命名实体识别在中医药领域的研究进展

沈蓉蓉[1]，夏帅帅[2]，晏峻峰[1*]
（[1]湖南中医药大学信息科学与工程学院，长沙 410208；2 湖南中医药大学中医诊断研究所，长沙 410208）

摘要：本文阐述了中医药领域命名实体识别的研究进程，总结了中医药领域命名实体识别的数据特征、评价指标和应用情况，指出了当前中医药领域命名实体识别研究的难点和不足，并对未来中医药领域命名实体识别的发展进行展望，旨在为未来中医药领域命名实体识别研究提供参考。

关键词：中医药；命名实体识别；知识抽取；综述

中图法分类号：TP391.1; R319

Review on Research of Named Entity Recognition in Chinese Medicine

Shen Rongrong[1], Xia Shuaishuai[2], Yan Junfeng[1]*
（1 School of Informatics, Hunan University of Chinese Medicine, Changsha, Hunan, 410208, China; 2 Provincial Key Laboratory of TCM Diagnostics, Hunan University of Chinese medicine, Changsha, Hunan Province, 410208）

Abstract: This paper described the research methods and contents of named entity recognition (NER) in the field of traditional Chinese Medicine (TCM), summarized the data features, evaluation criteria and application development in the research of NER in TCM. Finally, we pointed out the difficulties and shortcomings of the current research of NER in TCM, and made a prospect, aiming to provide reference for the future research of named entity recognition in TCM.

Keyword: Chinese Medicine; Named Entity Recognition; Knowledge Extraction; Review

实体作为物质个体的哲学概念被引入计算机科学领域，命名实体识别(Named Entity Recognition, NER)即利用自然语言处理技术识别文本中具有特定意义的实体。中医药领域命名实体识别是 NER 在中医药领域的应用研究，是指识别中医药文本中具有特定意义和一定研究价值的实体，如“感冒”“气虚证”“咳嗽”“人参”等。中医药实体是构成中医药文本的基本元素，对其内在关系的探索是中医药研究的本质工作。早期，中医药科学研究多以人工方式提取中医药文献、中医临床诊疗记录等非结构化文本中的关键信息，效率低、代价高，难以满足中医药领域的发展要求。随着自然语言处理技术的飞速发展，

1基金项目：the National Key R&D Program of China (No.2018AAA0102100)；湖南省教育厅科学研究重点项目（18A219）；湖南省卫健委委托项目（HNZH-DY-20191217001）；科技创新 2030—“新一代人工智能" 重大项目(批准号：2018AAA0102100)作者简介：沈蓉蓉（1995-），女，江苏盐城，硕士研究生，主要研究方向为大数据与计算中医学；晏峻峰（1965-），女，湖南长沙，教授，博导，博士，主要研究方向为大数据与计算中医学，E-mail: junfengyan@hnucm.edu.cn。

NER 和中医药领域的结合研究逐渐兴起，NER 以机器代替人工提取中医药实体，降低了时间和人力成本，加快了中医药领域研究的研究效率。

目前，NER 已广泛用于中医药文献研究、基础研究、临床研究等，并为中医知识库、知识图谱、智能问答等研究带来了助力。基于 NER 巨大的研究价值及在中医药领域的众多研究成果，本文对中医药领域 NER 相关研究进行梳理，以期为后续中医药领域 NER 研究提供参考。

1. 中医药领域 NER 概况

1.1 中医药实体类型 识别实体的边界和预测实体的类别是 NER 的本质工作。在中医药领域，命名实体可以是时间、地名等基本实体，也可以是疾病、证型等领域实体。检索中国知网及 Google Schola 数据库，获取主题与“中医药命名实体识别”密切相关的文献。以人工阅读全文的方式对文献进行筛选，剔除重复及非实验类文献，对剩余 68 篇文献的发表时间、相关技术方法、中医药实体类型等信息进行统计。由图 1 可见，中医药实体主要包括症状、中药、疾病等 21 类，其中症状的研究占比最高，为 24.06%。

