·专　　题·

中医文本命名实体识别研究综述

时倩如[1]　李　贺[1*]　于雯倩[1]　沈　旺[1]　张承坤[2]
（1. 吉林大学商学与管理学院，吉林 长春 130012；2. 南京中医药大学中医学院，江苏 南京 210023）

摘　要：［目的/意义］中医文本中包含了大量领域相关知识，可为准确诊断和有效的疾病防治提供指导。本文对中医文本命名实体识别(NER)研究进行系统性综述。［方法/过程］从中医文本的特征出发，探讨了中医文本 NER 在知识体系、语料构建和技术算法层次面临的挑战；梳理中医文本 NER 语料构建中可用的术语标准、实体类型和标注原则与方法；归纳中医文本 NER 技术的一般框架、常用方法和近期趋势，并总结评估指标。［结果/结论］建议未来研究可从以下方向开展：在语料层面制定标注规范并构建高质量数据集，在算法层面探索针对小样本问题的数据优化、针对复杂实体的识别模型和增强模型解释性，以提高中医 NER 的效果。

关键词：命名实体识别；中医；深度学习；自然语言处理；综述
DOI:10.3969/j.issn.1008-0821.2025.02.001
〔中图分类号〕G250.2；TP291.1　〔文献标识码〕A　〔文章编号〕1008-0821（2025）02-0004-13

Review of Research on Named Entity Recognition in Traditional Chinese Medicine（TCM） Texts

Shi Qianru[1]　Li He[1*]　Yu Wenqian[1]　Shen Wang[1]　Zhang Chengkun[2]
（1. School of Business and Management，Jilin University，Changchun 130012，China；
2. School of Traditional Chinese Medicine，Nanjing University of Chinese Medicine，Nanjing 210023，China）

Abstract：［Purpose/Significance］Traditional Chinese medicine（TCM） texts contain a wealth of domain-specific knowledge that can provide guidance for accurate diagnosis and effective disease prevention and treatment. This paper provides a systematic review of the research on named entity recognition（NER） in TCM.［Method/Process］From the characteristics of TCM texts，the paper discussed the challenges of TCM texts NER including knowledge system，corpus construction，and technical algorithms aspects. It sorted out the terminological standards，entity types，and annotation principles and methods available in the construction of TCM texts NER corpora. The paper also reviewed the general framework，common methods，and recent trends of TCM texts NER techniques and summarized the evaluation indicators.［Result/Conclusion］Future research can be conducted in the following directions：establishing annotation standards and constructing high-quality datasets at the corpus level，exploring data optimization for small sample problems，developing recognition models for complex entities，and enhancing model interpretability at the algorithm level to improve the effectiveness of TCM NER.

Key words：named entity recognition；traditional Chinese medicine；deep learning；natural language processing；review

中医学是在中华几千年的历史长河中形成的独具特色的医学体系，对中华民族的繁衍昌盛起到积极作用。信息技术的发展推动了中医知识的现代化应用。随着对中医知识深度研究的需求日益增长，

收稿日期：2024-10-31
基金项目：国家社会科学基金冷门绝学专项研究项目“本草典籍整理、知识组织与智慧化建设研究”（项目编号：23VJXT024）。
作者简介：**时倩如**（1996-），女，博士研究生，研究方向：数据挖掘。**于雯倩**（1996-），女，博士研究生，研究方向：知识组织。**沈旺**（1983-），女，教授，博士生导师，研究方向：数据挖掘。**张承坤**（1992-），男，讲师，研究方向：中医医史文献。
通信作者：**李贺**（1964-），女，教授，博士生导师，研究方向：知识管理，大数据分析。

命名实体识别技术(NER)越来越多的应用于中医文本挖掘。

中医文本 NER 属于领域 NER 范畴。不同领域的文本通常包含特定的专业术语、实体类型和上下文信息，因此领域 NER 需要根据这些特征进行模型的优化，以提高实体识别的准确性。中医文本不仅涉及传统医学知识，还融合了哲学、文化和历史背景。同时，中医文本常以古文或半文言文形式呈现，其语法结构与现代汉语有显著差异。由于缺乏标准化的术语规范，不同文献可能对同一概念有不同的表达。此外，中医领域的高质量标注数据相对稀缺。这些因素共同导致了中医文本 NER 研究面临诸多挑战。中医 NER 是典型的交叉课题，受到中医药、计算机、数字人文等多领域学者的关注，积累了较多的研究成果。在其发展过程中，不乏一些从不同视角进行归纳总结的综述性论文，例如，中医术语研究文献计量分析[1]、实体抽取在中医药领域的应用综述[2]、中医症状信息抽取研究进展[3]等。然而，当前中医文本 NER 缺乏系统性综述，这阻碍了研究人员对已有工作的全面理解和对未来研究方向的把握。本文旨在填补这一空白，为该领域的发展提供有价值的参考。

本文在分析中医文本特征的基础上，提出中医文本 NER 在知识体系、语料构建和技术算法层次的研究挑战；系统性地梳理中医文本 NER 语料构建中的术语标准、实体类型和标注原则与方法；从基于词典和规则的模式匹配方法、基于统计原理的机器学习方法和基于深度学习的方法 3 个方面介绍中医文本 NER 的技术发展历程，并详细介绍基于深度学习的中医文本 NER 方法的一般框架和 3 种主流架构；最后，基于研究现状，对未来发展进行展望，以期为该领域的进一步研究提供参考。

1 中医文本特征及命名实体识别挑战

中医文本形式多样，不同类型的文本具有共同之处，也不乏差异性。本节首先梳理不同类型中医文本的通用特征，然后针对典型的中医文本形式进行各自特征的剖析，最后提出中医文本 NER 的挑战。

1.1 中医文本通用特征

中医文本具有抽象性、经济性和复杂性等特点。中医包含的部分概念无法对应到客观世界的具体事物，且常使用隐喻、象征等修辞手法。例如，“脏腑”并不仅仅指现代医学的某个具体器官，而是涵盖了人体内脏的功能、相互关系以及与外界环境的互动等多个层面的系统概念[4]。这种独特的表达方式使中医文本较为抽象且模糊，常常给人深奥晦涩之感。中医语言的经济性与其抽象性是高度一致的。中医行文倾向于删繁去冗，省略某些词句的情况比比皆是。《重广补注黄帝内经素问》序认为：“其文简，其意博，其理奥，其趣深。”凸显的就是中医语言的经济性。这种经济性使其信息密度较高，某些字、词甚至句子的语义极度依赖于上下文语境。

此外，中医语言的字词含义在其发展历程中不断扩充，经常存在一词多义和多词一义的现象，同名异物和同物异名情况也较为普遍。中医用语继承了古汉语的特点，保留了较多生僻字和通假字。这些特征使中医语言具有高度的复杂性。

1.2 不同类型中医文本的特征

1.2.1 中医诊疗文本

中医诊疗文本是中医临床实践的重要记录形式。中医诊疗文本的内容较为完整、有相对稳定的结构、要素结构也相对简单。在内容上，诊疗文本力求简明，只记录关键信息，如重要的症状、诊断等。在语言上，诊疗文本古今汉语混用情况十分常见，具有叙事性强、口语化重等特点[5]。

