doi:10.3772/j.issn.2095-915x.2023.05.009

基于 Lattice LSTM 的中医药古文献命名实体识别与应用研究

开放科学（资源服务）标识码（OSID）

曾江峰[1] 庞雨静[2] 高鹏钰[1] 冯昌扬[1,3]

1. 华中师范大学信息管理学院 湖北 武汉 430079；
2. 北京理工大学管理与经济学院 北京 100081；
3. 富媒体数字出版内容组织与知识服务重点实验室 北京 100038

摘要：[目的 / 意义] 为进一步提升中医药古文献命名实体识别的准确性，以信息化手段辅助现代中医学者进行医学诊断与临床决策，促进中医学的传承与创新。[方法 / 过程] 提出一种集成字符与词汇信息的中医药古文献命名实体识别的 Lattice LSTM 模型，对《伤寒论》的疾病、证候、方剂、症状和药材五类实体进行抽取；在抽取出的实体基础上，人工提取实体间关系，利用 Neo4j 搭建了中医药知识图谱；最后以新冠肺炎为例，在图谱上完成相关检索。[结果 / 结论] 实验结果表明，Lattice LSTM 在中医术语识别上性能最优，F1 值达到 95.66%，比主流模型 BiLSTM-CRF 提升了 1.68%，可用于中医药古文献的实体识别；搭建的中医药知识图谱也验证了主模型的现实价值。

关键词：Lattice LSTM；中医药古文献；命名实体识别；知识图谱

中图分类号：G35; R2-03; TP391.1

Research on Named Entity Recognition and Application of Traditional Chinese Medicine Ancient Literature Based on Lattice LSTM

ZENG Jiangfeng[1] PANG Yujing[2] GAO Pengyu[1] FENG Changyang[1,3]

1. Central China Normal University, School of Information Management, Wuhan 430079, China;
2. Beijing Institute of Technology, School of Management and Economics, Beijing 100081, China;
3. The Key Laboratory of Rich-media Knowledge Organization and Service of Digital Publishing Content Institute of Scientific, Beijing 100038, China

基金项目 教育部人文社会科学研究青年基金“情境大数据驱动的社交媒体虚假信息识别模型与治理策略研究”（21YJC870002）；武汉市知识创新专项项目曙光计划项目“多源知识驱动的社交媒体虚假新闻检测研究”(2022010801020287)；富媒体数字出版内容组织与知识服务重点实验室开放基金项目“面向融合出版的前沿技术主题演化及发展趋势预测研究”。

作者简介 曾江峰（1988-），博士，副教授，研究方向为数据挖掘、自然语言处理等；庞雨静（2000-），硕士研究生，研究方向为数据挖掘、自然语言处理；高鹏钰（2000-），硕士研究生，研究方向为数据挖掘、自然语言处理；冯昌扬（1991-），讲师，研究方向为信息政策分析，E-mail: cyfeng@ccnu.edu.cn。

引用格式 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.

Abstract: [Objective/Significance] In order to further improve the accuracy of named entity recognition in ancient Chinese medicine literatures, using information tools to assist modern Chinese medicine practitioners in medical diagnosis and clinical decision-making, promote the inheritance and innovation of traditional Chinese medicine. [Methods/Processes] This paper proposes a named entity recognition model of ancient Chinese medicine literatures called Lattice LSTM model that integrates character information and lexical information to extract five entities: disease, syndrome, prescription, symptom and medicine in "Treatise on Febrile Diseases". Then on the basis of these extracted entities, the relationships between them are manually extracted, and Neo4j is used to build a knowledge graph of traditional Chinese medicine. Finally, taking the COVID-19 as an example, the graph is used to complete the relevant information retrieval. [Results/Conclusions] The experimental results show that Lattice LSTM has the best performance in the recognition of traditional Chinese medicine terms, with an F1 value of 95.66%, which is 1.68% higher than that of the mainstream model BiLSTM-CRF, so Lattice LSTM can be used for named entity recognition of ancient Chinese medicine literatures. In addition, the constructed knowledge graph of traditional Chinese medicine verifies the realistic value of the main model.

