，正字）三元组作为计算单位。

子任务	输入	输出
通假字检测	{"sentence": "北庭使劉涣躬行勃逆，委公斬之。"}	[7]
正字识别	{"sentence": "不韋毁身燋慮，出于百死。", "pos": 4}	"焦"

表 4. 评测任务示例

第二十二届中国计算语言学大会论文集，第535页-第546页，哈尔滨，中国，2023年8月3日至5日。

2.3.2 评测数据集的构建

考虑到通假字标注语料库主要收录目标字作为通假字使用的数据，为了评测模型的判别能力，兼使其适应真实应用情境，我们从词典中的其他义项例句中补充了目标字非通假用法的数据，构成正负例，如下面两例所示。

例1. **惠**心燭千仞，雄風扇八區。（正例，通“慧”，表“明慧”含义。）

例2. 必也君亂之，君終之，君之**惠**也。（负例，表“恩惠”含义）

考虑到通假字的常用度存在差异，且有必要对模型的泛化能力进行评估，我们构建了基础版和拓展版两类评测数据集。基础版评测旨在识别常用通假用法，其中，每个通假字收录至少10条正例，最多不超过20条[6]。同时，尽量补充与正例数量相等的负例，即目标字非通假用法的例句。进一步地，将每个通假字的正例和负例均按照8:2的比例拆分，分别划入训练集和测试集，从而保证训练集与测试集的数据分布相同。最终，基础版数据集覆盖了279个常见通假字，包含7962条语料，其中，训练集6190条，测试集1772条。

针对用例并不充足的通假字，我们又额外构建了拓展版评测数据。拓展版训练集与基础版训练集保持一致，拓展版测试集则在基础版测试集的基础上，额外补充了2200个通假字的正例和负例数据，其中，每个通假字的正例少于10句，负例与其数量相当，共计增补了11716条语料，因此拓展版测试集共收录13488条语料。由于拓展版测试集中收录了大量训练集未覆盖的通假用法，这便要求模型结合语境识别出训练时未见过的通假字，无疑挑战性极高，也更加接近真实的应用情境。

3 通假字自动识别评测

基于上节介绍的评测任务和数据，我们就通假字的自动识别开展了初步探索，以期为未来学界的相关研究提供基线（Baseline）结果[7]。接下来，将首先介绍本文引入的基线方法，然后，将分别报告通假字检测（基础版）、通假字检测（拓展版）、正字识别（基础版）和正字识别（拓展版）任务的评测结果，并展开分析和讨论。

3.1 实验方法

为了服务于通假字探测和正字识别，我们首先参考文本纠错的实验设定构建了一个“通假字-正字”混淆集。混淆集数据采集自通假字知识库和评测训练集中的“通假字-正字”字对。由于测试集中的语料来自《汉语大词典》，为避免测试数据的字对信息泄露，我们在使用通假字知识库中的字对数据时，排除了来自《汉语大词典》的数据。

3.1.1 通假字检测任务

在通假字检测任务中，我们采用了四类基线模型：Ngram语言模型、GPT2语言模型、BERT MLM语言模型和基于BERT的通假字检测微调模型。

对于Ngram语言模型来说，我们利用KenLM计算句子的概率得分p，并代入公式$p^{-\frac{1}{N}}$得到句子困惑度P。困惑度越低，句子表达越合理。对于句中每个位置的汉字，如果它位于混淆集中，将分别计算用它的混淆字替换原字后的句子困惑度，并与句子的初始困惑度进行比较。如果替换后的句子困惑度比原句低，则将该字所在的位置标记为通假字位置。实验中，我们分别基于殆知阁古代文献藏书2.0版和文渊阁版四库全书（繁体）语料库训练bigram和trigram模型，得到了四组结果，后文分别用DaizhigeBigram、DaizhigeTrigram、SikuBigram和SikuTrigram指代。

基于GPT2语言模型的通假字检测方法与Ngram模型类似，即利用困惑度和混淆集信息标记通假字位置。实验采用了Huggingface中两个开源的古汉语GPT2模型，分别基于殆知阁和四库全书语料训练，后文用DaizhigeGPT2[8]和SikuGPT2[9]指代。

利用Bert MLM语言模型进行实验时，我们依次判断句中的每一个字是否位于混淆集中。若在，则将该位置用[MASK]遮罩，并输出Mask LM的预测结果，从而得以比较原字与混淆集中对应字的预测概率，如果存在混淆字的预测概率高于原字，则将该字所处位置标记为通假字位置。