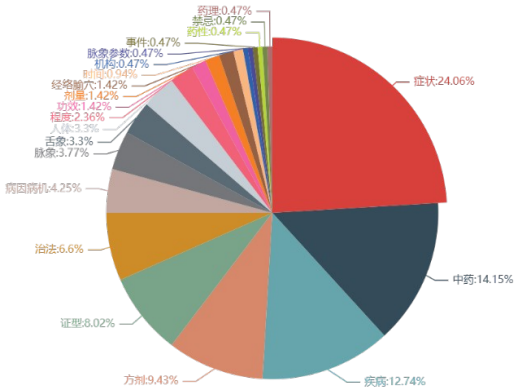

图 1 中医药领域命名实体类型

1.2 评价指标 通常，中医药领域 NER 以实体边界和实体类别是否同时预测正确作为评价标准，以精确率 P、召回率 R 和调和平均数 F_1 作为评价指标。除此之外，有学者[1, 2]还引入准确率、宏平均、微平均和实体内部标签评价进一步评价 NER 的识别效果。常用评价指标的计算公式如图 2 所示：

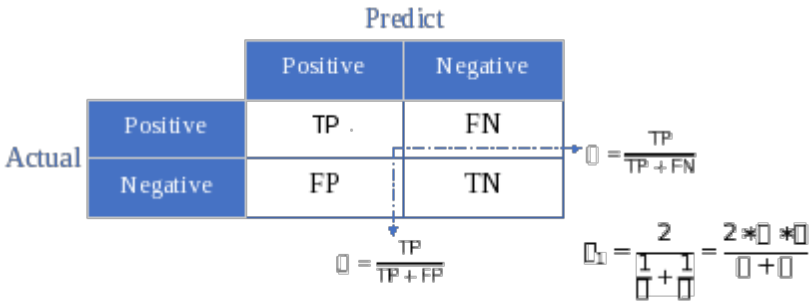

图 2 评价指标的计算公式

2. 命名实体识别在中医药领域的发展历程

文章将中医药领域 NER 研究分为机械识别研究、监督学习下研究及匮乏资源下研究三类，详细介绍了基于词典与规则、统计机器学习和深度学习三种常用的技术方法，方法的实现路线见图 3。