Keywords: Lattice LSTM; Ancient Chinese Medicine Literatures; Named Entity Recognition; Knowledge Graph

引言

中医学凝集着古老深奥的哲学思想和华夏千年来的养生之道，在我国漫长的历史长河中具有不可替代的地位，对世界人类健康作出了突出贡献，其传承和创新近年来受到党和政府的极大关注。习总书记曾在讲话中多次强调：要最大程度上利用和发挥中医药的独特优势，建设健康中国；《“十四五”中医药发展规划》指出，开展中医药振兴的重大工程，推进中医药与现代科学结合。中医药古文献是古中医思维成果的载体，承载了中医基础理论和临床核心知识，极具医学研究价值。近年来，人工智能、物联网等现代技术为中医药产业注入了新鲜活力，利用现代化手法发掘中医古文献资源，能够促进中医学的传承与创新。

中医术语识别是中医文本挖掘的重要方向，而基于自然语言处理的命名实体识别技术是信息化背景下术语识别的主流方法。命名实体识别是知识推荐、智能问答、机器翻译等场景的核心环节，在金融、医疗、教育等多个行业受到广泛关注，是计算机和人工智能领域的热点话题，术语的识别精度极大影响着后继任务的完成质量。然而中医数据量大而繁杂，特别是中医药古文献多为非结构化的文言形式，常存在一词多义及实体边界模糊的问题，相比现代文本识别难度更大，常规模型效果并不理想。探索适用于中医古文献的命名实体识别方法，提升中医术语的识别准确性，十分必要。

本文提出一种中医药古文献命名实体识别的 Lattice LSTM 模型，将字符信息和词汇信息共同输入，实现信息的充分利用且避免了分词错误，大大提升了中医术语识别的准确性，补充完善了自然语言处理技术在中医古籍文本挖掘中的相关理论；实现了现代化信息技术与传统中医学的有效结合，促进了中医学的传承与创新。另外，本文构建的中医药知识图谱，将非结构化的古籍文本以结构化方式存储，便于大规模数据的检索、更新与共享；辅助现代中医学者与临床医师进行医学诊断与临床决策，为广大患者提供更加快捷、优质的中医诊疗服务。

doi:10.3772/j.issn.2095-915x.2023.05.009

1 国内外研究现状

1.1 国外医学领域命名实体识别现状

命名实体识别[1]（Named Entity Recognition, NER）指挑选出给定句子中含义特殊的字或词语，如姓名、地区、职业单位等，是自然语言处理的基础性任务。近年来，国外医学领域的NER研究主要聚焦于深度学习。Gligic等[2]通过迁移学习引导神经网络，在大量未注释的电子健康记录数据集上进行预训练词嵌入，证实了迁移学习和无监督学习可以提升稀疏数据下神经网络的性能；Mulyar等[3]针对临床记录特定任务间信息不共享导致单个方案性能下降的问题，开发了基于深度学习的Multitask-Clinical BERT系统，同时执行实体抽取、模式识别等八项任务，显著提升了计算速度，证实了多任务学习在临床信息提取上的可行性；Kang等[4]结合统一医学语言系统（UMLS）知识，提出医学术语识别的数据增强方法UMLS-EDA，改善了句子分类模型的性能，识别效果优于BERT预训练分类器；Kormilitzin等[5]开发了开源的NER模型Med7，同时引入改进方法sense2vec，利用MIMIC-III数据集证实该模型可用于NER任务；Weber等[6]将HUNER标记器集成到Flair NLP框架中，提出HunFlair医学实体识别模型，在不同数据集的跨语料库评估中性能优于BioBERT、SciSpacy等现有工具；Ma等[7]提出集成两种弱监督学习方法的投票机制，在CCKS2017数据集上的实验效果优于监督学习，缓解了医学标注过少和标注困难的问题。

1.2 我国中医药领域命名实体识别现状

近年来，我国中医药领域的命名实体识别，按照研究对象可分为电子病历、在线医疗社区问答文本、临床医案、中医古籍等。电子病历方面，陈美杉等[8]针对数据标注稀缺的情况，提出KNN-BERT-BiLSTM-CRF，证实知识迁移可以改善小数据集下智能模型的识别效果。临床医案方面，肖瑞等[9]使用BiLSTM-CRF识别部分老中医医案中的术语。中医古籍方面，高甦等[10]以《黄帝内经》为实验材料，提出结合字向量的深度学习模型，解决了古籍实体识别困难的问题；屈倩倩[11]基于BERT-BiLSTM-CRF识别《伤寒论》实体，为后续任务提供了高质量的数据来源；卢克治[12]将Word2vec、ELMo、BERT词嵌入模型引入BiLSTM-CRF，对662本中医古籍进行实体抽取，发现BERT参与的模型获得了最佳识别效果。

综上可知，中医药领域的命名实体识别以电子病历和临床医案为主要研究对象，对于中医药古籍的研究相对较少，且主流方法为BiLSTM-CRF。Zhang等[13]针对中文NER任务首次提出混合字符和词汇信息的Lattice LSTM模型，证实其在现代中文数据集上的卓越性能；崔丹丹等[14]在《四库全书》的古汉语研究中首次运用这一模型，F1值比BiLSTM-CRF提升了3.95%，证实其可用于古籍研究；张笑天[15]将Lattice LSTM用于四川省肿瘤医院病程记录的实体识别，证实其在医学领域NER任务上表现良好，但这一模型目前还未应用于中医领域古文献的研究。基于此，本文提出一种面向中医药古文献术语识别的Lattice LSTM模型，并对

