

马王堆医书信息化和智能化研究展望^{*}

卢彦杰 盛 威 刘 伟

(湖南中医药大学信息科学与工程学院 长沙 410208)

〔摘要〕 目的/意义 展望马王堆医书信息化和智能化研究的方向和技术路线,为相关研究提供参考。方法/过程 综述马王堆医书的研究历程,分析已有研究存在的问题,探讨马王堆医书信息化和智能化研究中3个研究方向的可能性及技术路线。结果/结论 可从文本语料库构建、古医书修复、知识抽取3个方向,研究将人工智能技术与传统古籍相结合的可能性,充分发挥马王堆医书中蕴含的宝贵价值。

〔关键词〕 马王堆医书;古医书修复;知识抽取;信息化;智能化

〔中图分类号〕 R-058 〔文献标识码〕 A 〔DOI〕 10.3969/j.issn.1673-6036.2024.03.016

Research Prospect of Informatization and Intelligence of Mawangdui Medical Books

LU Yanjie, SHENG Wei, LIU Wei

School of Informatics, Hunan University of Chinese Medicine, Changsha 410208, China

〔Abstract〕 **Purpose/Significance** The research direction and technical route of Mawangdui medical books informatization and intelligence are prospected to provide references for related research. **Method/Process** The paper reviews the research process of Mawangdui medical books, analyzes the problems existing in the existing studies, and probes into the possibility and technical route of three directions in the research of informatization and intelligence of Mawangdui medical books. **Result/Conclusion** The possibility of combining artificial intelligence technology with traditional ancient books can be studied from three directions; the construction of text corpus, the restoration of ancient medical books and knowledge extraction, so as to give full play to the value contained in Mawangdui medical books.

〔Keywords〕 Mawangdui medical books; restoration of ancient medical books; knowledge extraction; informatization; intelligence

1 引言

1973 年湖南省长沙马王堆三号西汉古墓出土

14 部古医书,迅速引起考古、医史文献、中医药、古汉语文学等众多领域学者的高度重视^[1-2]。马王堆医书在中医中药、养生保健、文化传承、语言文字等方面具有重要意义和价值。50 余年来,围绕马王堆医书开展了大量研究,但仍有很多问题尚未解决,很多先进方法尚未充分应用。出土医书与经典传世医书的互动,新、旧出土医书的互动,是未来医学研究趋势,也是马王堆医书研究的着力点。

由于古汉语文本的整体信息化程度较低,古汉语特别是医古文与信息处理技术缺乏有机结合,关于马王堆医书的信息化和智能化研究鲜有

〔修回日期〕 2023-08-03

〔作者简介〕 卢彦杰,硕士研究生,发表论文9篇;通信作者:刘伟,教授,硕士生导师。

〔基金项目〕 湖南省中医药科研项目(项目编号:B2023039);湖南省长沙市自然科学基金项目(项目编号:kq2022260);湖南中医药大学研究生创新课题项目(项目编号:2022CX121)。

报道。本文基于马王堆医书兼具的古汉语复杂特征和医学文献特点,探讨其信息化和智能化过程中的研究方法,研究将现代人工智能(artificial intelligence, AI)技术应用于传统古籍研究的可能性,不仅可以充分发挥马王堆医书中蕴含的宝贵价值,而且可以为其他古籍的数字化提供参考。

2 国内外研究现状

2.1 古汉语语料库构建

古汉语信息处理是中文信息处理领域的重要研究方向之一,对古籍文献进行数字化和信息化的首要任务就是构建古汉语语料库。严顺^[1]初步探究先秦文献语料库构建,阐述古文献语料库构建过程中涉及的古汉语分词以及词性标注等基本知识。在构建古籍语料库时需要制定对应的分词规范和标注集,南京师范大学语言科技研究所制定了先秦汉语词类标注基本集并构建先秦典籍人工语料库^[2]。留金腾等^[3]以《淮南子》为例介绍上古汉语分词及词性标注语料库及其构建过程。石民等^[4]运用条件随机场(conditional random fields, CRF)方法研究先秦汉语的分词和标注一体化工作。但上述研究未充分考虑中医古籍特点。白玲玲^[5]阐述构建中医药古文献语料库的意义和思路,详细介绍语料分类问题。付璐等^[6]探索性研究和总结了中医古籍的分词规范,提出中医古籍分词规范建议,并以人工标注方式构建一个小型的清代医籍分词语料库,但未能实现计算机自动分词,且语料库规模较小,不能推广和应用。目前暂无关于马王堆医书语料库构建方面的研究成果报道,也尚无通过信息技术分析马王堆医书和传世中医典籍之间相互关系的研究。