[6]为确保数据分布的均衡性，如果通假字在标注语料库中的例句大于20条，则随机抽取20句。

[7]https://github.com/frederick-wang/tongjiazi-evaluation

[8]https://huggingface.co/uer/gpt2-chinese-ancient

[9]https://huggingface.co/JeffreyLau/SikuGPT2

第二十二届中国计算语言学大会论文集，第535页-第546页，哈尔滨，中国，2023年8月3日至5日。

除了BERT MLM模型外，我们还引入了BERT微调方法。具体来说，通假字检测可建模为token序列标注任务，句中非通假字对应的标签为0，通假字的标签为1。微调阶段，采用BERT+全连接层的结构进行token标签学习[10]。推理阶段，如果句中某字既被模型标记为1，又是混淆集中收录的通假字，则将该字所在的位置标记为通假字位置。同时，我们也引入了一个无需混淆集的版本，只要该字被模型标记为1，就将对应位置标记为通假字位置。实验中，为了与前面三种方法对应，我们采用了基于殆知阁语料库训练的古汉语BERT模型和Huggingface上开源的SikuBERT模型[11]，经微调，得到了TongjiaziDetectionDaizhigeBert模型与TongjiaziDetectionSikuBert模型。

3.1.2 正字识别任务

与检测任务类似，正字识别任务也可基于Ngram语言模型、GPT2语言模型、BERT MLM语言模型和BERT微调模型实现。

对于Ngram、GPT2模型来说，我们将判断句中给定位置的字符是否在混淆集中，如果不在，将该字符直接作为识别的正字；如果在，则依次计算混淆字替换该字符后的句子困惑度，并与句子的初始困惑度进行比较，取句子困惑度最小的字作为识别的正字。BERT MLM模型的识别方法与之类似，如果给定位置的字符不在混淆集中，则将该字符作为识别的正字；如果在，则将该字符用[MASK]遮罩，利用Mask LM获取原字符与所有混淆字的预测概率，取预测概率最大的字作为识别的正字。

关于BERT微调方法，我们借鉴Mask LM任务的形式，要求模型预测出句中通假字所对应的正字，其余位置的字符不参与训练[12]。经过微调，模型加强了正字和上下文语境信息之间的关联，在推理阶段，采用与上述BERT MLM模型一致的方法获取正字识别结果。后文用ZhengziRecognition指代经微调训练的识别模型。

3.2 实验结果

3.2.1 通假字检测任务

表 5 和表 6 分别列出了通假字检测任务在基础版和拓展版数据集上的评测结果。在基础版测试集上，模型检测最优F1值达到66.94%，拓展版测试集的最优检测F1值为21.63%，可见通假字检测是一个极有挑战性的任务，在处理模型训练未见过的通假用法时尤为困难。通过对比不同模型，我们发现以下几点要素或对模型的检测性能产生影响。

序号	模型	精确率	召回率	F1
1	DaizhigeBigram	7.55%	22.26%	11.27%
2	SikuBigram	9.21%	18.52%	12.30%
3	DaizhigeTrigram	7.56%	18.62%	10.75%
4	SikuTrigram	10.00%	11.80%	10.83%
5	DaizhigeGPT2	8.59%	22.74%	12.47%
6	SikuGPT2	10.84%	17.47%	13.38%
7	DaizhigeBert	20.24%	55.28%	29.63%
8	SikuBert	29.09%	57.49%	38.63%
9	TongjiaziDetectionDaizhigeBert	65.02%	64.40%	64.71%
10	TongjiaziDetectionSikuBert	64.25%	69.87%	66.94%
11	TongjiaziDetectionDaizhigeBert（无混淆集）	62.10%	64.78%	63.41%
12	TongjiaziDetectionSikuBert（无混淆集）	61.96%	70.35%	65.89%

表 5. 通假字检测任务（基础版）实验结果

(1) 模型结构与复杂度

[10]训练模型时，Torch、Numpy和random模块的随机数种子为42，Batch大小设为8，Epoch数设为5，采用AdamW优化器，学习率设为5×10^{-5}。按照9：1的比例将训练数据划分为训练集与验证集，*RandomState*同样设为42。

[11]https://huggingface.co/SIKU-BERT/sikubert

[12]在微调模型时，Torch、Numpy和random模块的随机数种子、Batch大小、Epoch数、优化器，学习率、训练数据划分方法均与前文的TongjiaziDetectionBert模型相同。