其进行实证研究。

2 Lattice LSTM 模型

Lattice LSTM 在基于字符的 LSTM 模型基础上加入了词汇的输入和额外的门控单元以调控信息流动，其结构见图 1。对于输入的句子 s，既可看作是由单个字组成的字符序列，即 $s=c_1, c_2,..., c_m$，其中 c_j 表示 s 的第 j 个字；也可看作是以词为单位的词序列，即 $s=w_1,w_2,...,w_n$，其中 w_i 表示 s 的第 i 个词。以本文的研究材料《伤寒论》文本数据为例，设 $t(i,k)$ 表示句子中第 i 个词的第 k 个字的位置，如“太阳病发热汗出”一句中的“发”字，其位置可用 $t(3,1)=4$ 表示。

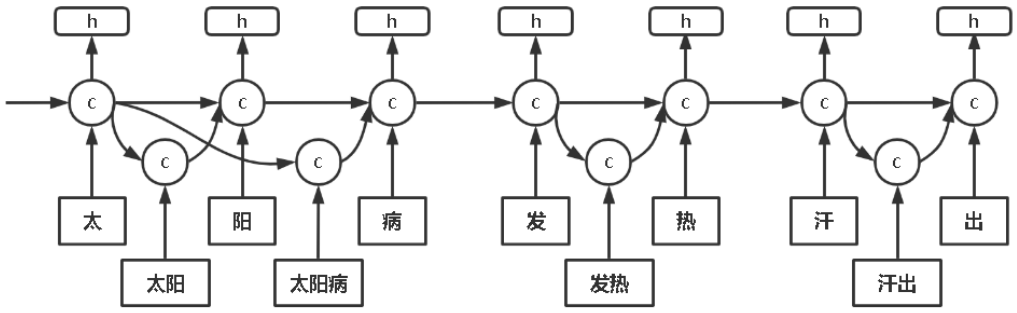

图 1 Lattice LSTM 结构图

Lattice LSTM 包含两种结构：基于字符的结构和基于词的结构。在基于字符的结构中，每个字 c_j 通过字符嵌入矩阵 e^c 表示为输入字向量 $x_j^c=e^c(c_j)$，然后依次通过输入门 i_j^c、遗忘门 f_j^c 和输出门 o_j^c，得到单元向量 c_j^c 和隐藏向量 h_j^c，前者控制句子的历史信息，后者作为 CRF 的输入，基于字符结构的计算如式 (1)、式 (2) 和式 (3) 所示：

$$\begin{bmatrix} i_j^c \\ o_j^c \\ f_j^c \\ \tilde{c}_j^c \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ \sigma \\ tanh \end{bmatrix} \left(w^{c^T} \begin{bmatrix} x_j^c \\ h_{j-1}^c \end{bmatrix} + b^c \right) \tag{1}$$

$$c_j^c = f_j^c \odot c_{j-1}^c + i_j^c \odot \tilde{c}_j^c \tag{2}$$

$$h_j^c = o_j^c \odot tanh(c_j^c) \tag{3}$$

其中 w^{cT} 和 b^c 是模型参数。

Lattice LSTM 的优势在于将潜在词汇信息融合进上述基于字符的结构，从而使得模型在获得字信息的同时，也可以有效地利用词的先验信息。基于词的结构中，每个词 $w_{b,e}^d$ 通过词嵌入矩阵 e^w 得到输入词向量 $x_{b,e}^w = e^w(w_{b,e}^d)$，其中 $w_{b,e}^d$ 表示索引从 b 到 e 的字符组成的词汇。因标记仅在字符级别执行，所以基于词的结构没有输出门，其余与基于字符的结构相似，计算如式 (4)、式 (5) 所示：

$$\begin{bmatrix} i_{b,e}^w \\ f_{b,e}^w \\ \tilde{c}_{b,e}^w \end{bmatrix} = \begin{bmatrix} \sigma \\ \sigma \\ tanh \end{bmatrix} \left(w^{w^T} \begin{bmatrix} x_{b,e}^w \\ h_b^c \end{bmatrix} + b^w \right) \tag{4}$$

$$c_{b,e}^w = f_{b,e}^w \odot c_b^c + i_{b,e}^w \odot \tilde{c}_{b,e}^w \tag{5}$$