2.2 古籍智能修复

马王堆出土医书共包含 14 种古代医书,其中帛书 10 种,竹木简 4 种,抄写年代为秦汉之际。由于年代久远,帛书残损非常严重,有诸多帛书碎片缺失,导致古医书文字残缺不全。对这些残

字的考释和修复是马王堆医书研究人员面临的重要问题,也是充分抽取其中蕴含的中医药知识的前提。近年来,有学者开始尝试将深度学习技术应用于古文本的修复并取得较好效果。2019 年牛津大学与 DeepMind 公司合作开展古文修复研究,联合开发名为 Pythia 的 AI 工具,利用深度学习神经网络技术识别和预测古希腊碑文中缺少的单词或字符,取得较好效果^[7]。2020 年 9 月国际顶级期刊 PNAS 发表了几位以色列学者运用深度学习中的循环神经网络恢复残缺古巴比伦文本的研究成果,实验结果表明该模型在预测残缺文本时效果良好^[8]。残缺图像的修复也成为 AI 和计算机视觉领域的一项重要研究内容,以深度学习为代表的机器学习技术在图像修复领域得到广泛应用^[9-10],但是关于帛书碎片的智能拼接和修复暂无相关报道。2014 年出版的《长沙马王堆汉墓简帛集成》一书发布了已有马王堆医书帛书残片的全部高清照片,通过 AI 技术可以设计图像拼接和图像修复模型,在一定程度上将降低简帛修复专家和研究人員修复古医书的难度,为后续马王堆医书的知识抽取和进一步研究提供便利。

2.3 知识抽取

知识抽取是 AI 的重要组成部分之一,也是构建知识图谱的基础。知识抽取包含命名实体识别、实体关系抽取和事件抽取 3 个核心任务。中医古籍处理难度远远大于现代汉语文本,主要难点是其中术语多且存在大量通假字、生僻字、一词多义等,部分句式结构复杂,且缺少标点和分隔符,中医术语缺乏统一的标准等。王世昆等^[11]基于 CRF 研究明清古医案中症状、病机的自动识别标注问题,提出数据清洗和缩减合并词性以减少特征空间规模。孟洪宇等^[12]基于 CRF 针对《伤寒论》文本中的症状、病名、脉象、方剂等中医术语进行识别,通过结合字本身、词性、词边界、术语类别标注的特征,分析不同特征组合对术语识别的影响,探讨最具有有效性的组合,该方法虽无需构建特有模式,但严重依赖人工构建特征的准确度,需大量标注语料,且受分词效果制约。张艺品等^[13]和高甦等^[14]

采用基于双向长短期记忆神经网络（bidirectional long short-term memory, BiLSTM）和 CRF 的实体识别模型对《黄帝内经》《备急千金方》《神农本草经》等中医典籍实体进行识别，获得不错的识别效果。朱玲等^[15]以文本信息抽取为关键技术，以关键词为节点，实现中医古籍中语义关系的半自动发现，将该方法运用于综合性医学专著《医学纲目》，可用于充实和细化中医古籍语言系统中的语义关系。叶辉等^[16]采用 CRF 结合词性和中医诊断标记集抽取《金匱要略》中的症状药物信息，构建相应的自动标签标注模型，结果表明多特征模型可以有效提高 CRF 算法对中医古籍的实体抽取能力。2019 年，微软亚洲研究院和清华大学联合提出一种新的基于卷积神经网络（convolutional neural networks, CNN）的命名实体识别方法：CNN-LSTM-CRF，可以进一步提高命名实体识别的准确率^[17]，具有良好的应用前景。