1.2.2 中医古代文献

在各类中医文本中，中医古代文献的书写风格最为晦涩难懂。中医古代文献使用的古汉语在词汇、句法和语法结构上与现代汉语存在显著差异。此类文献另一显著特征是流传版本繁多，呈现出同书异本、同书异名同版、同书异名异版等繁杂现象。一般认为，应选择底本优良且经过专家校注的权威版本，以保障数据标注和语料库的建设质量。

1.2.3 中医科技文献

中医科技文献包括专利、学术论文和专著等形式。在结构上，中医科技文献通常遵循一定的研究框架，其结构严谨、逻辑清晰，系统性记录了研究背景、方法、过程和结果等部分。在语言上，中医科技文献可能同时使用传统医学和现代医学专业术语，且不少文献包括大量的数据。

1.2.4 网络开放资源

网络开放中医资源来自于各医疗机构、研究机构和普通公众，相关文本语言风格多样化，不可一概而论。例如，社交媒体中的数据信息密度极低，而在线问诊数据则更为专业化。随着中医药国际化的推进，还出现了多语言的中医文本。这些资源不仅为公众提供了学习中医药知识的平台，也为研究人员提供了丰富的数据来源。

1.3 中医文本命名实体识别挑战

1.3.1 知识体系层次

中医知识体系是一个错综复杂的系统，融合了古代哲学思想、自然科学理论以及长期的实践经验。中医基于阴阳五行学说阐释人体与自然之间的和谐关系，以脏腑经络理论为核心，构建了生理病理模型。从纵向来看，中医学不断演化，知识体系也随之扩展。中医学深刻的哲学内涵、精细的理论架构以及不断发展使其知识体系呈现高度的复杂性，是中医文本 NER 研究在知识体系层次面临的重要挑战。

1.3.2 语料构建层次

语料库建设是中医药领域的一项重要工作，取得了显著的研究成果。然而，由于资源私有化、数据孤岛等问题，相关资源以个案形式分散分布，尚未整合成一个全面的语料库系统，难以满足大规模数据驱动的中医文本 NER 需求。此外，中医领域长期以来面临术语规范化不足的问题。相关术语标准无法覆盖所有的实体，仍有许多实体缺乏明确的规范名称。在实际应用中，标准的实施和推广面临困难，也制约了语料库建设的进程。受限于上述多种现实因素，高质量的中医语料库依旧相对稀缺，直接限制了中医文本 NER 模型的效果。

1.3.3 技术算法层次

中医文本中常存在实体嵌套、实体序列分散和实体过长等现象，导致实体边界较为模糊。实体嵌套情况要求 NER 算法具备层次化的区分能力，例如，“麻黄桂枝汤”中同时包含方剂名“麻黄桂枝汤”，又包含药物名“麻黄”和“桂枝”，识别结果应根据需要精准把握不同层级实体的边界。实体序列分散问题则需要算法具备足够的上下文信息利用能力。例如，在“脉沉无力”中，算法应识别出“脉沉”和“脉无力”两个独立实体。上述问题对中医 NER 算法的语义理解能力提出了极高的要求。

2 中医文本命名实体识别语料构建

2.1 中医文本命名实体识别术语标准

构建高质量的中医语料库，可以为 NER 模型提供丰富的训练数据。中医术语规范化是中医药标准化的基础性工作。本节对中医文本 NER 可参考的现行术语标准进行总结，如表 1 所示。在实际应用中，应优先以法典与国家标准为参照，其次以行业标准、工具书与教材为准。现有标准所收录的词，原则上不应进行切分。

表 1 中医文本命名实体识别术语标准

Tab. 1 Standard Terminology for NER in TCM Texts

类 别	名 称	编号/ISBN	实施日期	归口单位
国家法典	中华人民共和国药典(2020 年版)	9787506700778	2020-12-30	国家药品监督管理局
国家标准	中医基础理论术语	GB/T 20348-2006	2006-10-01	全国中医标准化技术委员会
	中医病证分类与代码	GB/T 15657-2021	2021-10-11	
	中医病证分类与代码(疾病部分、证候部分、治法部分)	GB/T 16751-2023	2023-03-17	
	中医临床名词术语 第 1~9 部分	GB/T 42467-2023	2023-03-17	
	中药编码规则及编码	GB/T 31774-2015	2015-12-01	
	中药方剂编码规则及编码	GB/T 31773-2015	2015-12-01	
	中医药学主题词表编制规则	GB/T 40670-2021	2021-10-11	
	健康信息学 中医药学语言系统语义网络框架	GB/T 38324-2019	2020-07-01	中国标准化研究院
	健康信息学 中医药数据集分类	GB/T 38327-2019	2020-07-01	

表 1（续）

类 别	名 称	编号/ISBN	实施日期	归口单位
国家标准	针灸学通用术语	GB/T 30232-2013	2014-12-01	全国针灸标准化技术委员会
	经穴名称与定位	GB/T 12346-2021	2021-11-26	
	耳穴名称与定位	GB/T 13734-2008	2008-07-01	
	经外奇穴名称与定位	GB/T 40997-2021	2021-11-26	
行业标准	中医病证诊断疗效标准(儿科、耳鼻喉科、妇科、肛肠科、骨伤科、内科、皮肤科、外科、眼科)	ZY/T 001-1994	1994-06-28	国家中医药管理局
团体标准	中医临床基本症状信息分类与代码	T/CIATCM 020-2019	2019-05-01	中国中医药信息学会
	中医特色治疗项目信息分类与代码	T/CIATCM 022-2019	2019-05-02	
	临床中药基本信息分类与代码	T/CIATCM 024-2019	2019-05-03	
	中医舌象诊断信息分类与代码	T/CIATCM 010-2019	2019-05-04	
	中医脉象诊断信息分类与代码	T/CIATCM 011-2019	2019-05-05	
	中医药信息化常用术语	T/CIATCM 001-2019	2019-05-01	
	中医病证术语属性描述基本模型	T/CIATCM 021-2019	2019-05-01	
	中药饮片处方用名规范	T/CACM 1361-2021	2021-06-30	中华中医药学会
	针刀医学临床基础术语	T/CACM 1063-2018	2019-09-01	
	中医治未病术语	T/CACM 1067-2018	2018-11-15	
词典词表	中国中医药学主题词表	GB/T 40670-2021	2021-10-11	全国中医标准化技术委员会
	中国医学大辞典	7530824791	1998-04-01	天津科学技术出版社
	中医药通假字字典	9787543919143	2001	上海科学技术文献出版社
	中医药常用名词术语辞典	9787801562074	2001	中国中医药出版社
	中医方剂大辞典(第 2 版)	9787117282598	2015-2019	人民卫生出版社
	中医大辞典(第 3 版)	9787521438512	2023	中国医药科技出版社
	中药大辞典(第 2 版)	9787532382712	2006	上海科学技术出版社
其 他	中医名词术语选释	9787500206040	1973	人民卫生出版社
	中国科学技术史：度量衡卷	7030078918	2001	科学出版社
	中国历代度量衡考	9787030032836	1992	科学出版社