其中 w^{wT} 和 b^w 是模型参数。

在上述结构的基础上，使用额外的门控 $i_{b,e}^c$ 将 $c_{b,e}^w$ 包含的词汇信息融入对应字符 c_e^c 结尾的字向量 x_e^c 中，得到当前字符的单元向量 c_j^c，计算方法如式 (6)、式 (7) 所示：

doi:10.3772/j.issn.2095-915x.2023.05.009

$$i_{b,e}^{c}=\sigma\left(w^{l^{T}}\begin{bmatrix}x_{e}^{c}\\ c_{b,e}^{w}\end{bmatrix}+b^{l}\right) \quad (6)$$

$$c_{j}^{c}=\sum_{b\in\{b'|w_{b'}^{d},j\in D\}}\alpha_{b,j}^{c}\odot c_{b,j}^{w}+\alpha_{j}^{c}\odot\tilde{c}_{j}^{c} \quad (7)$$

其中 $\alpha_{b,j}^{c}$ 和 α_{j}^{c} 是 $i_{b,j}^{c}$ 和 i_{j}^{c} 归一化的结果，归一化公式如式 (8)、式 (9) 所示：

$$\alpha_{b,j}^{c}=\frac{exp(i_{b,j}^{c})}{exp(i_{j}^{c})+\sum_{b'\in\{b''|w_{b''}^{d},j\in D\}}exp(i_{b',j}^{c})} \quad (8)$$

$$\alpha_{j}^{c}=\frac{exp(i_{j}^{c})}{exp(i_{j}^{c})+\sum_{b'\in\{b''|w_{b''}^{d},j\in D\}}exp(i_{b',j}^{c})} \quad (9)$$

最终的隐藏向量 h_{j}^{c} 的表达式如 (10) 所示。

$$h_{j}^{c}=c_{j}^{c}\odot tanh(c_{j}^{c}) \quad (10)$$

得到所有输出的隐藏向量 $h_1,h_2,...,h_l$，其中 l 是句子包含的字数，通过 CRF 层得到标记序列 $y=l_1,l_2,...,l_\tau$ 的概率，使用 Viterbi 算法计算得到最大概率值，其对应的标记即为预测标记。标记序列概率的表示方法如式 (11) 所示：

$$P(y|s)=\frac{\exp\left(\sum_{i}\left(w_{CRF}^{l_i}h_i+b_{CRF}^{(l_{i-1},l_i)}\right)\right)}{\sum_{y'}\exp\left(\sum_{i}\left(w_{CRF}^{l'_i}h_i+b_{CRF}^{(l'_{i-1},l'_i)}\right)\right)} \quad (11)$$

其中 $w_{CRF}^{l_i}$ 是针对 l_i 的模型参数，$b_{CRF}^{(l_{i-1},l_i)}$ 是 l_{i-1} 到 l_i 的偏差。

3 中医药古文献命名实体识别实验过程及结果分析

3.1 研究流程

基于 Lattice LSTM 的中医药古文献命名实体识别研究可分为三个阶段：数据预处理阶段，对《伤寒论》文本数据进行规范化和实体标记，构建中医药主题词典，结合词典进行分词及去停用词处理；字词向量训练阶段，使用 python 的 gensim 包对文本进行 Word2vec 训练；命名实体识别阶段，通过实验评估模型性能。

3.2 实验环境及超参数设置

本实验部署于 Windows 系统，选取深度学习 Pytorch 框架开发，实验环境及模型的最优超参数设置见表 1、表 2。

表 1 实验环境

环境	配置
GPU	NVIDIA GeForce GTX 1660 Ti (6G)
内存	8G
操作系统	Windows 10 64 位
Python	3.8
Pytorch	1.9.1

表 2 模型超参数

参数名称	参数含义	参数值
char/word_emb_dim	字（词）向量维度	300
hidden_dim	隐含层维度	200
learning_rate	学习率	0.015
lr_decay	学习率衰减	0.05
dropout	丢弃率	0.5

3.3 数据预处理

本研究的数据集采用明代赵开美翻刻的宋本《伤寒论》[11]，原作者为东汉医圣张仲景，全书原 12 卷，现存 10 卷 22 篇，共 398 条，条文共包含 464 句。

用 python 读取《伤寒论》全文，去除多余符号后，根据 BIOES 规则，使用基于 Linux 系统的 brat[16] 软件标记文本中的五类实体：疾病（disease）、证候（syndrome）、方剂（prescription）、症状（symptom）、药材（medicine），

其中单字符实体标注为 S，多字符实体首字标记为 B，中部标记为 I，末尾标记为 E，非实体标记为 O。如条文第 224 条可标注为：“阳 /B-dis 明 /I-dis 病 /E-dis，/O 汗 /B-sym 出 /I-sym 多 /E-sym 而 /O 渴 /S-sym 者 /O，/O 不 /O 可 /O 与 /O 猪 /B-pre 苓 /I-pre 汤 /E-pre。/O”，标注示例见图 2。