序号	模型	精确率	召回率	F1
1	DaizhigeBigram	4.63%	10.81%	6.48%
2	SikuBigram	5.62%	8.74%	6.84%
3	DaizhigeTrigram	4.50%	8.64%	5.92%
4	SikuTrigram	6.21%	5.35%	5.75%
5	DaizhigeGPT2	5.25%	11.23%	7.16%
6	SikuGPT2	7.39%	8.60%	7.95%
7	DaizhigeBert	9.80%	22.73%	13.69%
8	SikuBert	15.54%	23.50%	18.71%
9	TongjiaziDetectionDaizhigeBert	31.54%	11.68%	17.05%
10	TongjiaziDetectionSikuBert	27.78%	12.12%	16.88%
11	TongjiaziDetectionDaizhigeBert（无混淆集）	32.94%	16.10%	21.63%
12	TongjiaziDetectionSikuBert（无混淆集）	29.48%	16.53%	21.18%

表 6. 通假字检测任务（拓展版）实验结果

实验结果显示，预训练语言模型具有较好的语境信息编码能力，在一定程度上能够辅助探测通假字，其中，基于BERT模型的方法效果普遍最优，GPT2模型次之，Ngram模型最弱。推测一方面与模型的复杂程度有关，Ngram模型最为简单，对上下文信息的捕捉能力最弱，另一方面也和模型结构有关，与GPT2单向自回归训练机制不同，BERT在预训练阶段的双向编码机制使其更擅长利用上下文语境信息进行字符判断。

(2) 预训练数据

在不同类型的模型上，基于文渊阁版繁体四库全书数据训练的模型表现普遍优于基于殆知阁数据训练的模型。殆知阁语料库规模更大，繁简混合，而文渊阁版四库全书（繁体）数据规模偏小，全部为繁体。考虑到我们的评测数据均为繁体中文，这与四库版预训练模型更为匹配。

(3) 微调机制的引入

在基础版评测数据集上，无论是DaizhigeBert还是SikuBert，微调后精确率和召回率均有显著提升，相较之下，精确率提升幅度更为突出，这意味着微调前，模型倾向于将非通假用法识别为通假字，而经过训练数据上的微调，模型熟悉了常见通假字用法，探测精确率得到显著改善。

在拓展版评测数据集上，微调同样提升了BERT模型的精确率，但也使其召回率出现了明显下降，推测这主要是由于拓展版测试集中收录了大量训练集未覆盖的通假用法，在训练集上微调使得模型聚焦于用例较多的常见通假字，对训练中未见过的通假用法不再关注，从而降低了识别的召回率。

(4) 混淆集的使用

在通假字检测任务（基础）中，使用混淆集的TongjiaziDetectionBert精确率略高于无混淆集版，召回率二者几乎一致。但是，在拓展版任务中，无混淆集的TongjiaziDetectionBert不论是精确率还是召回率都优于带混淆集版，这主要是由于拓展版数据集中存在不少混淆集未覆盖的通假用法，使用混淆集反而在一定程度上限制了模型的识别效果。

3.2.2 正字识别任务

表 7 示出了正字识别的实验结果，在基础版测试集上，模型最优准确率为65.64%，在拓展版评测集上，模型最优准确率为19.88%。与通假字检测任务类似，BERT系列模型普遍表现最优，同时，引入微调机制能够进一步提升识别效果，微调给基础版测试集带来的提升比拓展版更为显著。对于未经微调的模型来说，基于四库全书训练的模型效果普遍优于基于殆知阁语料训练的模型。

3.3 实验分析

由前文实验结果可见，对现有基线模型来说，通假字检测和正字识别均为十分具有挑战性的任务，拓展版评测集的难度大大高于基础版。为了进一步探析模型的识别和泛化能力，我们

序号	模型	准确率（基础版）	准确率（拓展版）
1	DaizhigeBigram	34.55%	13.18%
2	SikuBigram	40.69%	14.31%
3	DaizhigeTrigram	33.11%	11.38%
4	SikuTrigram	30.71%	9.93%
5	DaizhigeGPT2	40.79%	13.99%
6	SikuGPT2	43.38%	14.78%
7	DaizhigeBert	35.22%	12.97%
8	SikuBert	42.32%	14.90%
9	ZhengziRecognitionDaizhigeBert	65.64%	19.88%
10	ZhengziRecognitionSikuBert	61.61%	18.96%