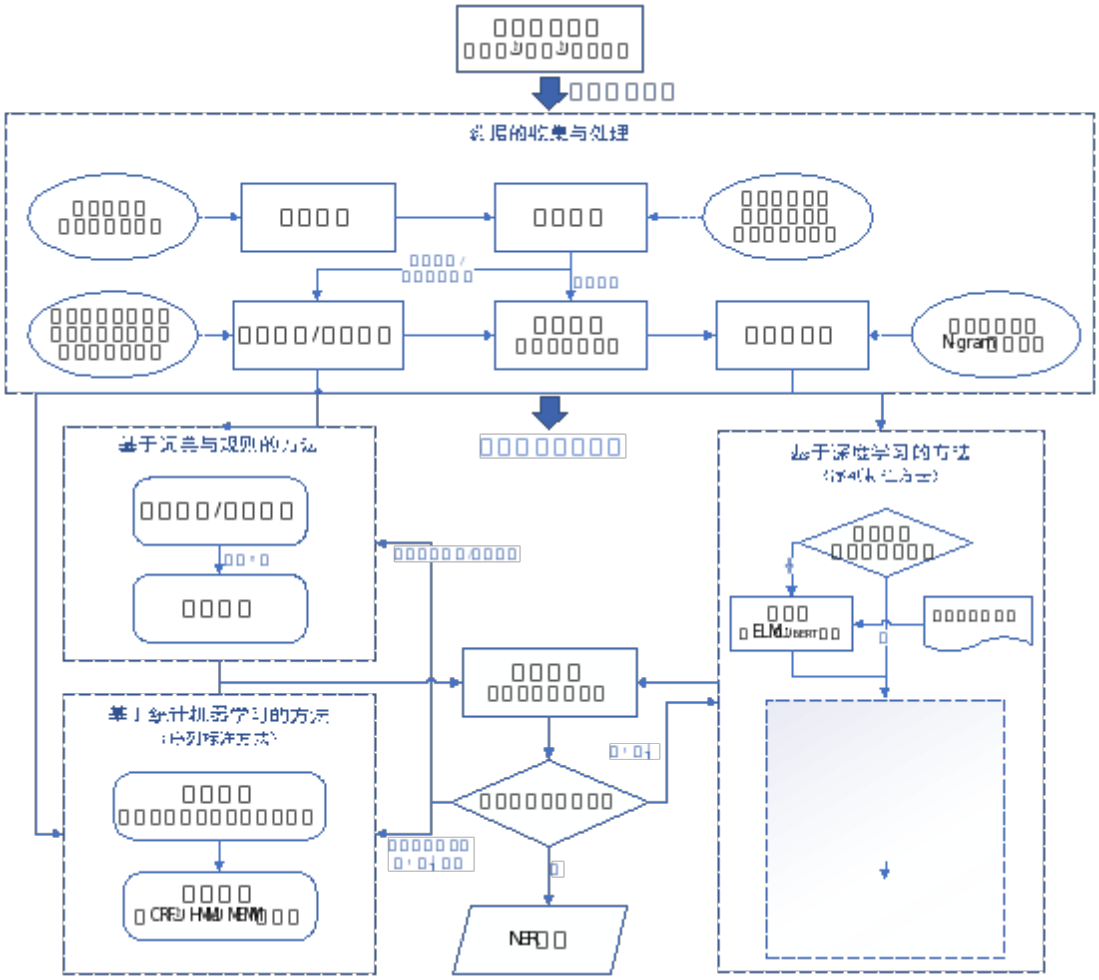

图 3　NER 实现路线图

2.1 机械识别研究

机械识别研究多指基于词典与规则的方法，研究者需先构建实体词典或规则词典[3]，再以模式匹配的方法进行实体抽取。常用的匹配方法如最大匹配算法、正则表达式匹配等；制定规则词典的方法如人工归纳总结、自动统计归纳。其中，自动统计归纳法可采用以种子词（实体）集迭代搜索更新规则模板的方式，在种子词集更新时，还可利用实体及上下文特征筛选候选种子词，以减少规则模板非专业性带来的影响，该模式下方剂 F_1 值达 90.9%[4]。

2.2 监督学习下的 NER 研究

2.2.1　基于统计机器学习的方法

统计机器学习方法的实现思路可分为两种，一是依次完成实体边界识别和实体类型预测两项任务，如刘一斌等[5]先借助中文分词确定实体内容，再训练分类模型确定实体类别，F_1 值为 80.74%。二是将 NER 视作序列标注任务，通过构建序列标注模型，预测序列各位置对应的标签，根据标签确定实体的边界及类别，该方法是监督学习下实现 NER 的常用思路。常用的统计机器学习模型如条件随机场[6]（CRF）、结构化支持向量机[7]（SSVM）等。

特征工程是构建统计机器学习模型的关键工程活动，研究者根据中医药实体及文本特点，人工选择有价值的特征、构建合适的特征模板，以提高模型的识别效果，如刘凯等[6]基于 CRF 模型实现中医电子病历 NER，该模型引入了上下文指示词、词典等特征，症状 F_1 值达 80%。

2.2.2　基于深度学习的方法

深度学习方法以“End-to-End”的学习方式求解序列标注问题，即“原始数据”输入模型后，模型自动完成特征学习和标签预测任务。通常，深度学习模型包含嵌入层、编码层和解

码层，嵌入层用于获取文本序列的嵌入表示，常见的方法见表 1；编码层用于文本特征提取和标签预测，如双向长短期记忆神经网络[8]（BiLSTM），BERT 等神经网络模型；解码层用于解析最优标签序列。

文章根据“嵌入层-编码层”实现方式的不同将深度学习模型分为基于特征（Feature-based）和基于微调（Fine-tuning）的实现。特征法将嵌入层获取的词向量作为特征输入编码层训练，训练时嵌入层模型参数不更新[1]；微调法基于迁移学习策略，将预训练好的模型迁移至目标任务中，训练时预训练模型的参数随训练数据的读入不断微调，是预训练方法的常用思路[9]。