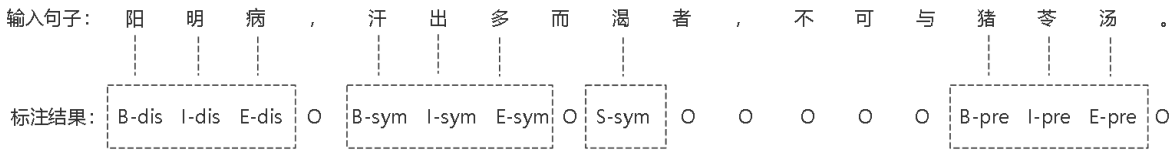

图 2　数据标注示例

以句为单位，将处理后的《伤寒论》数据按照 60%、20%、20% 的占比随机分配得到实验所用的训练集、验证集和测试集，具体情况统计见表 3。

表 3　实体类别统计

实体类别	实体数量		
	训练集	验证集	测试集
疾病	169	77	75
证候	171	57	52
症状	1259	532	416
方剂	662	214	210
药材	642	196	262

为进一步提高术语识别的准确性，本研究基于《中医病证分类与代码》（GB/T 15657—1995）、《中药方剂编码规则及编码》（GB/T 31773—2015）[17] 等国家标准文件，同时以中草药专业知识服务系统（http://zcy.ckcest.cn/tcm/）提供的数据资源为补充，构建中医药领域实体词典，共包含 5960 个实体，涉及疾病、方剂、证候、药材、症状等类别。利用词典对中医药古籍文本进行分词及去停用词处理。

3.4　实验结果分析

将 Lattice LSTM 与 NER 任务的多个经典模型进行对比，实验结果见表 4。可观察到如下特点：首先，深度学习的各模型整体表现出更好的性能；其次，BiLSTM 加入 CRF 层后各评价指标有了明显提升，这是因为 CRF 可以自动学习标记间的约束，进而修正模型的输出，保证了预测结果的合理性。再者，根据 F1 值对各模型性能进行排序：Lattice LSTM>BiLSTM-CRF>CRF>BiLSTM>HMM，可知 Lattice LSTM 性能最优，比 BiLSTM-CRF 总体提升了 1.68%，这是因为该模型在基于字的模型基础上实现了潜在词汇信息的充分利用，解决了语义缺失的问题，避免了分词错误，极大提升了识别效果。

表 4　模型效果对比

模型	P	R	F1
HMM	90.39%	89.36%	89.66%
CRF	93.89%	94.11%	93.86%
BiLSTM	93.66%	93.64%	93.59%
BiLSTM-CRF	94.18%	94.18%	94.08%
Lattice LSTM	95.69%	95.85%	95.66%

探究模型在不同类别实体识别上的表现，各类实体数量及 F1 值见图 3。可观察到如下特

doi:10.3772/j.issn.2095-915x.2023.05.009

点：首先，深度学习模型对证候的识别效果比浅层机器学习模型差，主要是因为可供模型学习的实体数过少，导致其优势无法体现；其次，各模型对方剂和药材的识别效果较好，主要是因为二者的数据特征较强，在文本中方剂均以“方”字结尾，而药材的名称全文统一、易于识别。疾病和症状的识别情况却不尽相同：疾病实体虽然数量少，但因其名称较为统一，所以识别效果很好；而症状实体虽然数量最多，但因其组成复杂、形式多样，所以识别效果较差。另外，Lattice LSTM 模型在除证候外的其他实体上识别效果均最佳，可用于中医药古文献术语识别。