2.4 知识表示

构建中医古籍知识库并可视化知识是中医古籍知识工程研究的一个重要领域。柳长华^[18]最早将基于知识元的知识管理理论引入中医典籍数字化研究，提出基于知识元的中医典籍计算机知识表示方法。此后，基于知识元的中医古籍数字化陆续展开。丁侃^[19]以中医古籍方剂知识元为研究核心，通过对古籍中方剂知识元的抽取与标引，试图构建以知识元为单元的知识组织体系，通过对语义关系的标引，关联起中医古籍方剂的知识语义网络，从而构建起以知识元为核心的知识表示体系，并尝试采用形式化的手段进行表达。但是该研究只对方剂构成的部分知识元进行基于语义关系的标引与形式化表达研究，缺乏系统性与完整性。朱玲等^[20]从古籍文献中人工抽取出崩漏的病因病机、并发症、疾病治疗等相关语义并构建语义网络，开发一个简单版的知识抽取及可视化平台。近年来，知识图谱技术在中医药领域得到较为广泛的应用^[21-22]。知识图谱的本质是一种结构化的知识表示形式，涉及知识表示、知识存储、知识抽取、知识融合、知识推理、问答和分

析等多方面的技术要素，广泛应用于辅助搜索、辅助问答、辅助大数据分析、辅助语言理解以及辅助设备互联等方面。贾李蓉等^[23]从数据来源、研究内容、图形化展示等方面探讨如何构建中医药知识图谱。张德政等^[24]提出一种基于本体的中医核心知识图谱表示及其构建方法。李兵等^[25]对当前中医古籍知识化研究现状和常用的知识分析与挖掘方法进行综述，梳理中医古籍知识深度利用与知识库构建中存在的一些问题，同时对中医古籍知识的深度开发和利用提出对策，其中语料库的构建、知识的概念与语义关联、专题特色知识库的构建等都是中医古籍知识发现与知识服务的重要发展方向。

3 讨论

3.1 马王堆医书文本语料库构建研究展望

基于已有的中医典籍语料库和南京师范大学语言科技研究所的先秦典籍语料库，结合统计机器学习和深度学习技术，构建马王堆医书自动标注模型，对医书释文开展全面整理和深入分析；采用人工标注和机器自动标注相结合的方式，对马王堆医书文本开展分词、词性和词类标注，构建一个系统和完整的马王堆医书文本语料库，达到分词和词性标注级别，为后续研究奠定数据基础。具体技术路线，见图 1。

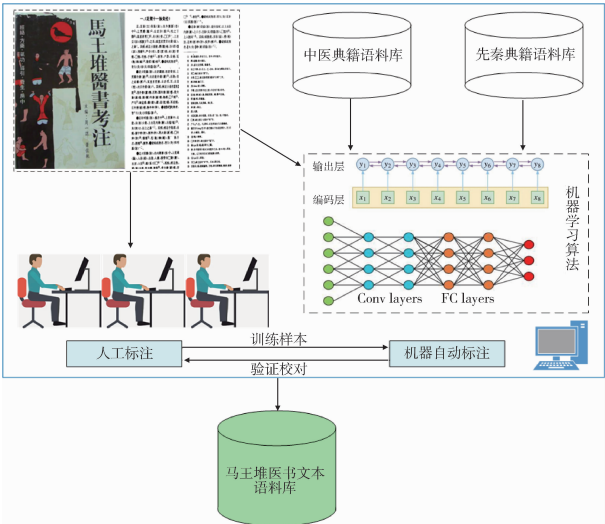


图 1 马王堆古医书文本语料库构建技术路线

3.2 马王堆医书智能修复研究展望

马王堆医书智能修复分为古医书残损文字修复和帛书图像修复。残损文字修复可以基于已有的《黄帝内经》等中医典籍语料库和先秦典籍语料库，结合所构建的马王堆医书文本语料库以及湖北江陵张家山医简、甘肃武威汉代医简等出土秦汉简帛文献文本库，运用传统统计机器学习方法和深度学习算法相结合的方式，修复和补全马王堆简帛医书中部分残损文字，构建一套具有普适性的古医书缺失文本修复方法。具体技术路线，见图 2。