表 1 表征学习方式对比

表征学习方式	举例	词/字嵌入获取方式	优/缺点
离散表示	One-hot	词/字在词典中的位置	无法表达词汇间的联系
词嵌入	Word2vec[1]	自监督学习的方法	无法解决一词多义问题
语境化嵌入	BERT[9]	自监督学习的方法	可根据语境的变化动态地更新词嵌入
混合方法	文献[10]	多种表示方法融合	词向量包含更丰富的特征信息

2.3 匮乏资源下的 NER 研究

监督学习方法下的 NER 研究依赖大规模、高质量的标注语料，但中医药领域语料资源匮乏、语料获取代价昂贵。为此，研究者一方面改进获取语料的方法，如范岩[11]将 Bootstrapping 算法生成的粗标注语料用作 CRF 的训练语料，F_1 值达 72.46%。另一方面改变模型方法，如 Zhang 等[12]提出了引入回标策略的远程监督算法，在仅使用中医领域词典的情况下，F_1 值达 69.76%；Qi 等[13]提出了跨层次的远程监督算法，该方法无需人工标注数据，平均 F_1 值达 77.34%。

2.4 研究方法的对比分析

研究方法的对比情况见表 2，图 4 展示了中医药领域 NER 技术方法随年份的变化情况，研究发现，2004 年伊始，中医药领域 NER 研究呈稳步增长趋势，近 3 年增长速度较快；技术方法的应用阶段有明显的时间特征，具体见表 2；除此之外，中医药领域 NER 研究以对比研究（1-3）和改进研究（4-10）为主，研究内容及结果见表 3；目前，基于迁移学习策略的深度学习方法是中医药领域 NER 研究的主流方法。

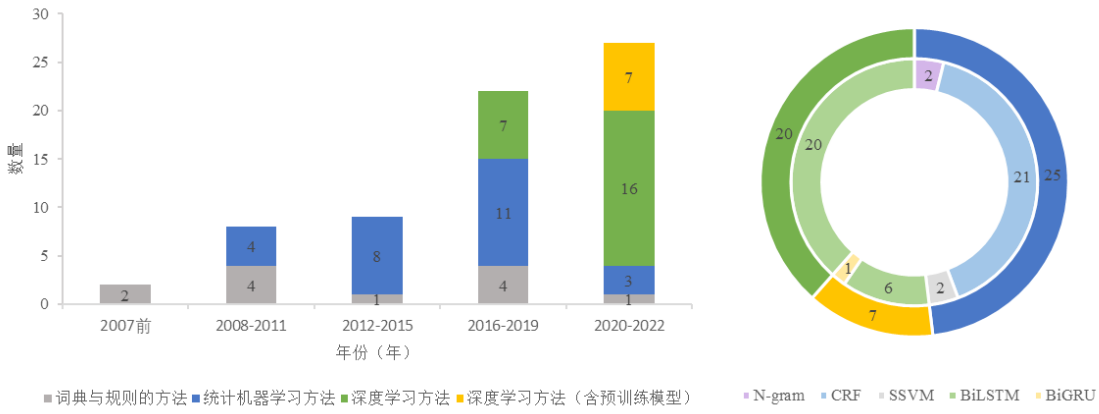

图 4 技术方法及年份统计图

表 2 不同研究方法对比

对比内容	机械识别	监督学习		匮乏资源
	词典与规则	统计机器学习	深度学习	-
工程依赖	规则模板	特征工程	数据驱动	-
数据集	规则词典	标注数据集	标注数据集	-
识别方法	模式匹配	序列标注模型	序列标注模型	-

应用阶段	2009 年以前	2009 至 2018 年	2018 年以后	-
优点	识别速率快，精确率高	能联合上下文信息，依赖样本的程度较小	长距离信息处理能力，灵活，依赖样本的程度高	依赖样本的程度小
缺点	歧义处理能力差，泛化能力差	依赖人工参与，抗干扰能力差	模型训练代价高可解释性差	模型性能不及监督学习方法