2.2 中医文本命名实体识别实体类型

中医文本 NER 的实体类型划分以中医理论为基础，围绕辨证论治的核心思想展开。相关研究中包含的实体类型，主要包括疾病、症状、治法、方剂、药物等，如表 2 所示。此外，部分研究抽取了中医认知方法、阴阳五行、运气学说等基础理论相关的实体。针对中医文本中蕴含的民俗、伦理观念及文化内涵等人文相关实体的研究则相对较少。

不同研究者往往根据各自的背景和需求，采用不同的实体分类体系。这一方面是由于中医文本具有多种类型，不同类型文本中包含的实体本身就存在较大差异。例如，针灸多涉及经络、腧穴等特定术语，而本草则主要涉及性味归经等。另一方面，

表 2 中医文本命名实体识别实体类型

Tab. 2 Entity Types for NER in TCM Texts

类 别	相关实体
体征	体征、舌象、舌苔、脉象、临床表象、辨证
疾病	疾病、证型、症候、病症、病因、病机(病理)、病位、病性、主诉、程度
治法	治法(疗法)、治则、针灸、检查、导引、针法
方剂	方剂、配伍、组分、处方
药物	药物、药材、名称、别名、功效、主治、剂量、性味、归经、部位、采收、鉴定、分类、炮制、用法、禁忌、产地
中医理论	运气学说、阴阳、五行术语、中医生理、中医自然
其他	医家、人物、古籍、时间、腧穴、经络、人群

即便是对同一类中医文本，不同学者在实体类型划分上也存在差异，特别是在实体的粒度的选择上。如表 3 所示，同样是针对中医古籍《神农本草经》的命名实体识别，各研究选择的实体类型存在显著的差异。这一差异主要源于中医领域缺乏公认的、系统化的实体标准。NER 方法可以在没有实体标准的情况下开发和实现，这允许研究者根据需求进行快速迭代和灵活调整。然而，缺乏实体标准使不同数据集和标注方案之间不可融合，导致了知识的孤岛化与碎片化问题，阻碍了模型之间的迁移。同时，研究者在进行模型评估时，无法采用一致的标准进行算法比较，影响了研究的可重复性。

表 3 《神农本草经》命名实体识别研究实体类型划分

Tab. 3 Entity Types for NER in the Shennong Ben Cao Jing

文 献	方 法	实 体 类 型
[6]	BiLSTM-CRF	中药材、方剂、病症
[7]	BERT	分类、名称、产地、性味、药用部位、采收、鉴定、主治、功效、炮制、用法、禁忌、古籍名称、医家
[8]	BERT-CRF	药物、药性、疾病、功效
[9]	词典增强的 Bert-BiGRU-CRF	药名、药性、药味、归经、别名、症状、功效、古籍

2.3 中医文本命名实体识别实体标注

2.3.1 标注原则

命名实体标注应遵循可分性、不可分性和一致性等原则，保障标注的准确性和可靠性。可分性原则指的是具备相对独立语义的词组应作为独立的实体进行标注。例如，“清热解毒”是中药常用治疗方法，其含义可以拆分为“清热”和“解毒”两个部分。不可分性原则强调某些专业术语和组合词应视为不可分割的整体。诸如《黄帝内经》和《神农本草经》等篇章名，作为中医领域广为接受的专业术语，在 NER 过程中不应被拆分。由两个或多个构词要素组成的组合词，如方剂名“四时加减柴胡饮子”和证型名“阴虚阳亢证”等，拆分将导致概念的丧失，无法准确传达其所承载的专业知识。因此，必须确保它们在标注和识别过程中的完整性。此外，应确保同一个实体在不同上下文中被一致地标注。一致性原则涵盖了多个方面，包括实体定义的一致性、标注规则的一致性和上下文应用的一致性等。

2.3.2 标注方法

目前，中医文本 NER 标注方法与通用领域 NER 的标注方法大致相同，主要包括 BIO、BIOS、BMES、BIESO 等。其中，最常使用的是 BIO 和 BIOS 标注。各标注方法的具体含义总结如表 4 所示。

表 4 中医文本命名实体识别常见标注方法

Tab. 4 Annotation Methods for NER in TCM Texts

标注方法	实体首字	实体中间字	实体结尾字	非实体	单字实体
BIO	B	I	I	O	-
BIOS	B	I	I	O	S
BIOES	B	I	E	O	S
BMES	B	M	E	O	S
BMEO	B	M	E	O	-

3 中医文本命名实体识别方法

中医文本 NER 技术沿着通用领域 NER 技术的发展路线演进，经历了基于词典和规则的模式匹配方法、基于统计原理的传统机器学习方法和基于神经网络的深度学习算法等发展阶段。

3.1 传统方法

3.1.1 基于词典和规则的模式匹配方法

在早期阶段，中医文本 NER 主要依赖于构建预定义的专业词典和规则进行模式匹配。领域词典包含的是准确的已知知识，为 NER 提供了可靠的参考依据。基于规则的方法在分析文本规律的基础上，制定规则集，利用最大匹配算法[10]、正则表达式[11]等进行实体抽取。模式匹配方法展现出了较高的准确率，但其局限性也不容忽视。领域词典需要不断更新和完善，以适应新出现的术语和概念，对于未登录实体的识别效果往往不佳。规则的制定依赖于领域专家的经验，在自由文本处理方面的能力相对较弱，且可扩展性有限。随着中医文本数据的不断增加和多样化，单纯依靠词典和规则的方法

难以满足实际应用的需求。

尽管存在这些问题，基于词典和规则的模式匹配方法具有可解释性强、易于理解的优点。对于规模较小且结构化程度较高的中医文本，该方法具有简单、准确的优势。当前，一种新的研究范式是词典、规则与深度学习技术相结合，在语料准备阶段，利用词典与规则进行自动标注，有效减少了人工标注的工作量，且提升了标注的准确性。

3.1.2 基于统计原理的传统机器学习方法

基于统计原理的传统机器学习方法是基于概率性的非确定性模型，依赖于数据的统计学特征进行预测。该方法的核心是特征工程，通过构造特征模板进行文本特征提取，然后由机器学习模型预测命名实体的概率。机器学习算法能够从数据中自动学习特征，往往比基于词典和规则的确定性信息抽取模型效果要好。在中医文本 NER 任务中常用的统计机器学习方法包括条件随机场(CRF)、隐马尔科夫模型(HMM)、支持向量机(SVM)、最大熵模型(ME)等。2009 年，王世昆等[12]率先提出基于 CRF 的中医文本 NER 方法，在医案数据上效果明显优于 ME 和 SVM 方法。自此，CRF 成为这一时期中医文本 NER 的主流模型，应用到网络信息[13]和古籍[14]等多种文本。

在文本规模相对有限的情况下，机器学习模型往往能取得较好的效果。机器学习算法的首要问题在于特征工程的复杂性，设计恰当的特征以从原始数据中有效提取代表性信息是一大挑战。此外，机器学习模型的泛化能力有限，在面对未见过的数据时，模型表现可能会大幅下降。

3.2 深度学习方法

基于神经网络的深度学习方法突破了传统机器学习算法的局限性，能够有效的利用深层次语义信息。本节首先介绍基于深度学习的中医文本 NER 模型的一般框架；随后分别详细探讨 3 种架构：基于序列标注的方法、基于跨度的方法和基于大语言模型的方法。