表 7. 正字识别任务实验结果

将拓展版测试集按照目标字是否在训练集中收录分为两部分，分别计算了通假字检测和正字识别的实验结果，分别如表 8 和表 9 所示。

通假字分类	字数	精确率	召回率	F1
常见通假字（有训练数据）	279	26.22%	68.10%	37.86%
拓展通假字（无训练数据）	2200	37.60%	7.14%	12.00%
全部通假字	2479	29.48%	16.53%	21.18%

表 8. TongjiaziDetectionSikuBert（无混淆集）模型的通假字检测任务（拓展版）实验结果

通假字分类	字数	准确率
常见通假字（有训练数据）	279	58.97%
拓展通假字（无训练数据）	2200	11.68%
全部通假字	2479	18.96%

表 9. ZhengziRecognitionSikuBert模型的正字识别任务（扩展版）实验结果

如对于通假字检测任务来说，据表 8 可以发现：首先，对于训练数据中未出现的通假字，模型也可以检测出来一部分，并且具有较高精确率，这说明模型具有一定的泛化能力，能够探测出少量训练阶段未见过的通假用法，如例3中的“考”字。第二，对于训练数据收录的常见通假字，模型探测的召回率较高，但精确率却不理想，经过进一步地分析，发现主要有两点原因：(1) 模型倾向于将在训练数据中见过的通假字的非通假用法也判定为通假字，如例4中的“皇”字；(2) 模型实际预测正确，《汉语大词典》中的例句仅针对字头标注通假用法，句中还可能包括其他通假字，数据标注存在少量缺失情况，如例5中的“皇皇”。

例3. 陳登者，善術，夜過吉甫家，即捕登掠**考**，上言吉甫陰事。（“考”通“拷”，“考”字通假用法在训练集中未出现，模型正确预测其为通假字）

例4. 真宗**皇**帝之嘉嘆，面可其奏。（训练集中收录了“皇”的通假用法，但此处“皇”字并非通假，模型错误预测其为通假字）

例5. 孔子三月無君，則**皇皇**如也，出疆必載質。（此处“皇”通“惶”，模型正确预测其为通假字，但由于该句取自《汉语大词典中》“質”通“贄”的例句，其中“皇”的通假用法未被标注，导致评测时误将此例计为误探测条目。）

在正字识别任务中，如表 9 所示，ZhengziRecognitionSikuBert模型同样具有一定的泛化能力。对于训练数据中未覆盖的通假字，来自通假字知识库的混淆集发挥了作用，帮助模型将它们识别了出来。对于未识别出的正字，经分析，发现主要包括两种错误类型：第一，模型认为该位置填通假字比填正字更合适，如表 10 所示，在识别句中“台”的正字时，只有常见的通“鮐”被成功识别，而相对罕见的通“嗣”之用法则未被识别；第二，一个通假字对应着多个正

评测语料	训练数据覆盖	通假字	正字	识别结果
黃耇台背，以引以翼。	否	台	鮐	鮐（正确）
黃髮台背，壽胥與試。	否	台	鮐	鮐（正确）
有于德不台淵穆之讓，靡號師矢敦奮撝之容。	否	台	嗣	台（错误）
聖人共手，時幾將矣。	是	共	拱	拱（正确）
非吾所以共承宗廟意也。	是	共	恭	恭（正确）
唯是桃弧、棘矢以共禦王事。	是	共	恭	供（错误）

表 10. 拓展评测集上的正字识别结果示例

字，模型错误地识别为其他正字，例如，在识别“共”的正字时，存在通“恭”和通“供”两种通假用法，模型将部分通“恭”用法识别为了通“供”，如表 10 中的最后一例:“唯是桃弧、棘矢以共王事。”进一步查阅文献发现，不同学者对通假释读方式存在差异：唐代陆德明《经典释文》注此句中“共”音“恭”，成为清代之前学者共识，《汉语大词典》亦用此说。而以清代俞樾《群经平议》为代表的清人观点认为该字通“供”，并为现代人所继承，如杨伯峻《春秋左传注》、中华书局版《左传》(郭丹等译注)皆同此观点。可见，模型判定虽不同于“标准答案”，但有其合理之处。