表 3 中医药领域 NER 研究

序号	研究内容	研究结果
1	对比不同模型	基于预训练的深度学习模型优于其他模型[14]
2	对比不同嵌入方式	字嵌入效果优于词嵌入[15]，语境化嵌入法优于其他单一嵌入方法[16]
3	对比不同模型内部结构[17]	-
4	引入多特征	如词边界，词性，键值对[18]，词典，实体关键词，实体结构[19]，拼音[20]，上下文指示词等
5	剔除无效信息[20]	可提升 CRF 模型的性能
6	引入分步提取策略[21]	先识别细粒度实体，再将其转为特征识别实体的方法可提升模型的准确率和查全率
7	优化序列的嵌入表示	引入局部特征编码（FOFE）[22]或引入症状关键字特征[17] 均可提升深度学习模型的性能
8	引入主动学习策略[23]	以较少的标注语料获取较高的收益
9	实体关系联合抽取[24]	可解决实体离散问题

3. 中医药领域 NER 的应用研究

3.1 构建中医药领域知识库

从 20 世纪 80 年代以构建专家系统为目的的中医辅助诊断数据库[25]、90 年代以《中国中医药数据库》为代表的中医文献数据库，到 21 世纪提出的以知识服务和知识组织为基础的智能医学知识库，事实知识一直是知识库的主要内容。早期，抽取事实知识多采用人工提取的方式，效率低、代价高，知识库的规模和水平难以达到应用标准；之后，有学者尝试采用机器抽取的方法，如 Ruan 等[26]基于统计机器学习的方法，建成了中文症状公共知识库。中医药领域 NER 的发展改善了事实知识的抽取效率，必将推动建成基于本体的中医学知识库和更多中医药领域专题知识库。

3.2 构建中医药领域知识图谱

中医药领域数字化进程的不断推进，使得中医药标准化、结构化、智能化的要求不断增长，知识图谱逐步成为中医药领域研究的焦点。知识图谱是一种知识表示的方式[27]，与知识库一样，NER 是构建知识图谱的重要技术支撑。目前，已有多名学者借助 NER 建成了中医诊断[28]、中医古籍[16]、专病[22]、方剂[29]等中医药专题知识图谱。值得注意的是，知识库与知识图谱并不相同，前者更多的是事实、规则等多方面知识的集合，后者更侧重于隐性知识的推理和挖掘；因此，知识图谱正逐步替代知识库成为中医智能问答、智能诊断等中医智慧医疗服务系统的最主要知识来源。

3.3 构建中医药领域问答系统

智能问答是智慧中医的重要组成部分，对推动中医教育模式的转变、中医治未病思想的推行有重要现实意义，其实现方式主要包括两种，一是基于语义理解[30]，以语义分析后的逻辑表达式检索问答库；二是基于信息检索[31]，以语句中的实体查询问答库；两种实现方式不尽相同，但一定程度上均依赖 NER。崔智颖[32]综合利用了 NER、中文分词、语义相似度分析等技术，建成了面向中医药功效的智能问答系统。在问答系统应用时，个体表述差异、字词使用不规范等问题普遍存在，提升系统的纠错能力、中医药领域 NER 的性能、语义分析能力，扩大智能问答库，加入智能语音技术等或是当前中医药智能问答系统优化和前进的方

向。

4. 总结与展望

4.1 不足与挑战

4.1.1 模型方法单一

中医药领域 NER 研究多采用监督学习方案，以模型的比较研究为主，模型的改进研究为辅，模型方法和改进形式过于单一，图 3 右侧给出了监督学习模型的使用情况（深度学习模型仅统计特征提取层模型）。值得注意的是，监督学习模型性能的优劣与训练数据集的规模、质量密不可分，而当前，由于标注语料获取难度大，中医药领域 NER 研究的训练数据集普遍较小，一定程度上影响了模型性能的判定和 NER 效果的提升。