3.2.1 基于深度学习的中医文本 NER 模型的一般框架

基于深度学习的 NER 模型通常由 3 个主要部分构成：嵌入层(Embedding Layer)、编码层(Encoding Layer)和预测层(Prediction Layer)，其总体架构如图 1 所示。

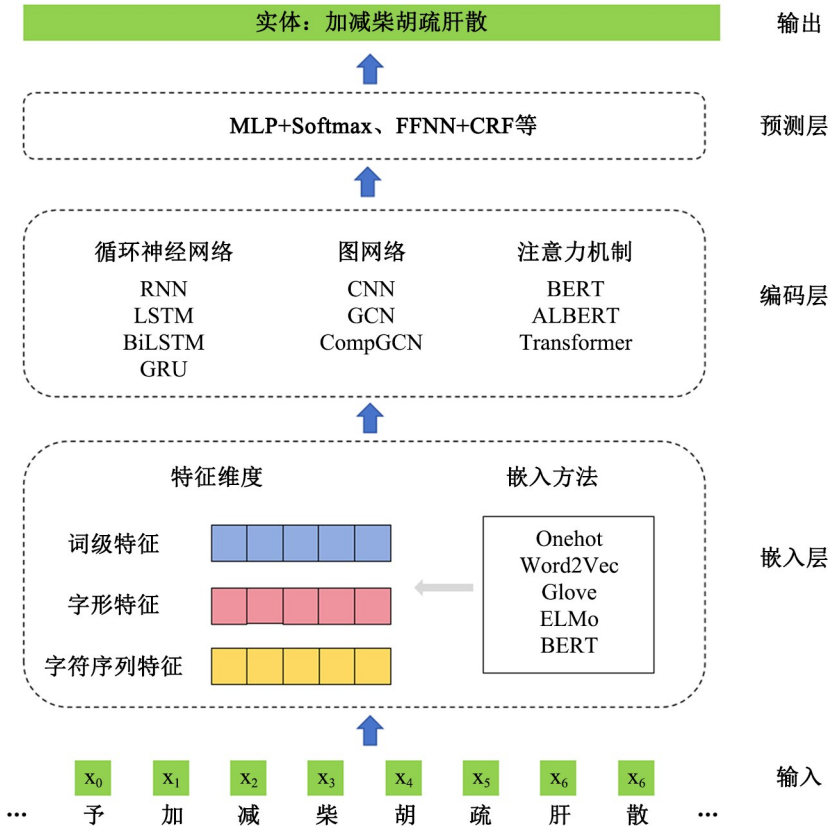

图 1 基于深度学习的中医文本 NER 模型的一般架构

Fig. 1 General Architecture of Deep Learning-Based TCM Texts NER Model

1）嵌入层：主要任务是将原始文本转化为可供模型处理的低维稠密表示。嵌入向量的生成方法主要分为两类：基于特征的实现和基于微调的实现。在基于特征的实现中，嵌入向量在模型训练期间中不参与参数更新，如 One-hot、Word2Vec 和 GloVe 等。基于微调的方法使用上下文相关的嵌入表示，在训练模型过程中不断优化参数，动态调整表示向量，其典型代表是 BERT 嵌入。BERT 基于 Transformer 架构，通过自注意力机制捕捉序列中的长距离依赖关系，能够有效捕捉语言的上下文信息和语义特征。

BERT 等通用预训练模型得到的文本嵌入结果携带的是通用语料中的语义信息，难以充分理解中医领域的专业术语。近年来，构建更为贴近中医领域的预训练模型逐渐受到重视。谢靖等[15]使用基于繁体《四库全书》的 SikuBERT 和 SikuRoBERTa，以 Flat-Lattice Transformer(FLAT)结构为微调模型，验证了基于古文的预训练模型在中医 NER 任务中明显优于通用 BERT 模型。更进一步的，直接利用中医领域数据训练模型，能更有效捕捉中医文本中独特的语义信息。王亚强等[16]构建了中医临床记录语料库，对 MC-BERT 进行领域微调，验证了专有预训练模型对中医文本 NER 效果的提升作用。

2）编码层：编码层对嵌入层输出的向量进行处理，捕捉输入序列中的上下文信息。常见模型包括多卷积神经网络(CNN)、循环神经网络(RNN)和双向长短期记忆神经网络(BiLSTM)等。BiLSTM 网络在中医 NER 任务中得到最为广泛的应用。BiLSTM 由两个 LSTM 层组成，同时捕捉序列的正向和反向信息，通过门控机制保留长距离依赖关系，能克服传统 RNN 在处理长期依赖时存在的梯度消失问题。然而，BiLSTM 在局部特征建模上的能力往往不如 CNN。Ma Y 等[17]认为，古代中医文本中一些知识的表达式是以短语的形式呈现出来的，且缺乏完整的语法结构，在构建中医古籍 NER 模型时应该同时考虑输入文本的上下文语义特征和局部语义特征，增强模型的语义判别能力。

近年来，除了基于 CNN 和 RNN 的编码方法外，还涌现了一些新的方法。例如，基于条件生成对抗网络(cGAN)的 NER 模型能够应对中医 NER 标注数据较少的问题[18]。该模型通过改进的 U-Net 结构，从单词、句子、段落和章节中提取多粒度的语法和语义特征，并通过跳跃连接结合低层和高层特征，增强了生成过程中的特征表达。

3）预测层：负责将经过编码处理后的特征映射到具体的实体类别。中医文本 NER 最常采用 CRF 作为逻辑回归层，在标签之间建立关联性约束，保证预测标签的合理性，使模型生成全局最优序列标注[19]。此外，预测层还可以结合多种模型优化策略。Zhao Z 等[20]提出的基于动态优化的集成学习方法根据预测损失调整模型集成学习的实体类别和融合权重，并在实体稀疏时减少参数更新的幅度，防止模型受到非实体信息的不当干扰。Feng Y 等[21]提出的 ANeTCM 模型将序列标注转换为机器阅读理解任务，结合门控线性单元(GLU)提高模型的特征学习能力，在预测层利用正态分布来调整样本的权值，以解决实体类的不平衡问题。

3.2.2 基于序列标注的方法

基于序列标注的 NER 方法根据上下文为输入序列中的每个元素分配一个标签，表示其是否属于某个实体及其在实体中的位置。BERT-BiLSTM-CRF 是应用最为广泛的基于序列标注 NER 架构。该方法通过 BERT 将输入文本转换为嵌入向量，使用 BiLSTM 进行编码，通过全连接层进行分类，最后利用 CRF 最大化标签序列的联合概率。部分研究基于这一架构进行算法的局部改进。Hou J 等[22]提出的 Dyn-AttNet 模型引入了动态注意力和并行结构。针对中医领域存在的生僻词识别率较低的问题，Jin Z 等[23]提出的 TCMKG-LSTM-CRF 模型利用知识图谱信息进行增强学习，引入知识注意力向量模型，增强模型学习和识别生僻词的能力。此外，还有结合残差网络和归一化模型的 BERT-BiLSTM-CRF 医案症状及药物实体抽取模型[24]、基于自适应词嵌入 RoBERTa-WWM-BiLSTM-CRF 的名中医临床病例 NER 模型[25]等。