总之，通假字的检测和识别是一个复杂的的问题，本文搭建的基线模型能够识别部分通假用法，但泛化能力尚显不足，对微调训练时未能覆盖的通假字，往往无法检测到或准确识别出本字。在识别本字时，对于不常见的通假关系，模型也往往无法正确识别。未来我们仍需要在设计模型时充分集合上下文语义信息与通假字、正字的释义信息，提升模型泛化能力，加强其对不常见通假关系的识别能力。

4 总结

通假是古汉语中的一种常见用字现象，为了服务于通假字的人工判别和机器处理，本文构建了一个涵盖标注语料库、知识库和评测数据集的多维度通假字资源库。在此基础上，本文基于Ngram、BERT、GPT2等主流语言模型开展了通假字自动检测和正字识别实验，为通假字检测和正字识别任务提供了基线结果：在收录常见通假字用法的基础版测试集上，通假字检测的F1值达到66.94%，正字识别的准确率达到65.64%；在拓展版测试集上，模型具备一定泛化能力，能够识别出少量在训练集中未见过的通假字及其正字，但识别效果远远低于基础版评测集。通过对比不同的基线模型，本文发现，模型结构、预训练数据、微调机制和混淆集的使用均会对两个子任务产生不同程度的影响。进一步地，本文对模型的预测误例及原因进行了初步分析。

需要指出的是，本文所开展的通假字资源库建设和通假字识别算法的研究只是该领域的初步探索性工作，研究还存在不少待改进之处。例如，在资源库的建设上，本研究基于《汉语大词典》采集基础性标注语料，词典仅针对字头给出通假例句，例句中仍可能存在其他通假字，有待在后续工作中通过人工标注进行补充；同时，《汉语大词典》所收录的通假用法旨在覆盖基础性、一般性需求，未来还有必要基于面向出土文献和传世文献的通假字辞书资源引入更大范围的通假用例数据，对现有的语料库和知识库进行扩充，从而更好地辅助汉语史领域的相关研究。在自动识别技术上，本研究搭建了通假字检测和正字识别的基线方法，由实验结果可见，通假字检测和正字识别是极具挑战性的自然语言处理任务，目前模型具有一定识别能力，但其准确性和泛化能力还有待进一步提升。此外，基于ChatGPT、GPT4等大模型开展通假字识别是一个值得探索的方向。

最后，在资源库和识别技术的应用上，仍有不少可以开展的工作。例如，通假字资源库及识别算法可以接入古籍整理或古文献检索平台，为该领域研究者提供可能的通假字用例及相关语料信息，辅助专家释读文献，提升古籍整理效率。如前文所述，基于图结构的知识库能够提供传统辞书无法呈现的大规模汉字通假关系网络信息，从而可为古汉语字用现象、词汇发展、词义关联、语音演变等研究提供新视角、新方法。此外，资源库中的高频常用通假字数据可以为文言文教学材料编写、考试命题提供参考，基于该库和其他古汉语领域现有语言资源（如词性标注语料库、词义标注语料库、文白翻译平行语料库等）还可进一步研发辅助文言文学习的工具应用，提升学生的文言文阅读理解能力。

第二十二届中国计算语言学大会论文集，第535页-第546页，哈尔滨，中国，2023年8月3日至5日。

致谢

本研究得到国家语委重大项目“古籍整理智能化关键技术研究”(ZDA145-9)、国家自然科学基金青年项目“面向古籍整理智能化的知识表示与加工研究”(62006021)、北京市社科重点项目“古典文献的智能化分析与关联技术研究”(21DTR037)资助。北京师范大学李隽琪、陈青、孟琢等师友为资源库设计提出了宝贵的建议，在此表示感谢。

参考文献

党怀兴. 1998. 通假成因说略. 陕西师范大学学报：哲学社会科学版, (1):61–65.

邓三鸿, 胡昊天, 王昊, and 王东波. 2021. 古文自动处理研究现状与新时代发展趋势展望. 科技情报研究, 3(1):1–20.

胡韧奋, 李绅, and 诸雨辰. 2021. 基于深层语言模型的古汉语知识表示及自动断句研究. 中文信息学报, 35(4):8–15.

胡韧奋, 曹冰, and 杜健一. 2013. 现代汉字形声字声符在普通话中的表音度测查. 中文信息学报, 27(3):41–48.

孔德明. 1993. 通假字概说. 北京广播学院出版社.

Dayiheng Liu, Kexin Yang, Qian Qu, and Jiancheng Lv. 2019. Ancient–modern chinese translation with a new large training dataset. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 19(1):1–13.