4.1.2 尚无统一的实体规范

领域标注数据集是开展中医药领域 NER 研究的重要基础。目前，中医药领域尚无统一的实体标注规范，NER 研究均是基于研究者自建的标注语料，语料的数据源、标注规范等均由研究者自行定义；考虑到中医药学概念模糊，中医药实体多存在描述不统一、边界不清晰、类别易混淆等问题，易造成语料间出现实体类型和实体边界等标注差异，导致中医药领域 NER 研究重复度高、横向对比难、研究成果转化难等问题，阻碍了中医药领域 NER 的发展。

4.1.3 NER 技术的局限性

中医药古籍是中医药学的知识宝库，是中医药现代化不可或缺的组成部分，但因古汉语与现代汉语的巨大差异，不同时期古汉语的语言差异及中医药文本简洁、表述抽象，习惯性缩写、实体连写，古今词混用，文本倒装，歧义等问题，NER 的效果并不理想。除此之外，现阶段 NER 对于长度较大的实体的识别性能较差，难以处理包含标点的实体，且中医药领域 NER 要求的高精确率和高召回率，是现阶段机器学习算法尚且无法达到的高度。

4.2 展望

4.2.1 深入 NER 研究

在信息爆炸的今天，中医药结合信息技术是现实所趋，也是中医药发展的机遇和挑战。在技术层面，中医药领域 NER 研究应以尝试不同模型为辅，方法的改进为主，结合领域特点和领域实际，探索适合中医药领域的 NER 方法；同时，匮乏资源下研究 NER 的趋势短期内不会改变，可利用主动学习、迁移学习、强化学习、多任务学习[34, 35]、无监督学习等策略充分发挥有限资源的作用。

4.2.2 统一标注规范

制定统一、清晰、规范的标注标准是保证中医药领域 NER 长期健康、有效发展的重要举措，是未来中医药领域 NER 研究发展的首要工作之一；而加快推进古籍整理与数字化研究进程，逐步开放古籍资源是促进中医药数据规范化治理、标注规范制定的重要保证。

4.2.3 多学科协作合作

中医药 NER 研究是中医药人工智能研究的重要环节，应受到广大研究者的重视，积极参与解决双边问题；通过多学科合作的方式，充分发挥各学科优势，协同并进，共同解决中医药文本中长期无法解决的问题，是中医药领域 NER 快速发展的最有效途径。

4.2.4 推广应用

在实际应用中，NER 方法的选择应综合技术方法的优缺点和中医药实体的特点考虑，合理利用各模型方法的优势。NER 的应用如参考基于本体的中医药学语言系统[36]构建中医药领域语言系统；利用 NER 在广泛收集领域术语的基础上进行规范定名和制定术语标准，推动中医药术语的规范化治理和标准化使用；或用于中医药领域术语库建设、中医药领域信息检索、中医药文本翻译等。