基于序列标注的 NER 模型能够充分利用上下文信息，识别准确性较高，并且适用于多种文本类型，具有较好的通用性。这种方法存在的问题是在处理长文本、复杂边界实体和类型不平衡等问题方面可能受到挑战。

3.2.3 基于跨度的方法

基于跨度的 NER 方法将实体识别视为一个跨度分类任务。具体而言，该方法通过定义最长跨度 L 或使用某种策略生成可变长度跨度，列举所有可能的文本跨度(即连续的字符序列)，生成候选实体。对每个跨度进行特征提取，生成表示向量。在解码阶段，利用跨度分类器预测跨度的实体类别。基于跨度的方法没有明确的边界监督，可能导致边界信息利用不足的问题。Xu W 等[26]在跨度内部的词嵌入基础上，将跨度的起始和结束位置的隐藏特征作为显式特征加入到跨度表示中；同时，使用 BiLSTM 捕捉跨度上下文信息增强特征表示；最后使用多关系图卷积网络(CompGCN)进行跨度预测。

基于跨度的模型适应性较强，能够灵活处理嵌套实体等复杂实体类型。然而，枚举所有可能的跨度复杂性较高，导致大量低质量候选跨度，从而需要较多的计算资源来训练高性能的分类器。

3.2.4 基于大语言模型的方法

大语言模型(LLM)对 NLP 的发展产生了深远的影响。大语言模型的语义理解和常识推理能力较强，有助于充分分析上下文信息，从而更准确的进行实体抽取。大语言模型在 NER 中的应用仍然处于探索阶段。张颖怡等[27]的研究表明，基于 ChatGPT 的学术论文实体识别 F1 值高于由少量样本训练得到的神经网络模型。鲍彤等[28]在 MSRA 等常用数据集上测评了 ChatGPT 的信息抽取能力，结果其在 NER 中的表现不及 GlyceBERT 和 ERNIE3.0 模型，表明 ChatGPT 在典型的中文信息抽取任务上还有很大改进空间。

在中医 NER 研究中，李盼飞等[29]调用文言一心 API 进行了医案命名实体的自动化抽取，对结果进行了初步探索，但并未进行系统的效果评价。何宇浩等[30]则对比了 CasRel、GPLinker 与 GPTs 在抽取《中华医方》中“太阳病”方剂名、书名、中药名和剂量实体上的效果，结果显示，ChatGPT4.0 的表现最佳，其综合 F1 值达到 97.48%。在大语言模型的研究热潮下，许多研究团队、机构和企业相继推出了一系列中医领域专用的大语言模型，如“岐黄问道·大模型”“神农中医药大模型”“本草”等。尽管这些模型在中医 NER 中尚未有公开的研究成果，但它们仍然展现出巨大的潜力，有望在中医知识抽取上实现更高的效率。

LLM 可减少人工标注语料的工作，并且无需繁琐的训练过程和大量的计算资源。在使用时 ChatGPT 等 LLM 时，用户主要通过设计提示词(Prompt)来发出指令，引导模型关注特定的实体类型。提示词的优劣直接影响到模型的性能，如何设计出合适的提示词是 LLM 产生高质量回答的关键。

3.3 中医文本 NER 研究热点

3.3.1 多特征融合模型

多特征融合模型是近年中医文本 NER 研究的主要方向之一，相关研究如表 5 所示。多特征融合模型在字符的基础上，以词汇、拼音和形态学特征等作为补充，从模型底层减少中医语义信息的丢失。由于现有的分词技术对中医文本的处理效果不佳，中医文本 NER 通常以字作为基本的标记单元，避免分词带来的歧义性问题。词汇增强的算法引入高质量的领域词典，弥补基于字符向量的特征可能会导致文本序列中蕴含的词汇语义信息丢失的缺陷。词汇增强方法主要有适应嵌入和动态框架两种范式。适应嵌入范式仅在嵌入层对词汇信息进行自适应嵌入，不改变模型本身的结构，典型代表为基于 Soft-Lexicon 的词典匹配方法[46]。动态框架范式则通常需要设计相应的模型结构，以融入词汇信息，典型代表为 Lattice LSTM。Lattice LSTM 在基于字的 LSTM 模型上加入了词汇输入单元，可以有效地利用词的先验知识[31]。受到 Lattice LSTM 和 Transformer 的启发，Flat-Lattice Transformer 构建位置编码重构原有的 Lattice 结构。叶青等[32]采用 Flat-Lattice Transformer 模型融合了字、词和跨度特征，提高了模型对边界模糊实体的处理能力。

此外，汉语作为象形文字，其字形具有一定的规律性。具有相似偏旁或部首的汉字在语义上往往存在一定的相关性，例如，带有“艹”或“木”部首的字通常与本草相关(“芝”“藿”等)，而“疒”部首的字一般与疾病相关(“疟”“痛”等)。融合字形特征，能够使具有关联的字符在向量空间中更为接近。胡为等[33]通过融合汉字笔画、部首和词根等字形特征为字符赋予更丰富的语义特征，相比以往方法 F1 值提高了 3.0%。

表 5 中医文本命名实体识别多特征融合模型

Tab. 5 Multi-Feature Fusion Models for NER in TCM Texts

文献	特 征	模 型	数 据	实体类型	识别效果
[13]	字、词	BERT-BIGRU-CRF	古籍	药名、药性、药味、归经、别名、症状、功效、古籍	P：89.54%、R：91.56%、F1：90.54%
[25]	字、词	BERT-BiLSTM-CRF、IDCNN-CRF	医案	病因、方剂、药物、治疗方法、证型、症状	A：97.49、P：91.47%、R：93.73%、F1：92.59%
[34]	笔画、部首、词根	BERT-BiLSTM-CRF	医案	症状、辨证、治则、方药、人群、功效	P：93.2%、R：92.8%、F1：92.9%
[35]	字、词	Word2Vec-BiLSTM-CRF	古籍、教材	方剂、中药、证型、症状、病因	P：86.08%、R：88.33%、F1：87.19%
[36]	字、词	LEBERT-BiLSTM-CRF	医案	病名、性别、年龄、诱因、症状、舌象、脉象、证候、治法、治方	P：88.69%、R：87.4%、F1：88.1%
[37]	字、部首	BERT-BiLSTM-CRF	医案	身体部位、药物、症状、疾病	P：84.26%、R：85.37%、F1：84.81
[38]	字、词、部首	SiKuBERT-MECT	古籍	人名、中草药物名、病症名、病理名、经络名	P：86.95%、R：86.37%、F1：86.66%
[39]	字、词	ALBERT-BiLSTM-CRF	针刺文献	疾病、针刺部位、症状、证候、效应指标、针法、刺法	P：92.57%、R：91.42%、F1：91.85%
[40]	字、词、部首	Bert-ancient-Chinese-BiLSTM-CRF	医案	疾病、症状、舌象、脉象、证候、治法、方剂及中药	P：90.09%、R：90.61%、F1：90.35%