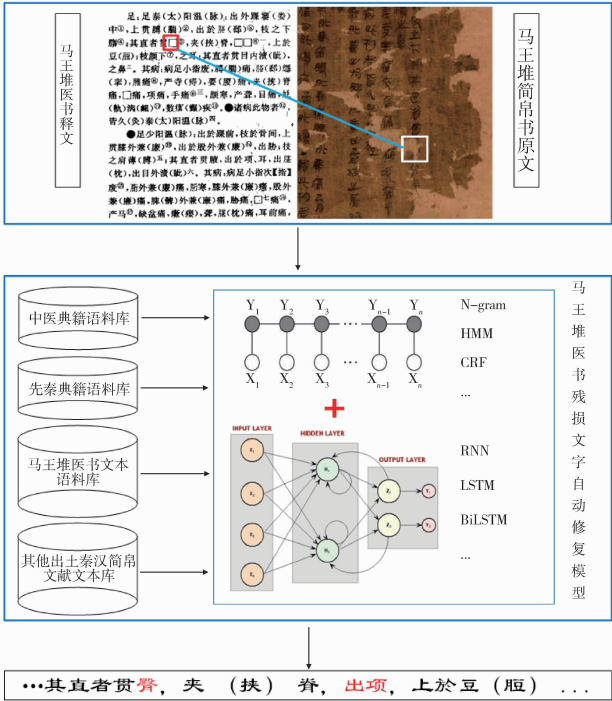


图 2 马王堆医书残损文字修复研究技术路线

深度学习中的语言模型技术可以根据文本的上下文信息预测，例如“九针之<mask>，各不同形”，“<mask>”代表缺失的内容，语言模型根据“<mask>”的上下文信息对“<mask>”进行预测，输出此处“<mask>”为“名”的概率最高。基于语言模型技术栈中 N 元（N - gram）模型、循环神经网络（recurrent neural network，RNN）模型、长短期记忆网络（long short - term memory，LSTM）模型、双向长短期记忆网络模型等技术，探讨适用于复杂医学古籍的文本补齐方法，降低古籍文本修

复专家考证古籍缺失文本的难度。帛书图像修复可以基于《长沙马王堆汉墓简帛集成》中马王堆医书帛书以及相关残片的全部高清照片，包括《五十二病方》《足臂十一脉灸经》《阴阳十一脉灸经》等古医书。随机分割已有的较完整的帛书，构造帛书图像拼接和修复训练数据集，提出一种多线索帛书残片图像拼接方法，基于残片边缘、文字笔画、帛书纹理等多个线索开展拼接研究，并结合深度学习技术整体评估拼接结果，提升拼接结果的准确率。对于拼接后存在缺损的帛书图像，采用卷积神经网络和生成对抗网络开展图像修复研究，在一定程度上修复存在残损的帛书医书，构建具有普适性的帛书图像修复模型。具体技术路线，见图 3。



图 3 马王堆医书帛书图像修复研究技术路线

3.3 马王堆医书知识抽取研究展望

马王堆医书知识抽取研究可以基于深度学习中的卷积神经网络和循环神经网络等技术，充分根据中医古籍特别是马王堆医书本身的特点，从字形（偏旁部首等）、字音（同音字等）、字义（古今异义和同义等）等多个维度对马王堆医书命名实体识别、命名实体链接和实体关系抽取开展研究，并改进已有方法，提出更适合中医典籍的通用处理方

法。基于深度学习的马王堆医书命名实体识别研究的技术路线，见图4。基于已抽取的命名实体、实体链接和实体关系，通过实体链接解决实体冲突和指向不明等问题，运用Python语言和开源Python库spaCy分析文本，再通过Neo4j构建图数据库，利用Cypher语言将结构化数据存储到Neo4j图数据库中，并使用Py2neo工具在Python中实现对Neo4j的管理，最后运用可视化库D3.js实现知识图谱的可视化，形成《五十二病方》《足臂十一脉灸经》《阴阳十一脉灸经》《十问》等马王堆医书的知识图谱，并与已有的《脉经》《黄帝内经》等医学古籍知识库链接，构建内容更加丰富且完备的知识推理和知识应用系统。马王堆医书知识图谱构建技术路线，见图5。