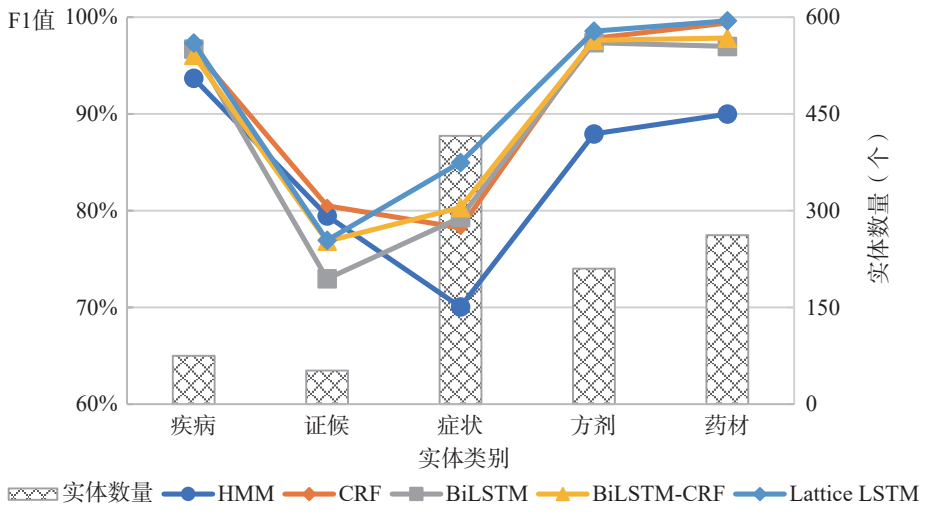

图 3　各类实体的数量及 F1 值

3.5　参数敏感性分析

字（词）向量维度（char/word_emb_dim）、隐含层维度（hidden_dim）和丢弃率（dropout）是 Lattice LSTM 模型的重要参数，评估参数变化对模型性能的影响，有利于发掘最佳参数组合，使模型表现出最优效果，结果见图 4。

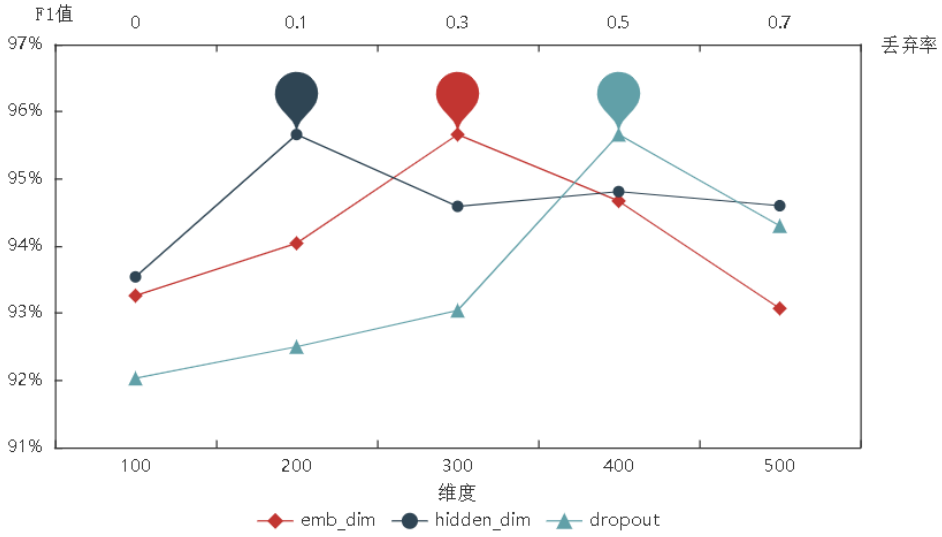

图 4　各参数对模型性能的影响

当不改变其他参数值，字（词）向量维度为 300 时模型的表现最佳；控制其他参数不变，加入 dropout 后模型的性能得到了较大提升，这是因为该技术在训练中将神经网络单元以某概率值暂时丢弃，可以防止模型出现过拟合的情况；当丢弃率值为 0.5 时，模型在实验中获得了最佳的效果，此后随着丢弃率值的增大，模型性能呈下降趋势。控制其他参数不变，改变 hidden_dim 后模型的性能受到较大影响；当 hidden_dim 值为 200 时，模型性能最优，此后随着 hidden_dim 值的增大，模型性能总体呈下降趋势。综上可知，本文的 Lattice LSTM 模型应选用 char/word_emb_dim 值为 300，hidden_dim 值为 200，dropout 值为 0.5 的参数组合。

4 中医药领域知识图谱及案例分析

4.1 模式构建

为检验本文提出模型的现实应用效果，本文在实验抽取出的实体基础上，构建了中医药知识图谱。知识图谱可以将零散、非结构化的知识以结构化形式存储，实现了知识的有效连接，提高了知识的管理和利用效率。构建知识图谱的核心环节是命名实体识别和实体关系抽取 [18]，首先是模式构建 [19]。本文的中医药知识图谱模式见图 5。