3.3.2 面向低资源问题的模型

目前的中医文本 NER 方法仍然对大规模的标记训练数据有很强的依赖性。中医领域的公开语料库较少。因此，解决低资源环境下的中医文本 NER 问题具有一定的挑战性。通用领域解决这一问题的主流方法是迁移学习，利用源域中已有的知识来帮助目标域的学习任务。中医文本与通用文本之间存在显著的差异，增大了将其他领域知识迁移到中医 NER 任务的难度，因此基于迁移学习的中医文本 NER 研究较少。目前，低资源环境下的中医文本 NER 相关研究主要包括数据增强、半监督学习和远程监督学习等方法。

数据增强在原始数据集的基础上，通过同义词替换、变换语序、随机删除、添加扰动等方法生成更多的训练样本。杨延云等[41]采用 EDA 数据增强方法进行数据扩充，结合半监督自训练学习，解决中医文本 NER 标注数据集较小的问题。Zhao Z 等[42]设计了一系列采样和数据增强策略，以缓解实体不平衡带来的挑战。远程监督方法利用领域实体词典和原始文本来自动生成“银标准”数据集(Silver Standard Datasets)，可以快速获得大规模标注数据，解决“黄金标准”数据集(Gold Standard Datasets)标注成本较高的问题。远程监督方法最关键的问题是假阴性样本的存在。Jia Q 等[43]将远程监督中医文本 NER 任务视作跨度检测任务，提出了一种针对银标准数据集的负采样策略。在训练阶段，该方法在每

个周期随机选择一定数量的非实体文本作为负样本，通过标签平滑减少假阴样本对训练的不良影响。

3.4 评价指标

中医文本 NER 的评价指标主要包括准确率、精确率、召回率和 F1 均值。假定 TP 表示模型成功识别的标记实体(真正例)；TN 表示模型未识别的非标记实体(真负例)；FP 表示模型错误识别的非标记实体(假正例)；FN 表示模型未识别的标记实体(假负例)。则评价各指标的定义如下：

准确率指模型正确识别的实体数占所有实体总数的比例，即式（1）：

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \tag{1}$$

精确率指模型正确识别的实体数占识别实体总数的比例，即式（2）：

$$Precision=\frac{TP}{TP+FP} \tag{2}$$

召回率指模型正确识别的实体数占标记实体总数的比例，即式（3）：

$$Recall=\frac{TP}{TP+FN} \tag{3}$$

F1 均值兼顾准确率和召回率之间的平衡，即式（4）：

$$F1=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{4}$$

各指标越接近于 1，表示 NER 模型的识别性能越好。

4 讨论与展望

深度学习算法的不断发展提升了中医文本 NER 的效果。然而，中医文本 NER 依然面临许多挑战。基于这些分析，我们建议未来研究可以从下述几个方面展开。

4.1 语料库建设

中医药领域可用于 NER 及其相关任务的高质量标注数据集相对匮乏，相关研究通常依赖于自行标注的小规模数据集进行封闭训练。新的 NER 研究往往需要重建语料库，导致模型缺乏可比性、可移植性和通用性。解决这一问题需要制定标注规范，并构建高质量数据集。

1）制定标注规范。缺乏统一的标注规范已成为制约领域数据迁移与融合的关键因素。制定公认的、系统化的实体规范，需要构建中医术语映射字典，确保标注的一致性。相应的，需合理的设计嵌套实体、不连续实体等复杂问题的标注方案，增强 NER 模型在真实场景中的适应性和鲁棒性。一种可能的方案是在标注上突破单一参数，构建多层次的标注结构。这不仅应包括兼容粗细粒度的分层标注，还可结合词性、句法功能与短语规则，利用多层次信息约束实体识别过程[44]。

2）构建高质量标注数据集。深度学习模型，特别是监督学习模型，需要大规模、高质量的标注数据集，构建普遍认可的标准数据集的重要性程度不言而喻。生物医学领域的数据集构建方式具有较强的参考价值。生物医学领域拥有丰富的公开资源，如 CCKS 数据集、CLUENER 数据集、NCBI-Disease 数据集以及瑞金医院糖尿病数据集等。这些数据集大多源于自然语言处理测评比赛任务，由专业团队进行数据的整理与标注，确保了数据的可靠性。相关数据集覆盖多种类型的实体，例如，CHIP 数据集支持嵌套实体的标注，而 ShARe 数据集则面向非连续实体的识别。相较而言，中医 NER 数据集的数量和质量均显著不足。2020 年，中医药天池大数据竞赛发布了中医药说明书实体识别数据集，共包含1 997份药品说明书，涵盖药品、药物成分、疾病、症状等 13 类实体。未来中医 NER 研究将依托高质量的标注数据集，进一步推动中医药信息化的发展。

4.2 小样本学习中的数据优化

在数据有限的情况下，利用数据优化技术能够提升模型的学习能力和泛化性能。

1）基于 GPT 的数据增强。GPT 模型的发展为数据增强提供了一种新的解决方案。GPT 模型能够对输入文本进行修改或重构，生成符合上下文逻辑的新样本。2023 年，Dai H 等[45]提出，基于 GPT 的增强算法 AugGPT，将训练样本中的每个句子重述为多个概念相似但语义不同的样本，该方法在测试精度和增强样本分布方面优于最先进的文本数据增强方法。2024 年，许钦亚等[46]应用 ChatGPT 对学术论文语步数据进行增强，提出 GPT 数据增强提示工程的角色设定与任务描述、任务要求描述、制定返回格式、设置任务示例和输入与评估六大步

骤。基于 GPT 的数据增强是未来中医 NER 数据增强研究的一个可能方向。

2）主动学习。主动学习算法通过选择价值密度最高的数据样本，筛选合适的候选集，再进行人工标记，减少所需标注数据量，降低标注成本。标注后的数据通过增量或重新学习的方式融入模型，再循环往复中提高模型的学习效果。Li T 等[47]提出一个对抗性的主动学习框架来选择最有价值的标注实例，结合 LSTM、BiLSTM 和注意力机制进行网络安全文本中的命名实体检测，以较低的标注成本增强了模型的效果。Tran V 等[48]使用基于实例的上下文和内容的多样性来选择信息最为丰富的实例，结合自学习算法筛选高度可靠的实例，在 Twitter 数据集 NER 中取得了较好的效果。这些研究证实了主动学习的有效性，主动学习与深度学习相结合可能是降低中医 NER 数据标注成本的一种可行方案。

4.3 针对复杂实体的识别模型

中医 NER 中嵌套实体、非连续实体和易混淆实体等复杂实体依然是主要的挑战。

1）嵌套实体。嵌套实体是命名实体中的一种特殊现象，指某个实体内部包含另一个实体的情况。假设输入文本序列 $X=\{x_1,x_2,\cdots,x_n\}$，其中 x_i 是序列中的第 i 个字，n 为序列的长度。对于非嵌套命名实体而言，每个字对应一个实体标签，标签集合可表示为 $Y=\{y_1,y_2,\cdots,y_n\}$。与此不同的是，嵌套实体中每个字可能对应多个标签，标签集合可表示为 $Y=\{\{y_1^1,y_1^2,\cdots,y_1^m\},\{y_2^1,y_2^2,\cdots,y_2^m\},\cdots,\{y_n^1,y_n^2,\cdots,y_n^m\}\}$，其中，$n$ 为序列的长度，m 为嵌套的层数。嵌套实体的嵌套结构复杂多变，嵌套颗粒度和嵌套层数缺乏规律性，例如，“麻黄桂枝汤”由多个非嵌套实体“麻黄”和“桂枝”并列构词，而“杏子汤”则是非嵌套实体“杏子”的扩展。嵌套实体包含的内部实体之间还可能存在依赖关系。因此，嵌套实体的识别难度较大，需要改进现有模型以提高识别准确率。嵌套命名实体识别是各领域信息抽取任务的一个研究热点，其他领域相关研究提出了基于超图表示[49]、状态转换[50]、二部平面图[51]等多种方法。Xu H 等[52]设计了针对中医 NER 的两层标注策略，对中医嵌套命名实体识别做出了探索。未来的研究需要在现有模型的基础上进行改进，更好地支持知识库构建等后续工作。