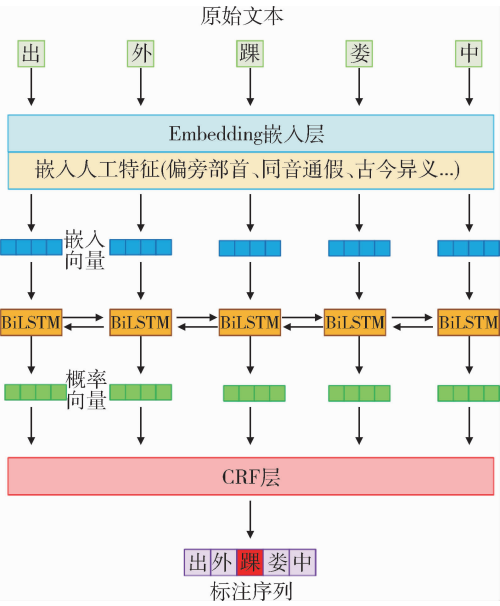


图4 基于深度学习的马王堆医书命名实体识别研究技术路线

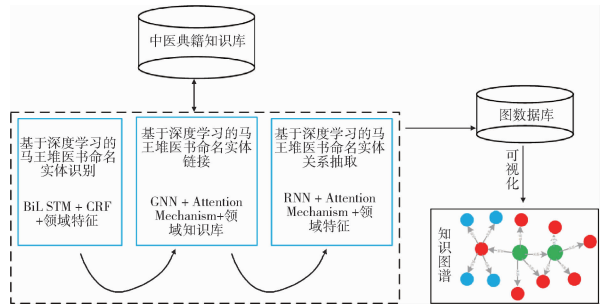


图5 马王堆医书知识图谱构建技术路线

4 结语

围绕马王堆医书、运用AI技术开展智能化和信息化研究，具有重要意义和价值，但也存在一些困难和挑战。首先是研究团队的人员配置问题，为了保证马王堆医书语料库构建的质量，知识抽取的专业性和准确性，研究团队中既需要有中医文献和典籍研究人员，又需要有深度学习和知识工程研究人员。团队成员需要具有较深厚的古汉语功底和中医知识背景，熟悉医古文的构词特点和文法规则，又掌握主流的深度学习框架和算法以及知识网络和图谱构建技术。其次，文本修复、图像修复、知识抽取3个阶段的数据集都需要团队自行整理构建，即使采取机器辅助标注，也是十分耗时耗力的过程。再者，将深度学习方法与马王堆医书的复杂特征有机结合，并在此过程中进行方法创新、应用，形成一套具有普适性的古籍信息化和智能化流程也具有一定难度。最后，深度学习模型的训练离不开强大的算力支持，购置满足一定条件的服务器需要相应经费支持。

国家高度重视中医药事业的发展，关于中医药的学术研究，特别是关于中医药的继承和创新研究已成为国内外关注的焦点问题。长沙马王堆出土的古医书既是一个中医药宝库，也是中国文化的重要代表之一。马王堆医书的信息化和智能化研究旨在运用AI技术，特别是近年来发展迅速的深度学习和知识图谱技术，为研究者提供高质量的文本语料库和系统化的知识图谱，降低马王堆简帛修复专家和研究人員修复古医书的难度，并为中医典籍研究者提供抽取知识的新方法，以可视化图谱的方式让更多人了解并熟悉马王堆医书中包含的中医、中药、养生和保健等知识，推广马王堆中医文化，促进中医药的传承和创新。

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 严顺. 先秦文献的语料库构建探究 [J]. 江苏科技信息, 2016, 489 (12): 32-33.