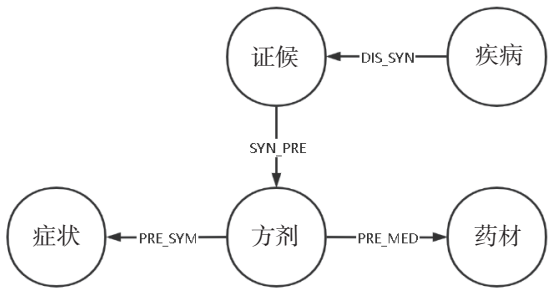

图 5 中医药知识图谱模式图

4.2 实体关系抽取

实体关系抽取即提取不同实体间存在的语义联系。本研究涉及疾病（disease）、证候（syndrome）、方剂（prescription）、症状（symptom）、药材（medicine）五类实体，定义实体间存在如下四类关系：疾病 _ 证候（DIS_SYN）、证候 _ 方剂（SYN_PRE）、方剂 _ 症状（PRE_SYM）、方剂 _ 药材（PRE_MED），对《伤寒论》文本进行人工关系抽取。如原文第 31 条的“项背强、无汗、恶风，葛根汤主之，葛根、麻黄、桂枝……”一句，“项背强”“无汗”“恶风”与“葛根汤”构成 PRE_SYM 关系，“葛根汤”与“葛根”“麻黄”“桂枝”等构成 PRE_MED 关系。

4.3 知识图谱构建

本研究使用 Neo4j[20] 构建中医药知识图谱。在 Neo4j 中，知识主要以节点和边的形式存储，节点代表意义特殊的实体，边代表实体间的关系。将《伤寒论》的实体及关系文件通过 LOAD CSV 语句导入 Neo4j 图数据库，生成 300 个节点和 825 条边，具体情况统计见表 5。

表 5 中医药知识图谱的节点及边统计

节点类型	节点数量	边类型	边数量
disease	8	DIS_SYN	39
syndrome	39	SYN_PRE	123
prescription	105	PRE_SYM	221
symptom	91	PRE_MED	442
medicine	57		

4.4 案例分析

“大疫出良方”，面对 2019 年底席卷全球、来势汹汹的新冠肺炎疫情，中医药通过临床试

doi:10.3772/j.issn.2095-915x.2023.05.009

验筛选出对疾病治疗起关键作用的“三药三方”，其中“清肺排毒汤”作为疫情期间临床首选的治疗药物在多省患者的实际治疗中有效率超过了 90%。该汤剂由《伤寒论》的麻杏石甘汤（即麻黄杏仁甘草石膏汤）、小柴胡汤、五苓散等经典名方组成，包含麻黄、杏仁、桂枝等 21 味中药，主要功效是宣肺透邪、清热化湿、健脾化饮，能够改善患者出现的发热、咳嗽、乏力、咽痛等症状。

在中医药知识图谱中，检索清肺排毒汤三大方剂的相关信息，结果见图 6。节点中橙色代指疾病，蓝色代指证候，绿色代指症状，红色代指方剂，浅褐色代指药材。可以看到，清肺排毒汤在《伤寒论》中涉及 4 种疾病、6 种证候、11 种症状和 14 味药材，如小柴胡汤包含柴胡、半夏、人参等药材，与少阳病、阴阳易差后劳复病有关，可用于治疗咽干、呕吐、不欲饮食等症状。此类查询展现了各中医实体间的关系，帮助医师迅速找到所需信息，为中医药智能问答及推荐奠定了基础。