柳建钰and 周晓文. 2017. 计算机辅助古籍版本校勘资源库建设浅议. 图书馆理论与实践, (3):54–58.

钱玄. 1980. 秦汉帛书简牍中的通借字. 南京师大学报(社会科学版), (3):44–48.

舒蕾, 郭懿鸾, 王慧萍, 张学涛, and 胡韧奋. 2022. 古汉语词义标注语料库的构建及应用研究. 中文信息学报, 36(5):21–30.

孙建伟. 2015. 假借和通假研究综论. 宁夏大学学报(人文社会科学版), (2):29–33.

苏祺, 胡韧奋, 诸雨辰, 严承希, and 王军. 2021. 古籍数字化关键技术评述. 数字人文研究, 1(3):83.

王宁. 2012. 古代汉语. 高等教育出版社.

Zinong Yang, Ke-jia Chen, and Jingqiang Chen. 2021. Guwen-unilm: Machine translation between ancient and modern chinese based on pre-trained models. In Natural Language Processing and Chinese Computing: 10th CCF International Conference, NLPCC 2021, Qingdao, China, October 13–17, 2021, Proceedings, Part I 10, pages 116–128. Springer.

Xiaoyong Yan, Ying Fan, Zengru Di, Shlomo Havlin, and Jinshan Wu. 2013. Efficient learning strategy of chinese characters based on network approach. PloS one, 8(8):e69745.

Xiaoyuan Yi, Maosong Sun, Ruoyu Li, and Wenhao Li. 2018. Automatic poetry generation with mutual reinforcement learning. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pages 3143–3153.

由明智. 2013. 谈人教版中学语文教材的通假字注释. 课程．教材．教法, 33(9):46–50.

张儒. 1988. 关于竹书、帛书通假字的考察. 山西大学学报：哲学社会科学版, (2):37–43+113–114.

附录A. 通假字知识库体例

通假字知识库采用图数据结构，以汉字为节点(node)，字节点之间有通假关系和形声关系两类连边(edge)，节点、边及其属性均以JSON Object形式存储。通假关系边属性会引用语料信息，这些语料没有像“通假字标注语料库”中的语料那样经过详细的标注与校对，只是将不同来源的语料去重后，解析为简单的结构化对象并存储。

字节点具有以下五个属性：

1. 节点ID：用于唯一标识字对象的编号，如“248”、“1764”。
2. 字形：字的书写形态，如“辟”、“譬”。
3. 部首：汉字的构造部分，用于分类和检索字，如“辛”、“言”。
4. 部件：汉字的基本构成单元，包括部首和其他部分，如“卩口辛”、“辟言”。
5. 结构：汉字的构造方式，如“左右结构”、“上下结构”等。

通假关系边具有以下八个属性：

1. 通假字关系ID：用于唯一标识通假关系对象的编号，如“638”。
2. 通假字：在该通假关系中通其他字的字，是有向边的起点，如“辟通譬”通假关系中的“辟”。
3. 正字：被通假的字，是有向边的终点，如“辟通譬”通假关系中的“譬”。
4. 拼音：该通假关系中字音的拼音表示，如“pì”。
5. 注音：该通假关系中字音的注音表示，如“ㄆㄧˋ”。
6. 古音：该通假关系中字音的古代发音，如“《廣韻》芳辟切，入昔，滂。《廣韻》房益切，入昔，並。”。
7. 释义：该通假关系中字的意义或用法解释，如“墨子提出的逻辑推理的方法之一。谓举旁例以喻所说的论题。”。
8. 关联语料ID：与通假关系对象相关的语料对象的编号列表，用逗号分隔，如“8440, 8804”。

形声关系边具有以下三个属性：

1. 形声关系ID：用于唯一标识形声关系对象的编号，如“644”。
2. 形声字：具有特定形声构造的汉字，是有向边的起点，如“譬”。
3. 声旁：形声字的声旁，是有向边的终点，如“辟”。

关联语料具有以下四个属性：

1. 语料ID：用于唯一标识语料对象的编号，如“8806”。
2. 语料文本：包含通假字与通假关系的文本内容，如“是過者也，過猶不及也；辟之是猶立直木而求其景之枉也。”。
3. 语料出处：语料的来源文献，如“《荀子·王霸》”。
4. 语料来源：语料的来源，为“汉语大词典”、“汉典”或“国学大师网汉语字典”。

第二十二届中国计算语言学大会论文集，第535页-第546页，哈尔滨，中国，2023年8月3日至5日。