- 2 石民. 先秦汉语自动分词及词性标注研究 [D]. 南京: 南京师范大学, 2010.
- 3 留金腾, 宋彦, 夏飞. 上古汉语分词及词性标注语料库的构建——以《淮南子》为范例 [J]. 中文信息学报, 2013, 27 (6): 6–15, 81.
- 4 石民, 李斌, 陈小荷. 基于 CRF 的先秦汉语分词标注一体化研究 [J]. 中文信息学报, 2010, 24 (2): 39–45.
- 5 白玲玲. 中医药古文文献语料库建设的语料分类问题研究 [D]. 济南: 山东中医药大学, 2007.
- 6 付璐, 李思, 李明正, 等. 以清代医籍为例探讨中医古籍分词规范标准 [J]. 中华中医药杂志, 2018, 33 (10): 4700–4705.
- 7 ASSAEL Y, SOMMERSCHIED T, SHILLINGFORD B, et al. Restoring and attributing ancient texts using deep neural networks [J]. Nature, 2022, 603 (7900): 280–283.
- 8 FETAYA E, LIFSHITZ Y, AARON E, et al. Restoration of fragmentary Babylonian texts using recurrent neural networks [J]. Proceedings of the national academy of sciences, 2020, 117 (37): 22743–22751.
- 9 强振平, 何丽波, 陈旭, 等. 深度学习图像修复方法综述 [J]. 中国图象图形学报, 2019, 24 (3): 447–463.
- 10 赵露露, 沈玲, 洪日昌. 图像修复研究进展综述 [J]. 计算机科学, 2021, 48 (3): 14–26.
- 11 王世昆, 李绍滋, 陈彤生. 基于条件随机场的中医命名实体识别 [J]. 厦门大学学报 (自然科学版), 2009, 48 (3): 359–364.
- 12 孟洪宇, 谢晴宇, 常虹, 等. 基于条件随机场的《伤寒论》中医术语自动识别 [J]. 北京中医药大学学报, 2015, 38 (9): 587–590.
- 13 张艺品, 关贝, 吕荫润, 等. 深度学习基础上的中医实体抽取方法研究 [J]. 医学信息学杂志, 2019, 40 (2): 62–67.
- 14 高甦, 金佩, 张德政. 基于深度学习的中医典籍命名实体识别研究 [J]. 情报工程, 2019, 5 (1): 113–123.
- 15 朱玲, 于彤, 杨峰. 基于关键动词的中医古籍概念实体间语义关系发现研究 [J]. 中国数字医学, 2016, 11 (5): 73–75.
- 16 叶辉, 姬东鸿. 基于多特征条件随机场的《金匱要略》症状药物信息抽取研究 [J]. 中国中医药图书情报杂志, 2016, 40 (5): 14–17.
- 17 WU F, LIU J, WU C, et al. Neural Chinese named entity recognition via CNN–LSTM–CRF and joint training with word segmentation [EB/OL]. [2023–05–03]. <https://arxiv.org/pdf/1905.01964.pdf>.
- 18 柳长华. 基于知识元的中医古籍计算机知识表示方法 [C]. 武汉: 中国生物医学工程学会第六次会员代表大会暨学术会议, 2004.
- 19 丁侃. 基于知识元的中医典籍方剂知识表示研究 [D]. 北京: 中国中医科学院, 2012.
- 20 朱玲, 朱彦, 杨峰. 基于中医疾病相关语义关系的正则表达式及知识抽取研究 [J]. 世界科学技术: 中医药现代化, 2016, 18 (8): 1241–1250.
- 21 李新龙, 刘岩, 何丽云, 等. 知识图谱研究概况及其在中医药领域的应用 [J]. 中国中医药信息杂志, 2017, 24 (7): 129–132.
- 22 孙华君, 李海燕, 聂莹, 等. 知识图谱及其在中医药领域应用研究进展 [J]. 世界科学技术–中医药现代化, 2020, 22 (6): 1969–1974.
- 23 贾李蓉, 刘静, 于彤, 等. 中医药知识图谱构建 [J]. 医学信息学杂志, 2015, 36 (8): 51–53, 59.
- 24 张德政, 谢永红, 李曼, 等. 基于本体的中医知识图谱构建 [J]. 情报工程, 2017, 3 (1): 35–42.
- 25 李兵, 张华敏, 李莎莎, 等. 中医古籍知识深度利用方法与知识库构建 [J]. 中国数字医学, 2018, 13 (8): 33–35.

敬告作者

《医学信息学杂志》网站现已开通, 投稿作者请登录期刊网站: <http://www.yxxxx.ac.cn>, 在线注册并投稿。

《医学信息学杂志》编辑部