参考文献

[1] 肖瑞, 胡冯菊, 裴卫. 基于 BiLSTM-CRF 的中医文本命名实体识别[J]. 世界科学技术-中医药现代化, 2020,22(07):2504-2510.
[2] 高佳奕, 刘震, 杨涛, 等. 基于条件随机场的中医临床医案症状命名实体抽取研究[J]. 世界科学技术-中医药现代化, 2020,22(06):1947-1954.
[3] 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(08):1241-1250.
[4] 周雪忠. 文本挖掘在中医药中的若干应用研究[D]. 浙江大学, 2004.
[5] 刘一斌, 叶辉, 易珺, 等. 基于朴素贝叶斯和 word2vec 的中医电子病历文本信息抽取[J]. 世界科学技术-中医药现代化, 2020,22(10):3563-3568.
[6] 刘凯, 周雪忠, 于剑, 等. 基于条件随机场的中医临床病历命名实体抽取[J]. 计算机工程, 2014,40(09):312-316.
[7] 原旎, 卢克治, 袁玉虎, 等. 基于深度表示的中医病历症状表型命名实体抽取研究[J]. 世界科学技术-中医药现代化, 2018,20(03):355-362.
[8] 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(01):113-123.
[9] 屈倩倩, 阚红星. 基于 Bert-BiLSTM-CRF 的中医文本命名实体识别[J]. 电子设计工程, 2021,29(19):40-43.
[10] 肖猛. 面向中医证候的健康领域知识图谱构建与应用研究[D]. 吉林大学, 2019.
[11] 范岩. 基于条件随机场模型的中医文献知识发现方法研究[D]. 北京交通大学, 2009.
[12] Zhang D, Xia C, Xu C, et al. Improving distantly-supervised named entity recognition for traditional Chinese medicine text via a novel back-labeling approach[J]. IEEE Access, 2020,8:145413-145421.
[13] Jia Q, Zhang D, Xu H, et al. Extraction of Traditional Chinese Medicine Entity: Design of a Novel Span-Level Named Entity Recognition Method With Distant Supervision[J]. JMIR medical informatics, 2021,9(6):e28219.DOI:10.2196/28219.
[14] 屈丹丹, 杨涛, 朱垚, 等. 基于字向量的 BiGRU-CRF 肺癌医案四诊信息实体抽取研究[J]. 世界科学技术-中医药现代化, 2021,23(09):3118-3125.
[15] 高佳奕, 杨涛, 董海艳, 等. 基于 LSTM-CRF 的中医医案症状命名实体抽取研究[J]. 中国中医药信息杂志, 2021,28(05):20-24.
[16] 卢克治. 基于中医古籍的知识图谱构建与应用[D]. 北京交通大学, 2020.
[17] 李明浩, 刘忠, 姚远哲. 基于 LSTM-CRF 的中医医案症状术语识别[J]. 计算机应用, 2018,38(S2):42-46.
[18] 王国龙, 杜建强, 郝竹林, 等. 中医诊断古文的词性标注与特征重组[J]. 计算机工程与设计, 2015,36(03):835-841.
[19] Zhao Y. Research on Entity Recognition in Traditional Chinese Medicine Diet[C]//: 2017 9th International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC), 2017. IEEE.
[20] Liang J, Xian X, He X, et al. A Novel Approach towards Medical Entity Recognition in Chinese Clinical Text[J]. Journal of Healthcare Engineering, 2017,2017:1-16.DOI:10.1155/2017/4898963.
[21] 万静, 涂喆, 冯晓. 基于条件随机场的医药领域症状信息抽取[J]. 北京化工大学学报(自然科学版), 2016,43(01):98-103.
[22] 郑子强. 面向慢性肾脏病中医医案的知识图谱学习与推理研究[D]. 电子科技大学, 2020.
[23] 李焕. 基于深度学习与主动学习的中医术语识别研究[D]. 北京工业大学, 2019.
[24] 庞震, 顾继昱, 吴宇飞, 等. 中医诊治高血压医疗实体提取问题研究[J]. 医学信息学杂志, 2021,42(09):45-51.
[25] 马利, 刘保延, 谢琪, 等. 我国中医知识库研究回顾与展望[J]. 中国数字医学, 2014,9(01):11-14.
[26] Ruan T, Wang M, Sun J, et al. An automatic approach for constructing a knowledge base of symptoms in Chinese[C]//: 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 2016. IEEE.
[27] 漆桂林, 高桓, 吴天星. 知识图谱研究进展[J]. 情报工程, 2017,3(01):4-25.
[28] 杜伯涵. 基于知识图谱的中医智能导诊系统[D]. 北京邮电大学, 2021.
[29] 刘禹琪. 中医名方知识图谱构建与链路预测模型的研究及应用[D]. 东北师范大学, 2021.
[30] 杜睿山, 张轶楠, 田枫, 等. 基于知识图谱的智能问答系统研究[J]. 计算机技术与发展, 2021,31(11):189-194.
[31] 陶永芹. 专业领域智能问答系统设计与实现[J]. 计算机应用与软件, 2018,35(05):95-101.
[32] 崔智颖. 面向中医药功效的智能问答系统[D]. 杭州师范大学, 2020.
[33] 贾李蓉, 于彤, 李海燕, 等. 中医药学语言系统的语义网络框架概述: 第一届中国中医药信息大会, 中国北京, 2014[C].