2）非连续实体。非连续实体是指在文本中由不相邻的字或词组成的实体。传统的 BIO 等标注规范无法有效支持非连续命名实体识别。针对这一问题，Tang B 等[53]提出了 BIOHD 标注法，该方法在 BIO 的基础上增加了 HB、HI、DB、DI 4 种标签，以表示不规则实体。其中 HB 和 HI 用于标注重叠实体，DB 和 DI 则用于标注非连续实体，DB 表示非连续实体的首字，DI 表示不连续实体的中间和尾部字。这一创新使非连续实体识别更为清晰和精确。Dai X 等[54]提出一种端对端的基于转移的神经编码模型，并利用专门的行动和注意力机制来确定特定跨度是否是非连续实体的组成部分，该方法能够在不牺牲连续实体识别准确性的前提下有效的识别不连续实体，对后续研究产生了极大的影响。中医非连续实体识别的研究较为缺乏，尤其是超图等新兴方法的有效性尚待验证。因此，未来研究应聚焦于这一领域，以探索更有效的识别策略。

3）易混淆实体。中医文本中不同实体的识别效果差异显著，相对而言，药物、部位的识别率较高，而症状、病名及病症等实体的识别精度较低。中医疾病术语往往散落在症状词中，且部分实体既可表示疾病，又可作为症状。这种模糊性大大增加了模型的识别难度，需要高度的上下文理解能力才能做出区分。症状词的表述极为丰富，在与不同的程度词结合后，表述更为复杂。此外，这些实体兼有上述的嵌套、不连续等复杂现象，实体边界不清，极大地影响了识别的精度。针对中医易混淆命名实体的识别，需要建立更为统一的数据标准和有效的标注策略，同时增强模型对上下文信息的理解能力。

4.4 增强模型解释性

深度学习方法的“黑箱”性质限制了相关人员对模型内部工作过程的理解。在临床诊断和治疗的现实过程中，错误的决定可能会产生非常严重的后果。模型的可解释性关系到模型结果的可信度，是确保其在实际应用中被采用的关键因素。因此，提升中医 NER 模型的可解释性是未来的一个研究重点。这需要从规则制定、内部模块解释、归因解释和实例分析等多个角度对模型进行解构。同时，也要制定科学的评价指标，衡量不同类型模型的解

释程度。

5 结 语

中医命名实体识别为中医知识组织和智慧医疗奠定了基础。由于中医文本固有的抽象性、经济性和复杂性等特征，中医文本 NER 面临知识体系复杂、语料库稀缺和技术算法效果有待提升等挑战。中医文本 NER 技术经历了从基于词典和规则的模式匹配方法、基于统计原理的机器学习方法和基于神经网络的深度学习方法的发展路径。目前，中医文本 NER 的主流方法是基于 BERT-BiLSTM-CRF 的序列标注方法，基于跨度的方法也有一定的研究。近年来，基于大语言模型的中医文本 NER 技术展现了一定的潜力。此外，中医领域的专有预训练模型、融合字词和字形等特征的融合模型和面向低资源问题的模型取得了不错的效果。未来的中医文本 NER 研究需要着重处理中医语料资源匮乏问题，制定统一的语料标注规范，构建高质量中医标注数据集；同时，小样本学习中的数据优化、针对复杂问题的识别模型和深度学习模型的解释性研究可能成为新的技术发展趋势。

参 考 文 献

[1] 刘丽莉，李明，罗晓兰，等. 基于自然语言处理智能技术的中医术语研究文献计量分析 [J]. 上海中医药杂志，2024，58（7）：1-6，14.

[2] 孔静静，于琦，李敬华，等. 实体抽取综述及其在中医药领域的应用 [J]. 世界科学技术-中医药现代化，2022，24（8）：2957-2963.

[3] 易钧汇，查青林. 中医症状信息抽取研究综述 [J]. 计算机工程与应用，2023，59（17）：35-47.

[4] 李虹. 中医语言的特点及其对中医英语表达的影响 [J]. 上海中医药大学学报，2006，（1）：69-71.

[5] 丁有伟，郭坤，胡孔法，等. 一种面向中医电子病历的实体抽取算法 [J]. 软件导刊，2021，20（12）：99-104.

[6] 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）：58-63.

[7] 佟琳，张华敏，佟旭，等. 基于命名实体识别的《神农本草经》知识图谱构建及可视化分析 [J]. 中国中医药信息杂志，2024，31（8）：37-43.

[8] 周嘉玮，王坤，吴雨璐，等. 基于 BiLSTM-CRF 的《神农本草经》命名实体识别研究 [J]. 成都中医药大学学报，2024，47（3）：54-59.

[9] 马月坤，吴国仲. 基于特征增强的中医本草命名实体识别方法 [J]. 河北大学学报（自然科学版），2024，44（2）：199-207.

[10] Wang Y，Yu Z，Jiang Y，et al. A Framework and its Empirical Study of Automatic Diagnosis of Traditional Chinese Medicine Utilizing Raw Free-Text Clinical Records [J]. Journal of Biomedical Informatics，2012，45（2）：210-223.

[11] 邓宇，张振铭，陈橙，等. 基于正则表达式的中医医案术语抽取方法研究 [J]. 湖南中医杂志，2023，39（5）：202-207.

[12] 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.

[13] 王莉军，李旭婕，刘志辉，等. 基于开放信息源的实体挖掘方法研究 [J]. 情报科学，2019，37（8）：139-144.

[14] 李贺，祝琳琳，刘嘉宇，等. 基于本体的简帛医药知识组织研究 [J]. 图书情报工作，2022，66（22）：16-27.

[15] 谢靖，刘江峰，王东波. 古代中国医学文献的命名实体识别研究——以 Flat-lattice 增强的 SikuBERT 预训练模型为例 [J]. 图书馆论坛，2022，42（10）：51-60.

[16] 王亚强，李凯伦，舒红平，等. 基于批数据过采样的中医临床记录四诊描述抽取方法 [J]. 中文信息学报，2024，38（2）：121-131.

[17] Ma Y，Liu H，Liu Y，et al. A Named Entity Recognition Method Enhanced with Lexicon Information and Text Local Feature [J]. Computer Science，Medicine，2023，20（3）：899-906.

[18] Ma Y，Liu Y，Zhang D，et al. A Multigranularity Text Driven Named Entity Recognition CGAN Model for Traditional Chinese Medicine Literatures [J]. Computational Intelligence and Neuroscience，2022：1495841.

[19] 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.

[20] Zhao Z，Qian Y，Liu Q，et al. A Dynamic Optimization-Based Ensemble Learning Method for Traditional Chinese Medicine Named Entity Recognition [J]. IEEE Access，2023，11：99101-99110.

[21] Feng Y，Zhou Y. ANeTCM：A Novel MRC Framework for Traditional Chinese Medicine Named Entity Recognition [J]. IEEE Access，2019，12：113235-113243.

[22] Hou J，Saad S，Omar，N. Enhancing Traditional Chinese Medical Named Entity Recognition with Dyn-Att Net：A Dynamic Attention Approach [J]. PeerJ Computer Science，2024，10：e2022.

[23] Jin Z，Zhang Y，Kuang H，et al. Named Entity Recognition in Traditional Chinese Medicine Clinical Cases Combining BiLSTM-CRF with Knowledge Graph [J]. Knowledge Science，Engineering and Management，2019，11775：537-548，2019.

[24] 王欣宇，高晓苑，杨涛，等. 名老中医诊治肺癌“症-药”关系自动化提取与分析模型构建及应用 [J]. 中华中医药杂志，2022，37（11）：6297-6301.

[25] 万泽宇，龚庆悦，李铁军，等. 基于自适应词嵌入 RoBERTa-wwm 的名中医临床病历命名实体识别研究 [J]. 软件导刊，2022，21（12）：58-62.

[26] Xu W，Wang L，Zhang M，et al. A Joint Entity Relation Extrac-

tion Method for Document Level Traditional Chinese Medicine texts [J]. Artificial Intelligence In Medicine, 2025, 154: 192015.

[27] 张颖怡, 章成志, 周毅, 等. 基于 ChatGPT 的多视角学术论文实体识别: 性能测评与可用性研究 [J]. 数据分析与知识发现, 2023, 7 (9): 12-24.

[28] 鲍彤, 章成志. ChatGPT 中文信息抽取能力测评——以三种典型的抽取任务为例 [J]. 数据分析与知识发现, 2023, 7 (9): 1-11.

[29] 李盼飞, 杨小康, 白逸晨, 等. 基于大语言模型的中医医案命名实体抽取研究 [J]. 中国中医药图书情报杂志, 2024, 48 (2): 108-113.

[30] 何宇浩, 李明, 罗晓兰, 等. 基于 GPTs 的中医知识图谱实体和关系抽取研究 [J]. 上海中医药杂志, 2024, 58 (8): 1-6.

[31] 曾江峰, 庞雨静, 高鹏钰, 等. 基于 Lattice LSTM 的中医药古文献命名实体识别与应用研究 [J]. 情报工程, 2023, 9 (5): 112-122.

[32] 叶青, 赖煊, 程春雷, 等. 融合词汇增强和跨度方法的中医药命名实体识别 [J/OL]. 计算机工程与应用, 1-10 [2024-08-27]. http://kns.cnki.net/kcms/detail/11.2127.tp.20240824.1025.002.html.

[33] 胡为, 刘伟, 盛威, 等. 融合字形特征的中医医案命名实体识别研究 [J]. 计算机时代, 2023, (7): 66-69, 73.

[34] 胡为, 刘伟, 盛威, 等. TcmYiAnBERT: 基于无监督学习的中医医案预训练模型 [J]. 医学信息学杂志, 2023, 44 (7): 63-67.

[35] 杨延云, 杜建强, 聂斌, 等. 一种面向中医文本的实体关系深度学习联合抽取方法 [J]. 计算机应用与软件, 2023, 40 (3): 217-222, 234.

[36] 李旻哲, 殷继彬. 融合 BERT 模型与词汇增强的中医命名实体识别模型 [J]. 计算机科学, 2024, 51 (S1): 134-139.

[37] 刘彬, 肖晓霞, 邹北骥, 等. 融合汉字部首的 BERT-BiLSTM-CRF 中医医案命名实体识别模型 [J]. 医学信息学杂志, 2023, 44 (6): 48-53.

[38] 张文东, 吴子炜, 宋国昌, 等. 基于 SiKuBERT 与多元数据嵌入的中医古籍命名实体识别 [J]. 华南理工大学学报 (自然科学版), 2024, 52 (6): 128-137.

[39] 王晰, 柯丽娟, 李海燕, 等. 基于“深度学习模型+词典”的针刺效应命名实体识别研究 [J]. 世界科学技术-中医药现代化, 2024, 26 (7): 1779-1785.

[40] 杨航, 彭叶辉, 杨伟, 等. 基于 BRL 神经网络模型的名家医案实体识别 [J]. 中国实验方剂学杂志, 2024, 30 (24): 167-173.

[41] 杨延云, 杜建强, 聂斌, 等. 融合数据增强和注意力机制的中医实体及关系联合抽取 [J]. 智能计算机与应用, 2023, 13 (8): 186-191, 196.

[42] Zhao Z, Tang Y, Cheng Z, et al. ABL-TCM: An Abductive Framework for Named Entity Recognition in Traditional Chinese Medicine [J]. IEEE Access, 2024: 3454278.

[43] 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.

[44] 闻永毅, 王治梅. 中医文献语料库建设与顶层设计刍议 [J]. 西部中医药, 2018, 31 (7): 62-65.

[45] Dai H, Liu Z, Liao W, et al. AugGPT: Leveraging ChatGPT for Text Data Augmentation [J]. arXiv: 2302.13007.

[46] 许钦亚, 薛秋红, 钱力, 等. 融合 ChatGPT 数据增强的学术论文语步识别方法研究 [J]. 图书情报工作, 2024, 68 (17): 84-94.

[47] Li T, Hu Y, Ju A, et al.. Adversarial Active Learning for Named Entity Recognition in Cybersecurity [J]. Computers, Materials & Continua, 2021, 66 (1): 407-420.

[48] Tran V, Nguyen N, Fujita H, et al. A Combination of Active Learning and Self-learning for Named Entity Recognition on Twitter Using Conditional Random Fields [J]. Knowledge-Based Systems, 2017, 132 (15): 179-17.

[49] Wang B, Lu W. Neural Segmental Hypergraphs for Overlapping Mention Recognition [C] //Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium: Association for Computational Linguistics, 2018: 204-214,

[50] Wang B, Lu W, Wang Y, et al. A Neural Transition-based Model for Nested Mention Recognition [C] //Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium: Association for Computational Linguistics, 2018: 1011-1017.

[51] Luo Y, Zhao H. Bipartite Flat-Graph Network for Nested Named Entity Recognition [C] //Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online: Association for Computational Linguistics, 2020: 6408-6418.

[52] Xu H, Liu H, Jia Q, et al. A Nested Named Entity Recognition Method for Traditional Chinese Medicine Records [J].

[53] Tang B, Hu J, Wang X, et al. Recognizing Continuous and Discontinuous Adverse Drug Reaction Mentions from Social Media Using LSTM-CRF [J]. Wireless Communications and Mobile Computing, 2018: 2379208.

[54] Dai X, Karimi S, Hachey B, et al. An Effective Transition-based Model for Discontinuous NER [C] //Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online: Association for Computational Linguistics, 2020: 5860-5870.

(责任编辑: 郭沫含)