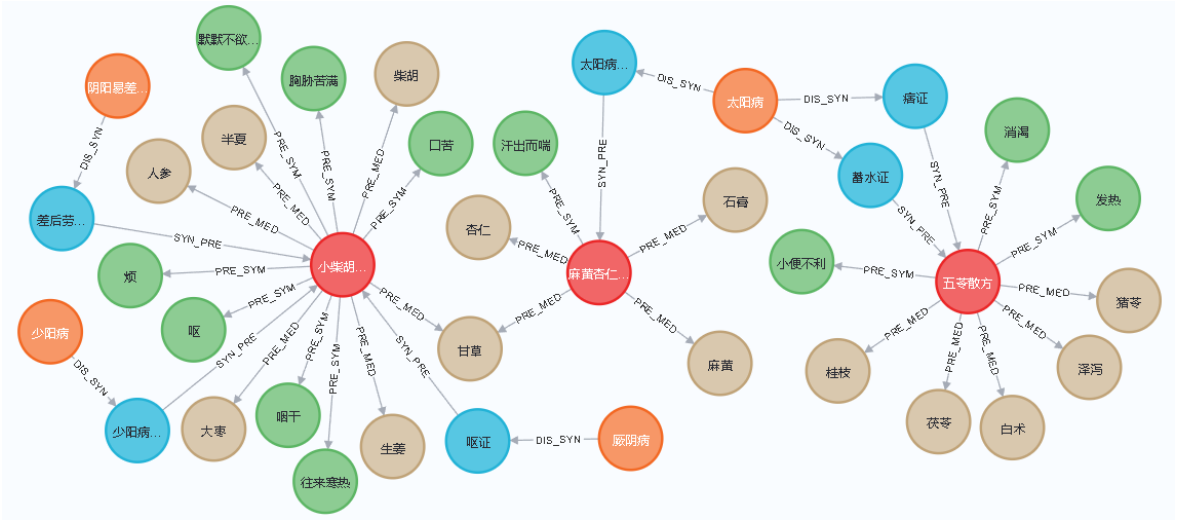

图 6　清肺排毒汤的相关检索结果

新冠肺炎的常见症状有：发热、呼吸不畅（对应《伤寒论》的“胸满”“胸中窒”）、咳、腹泻（对应《伤寒论》的“下利”）、头痛、头晕（对应《伤寒论》的“头眩”）、恶心（对应《伤寒论》的“呕”），在知识图谱中查询各症状对应的治疗方剂，结果见图 7。节点中绿色代指症状，红色代指方剂。可以看到，对于发热患者，可以使用调胃承气汤、大青龙汤、真武汤等治疗；对于发热和头痛兼具者，可以使用麻黄汤、理中丸、桂枝汤等治疗。此类查询为中医药医疗诊断提供了清晰的药方参考，有利于对症下药，辅助医师进行临床决策。

5　总结与展望

5.1　研究总结

本文创造性将 Lattice LSTM 模型应用于中医药古文献的命名实体识别任务中，以《伤寒论》为研究对象，对疾病、证候、方剂、症状和药材五类实体进行识别。研究发现，实体数量及构成的复杂程度均会影响模型的识别效果；实验结果表明：Lattice LSTM 性能最优，F1 值

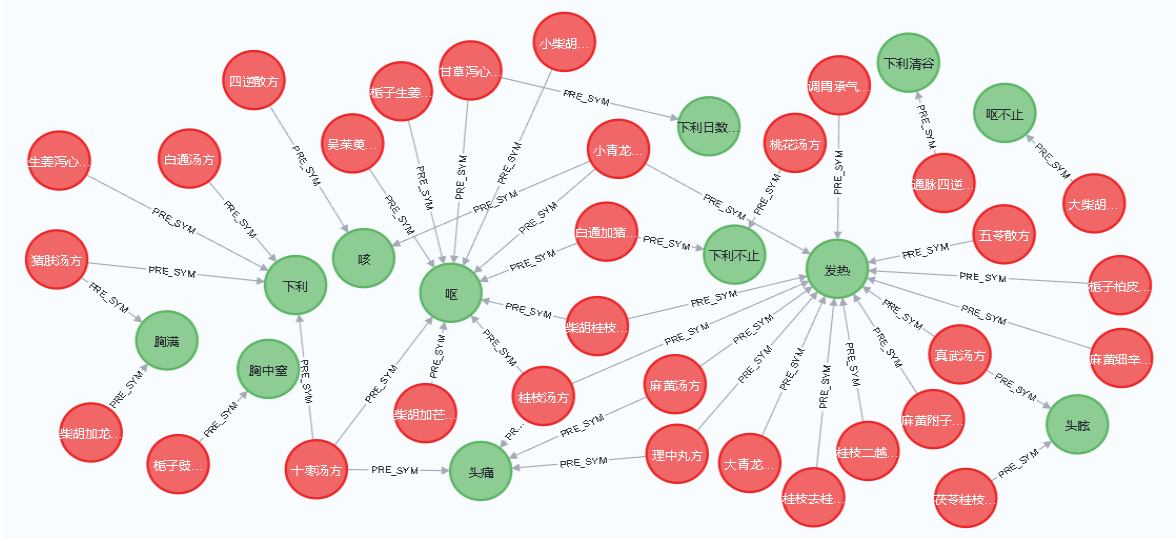

图 7　新冠肺炎症状及方剂的检索结果

达 95.66%，比主流模型 BiLSTM-CRF 提升了 1.68%，可用于中医药古文献的命名实体识别。此外，本文在提取出的实体基础上，人工抽取了实体间的四类关系，利用 Neo4j 搭建了中医药知识图谱；以新冠肺炎的中医诊疗方案为例，使用图谱进行了相关检索，证实了理论模型在实际应用中的效果。

5.2　未来展望

本文仍存在不完善和可改进之处，在今后的研究中将继续优化：

（1）用于中医古籍专业术语识别的语料库较小。本文仅使用《伤寒论》文本进行实证研究，数据量较小，难以充分发挥 Lattice LSTM 模型的优势。未来可以进一步扩充中医古文献的语料库，探究模型在大规模数据集上的表现；也可以进一步丰富图谱内容，如添加实体的属性描述（如各方剂中药材的配比），完善实体及关系类别。

（2）换用古汉语分词专用工具包。本文在 jieba 分词时使用了自定义的中医药领域主题词典，该工具一般用于现代汉语的分词，在古文分词方法的调研中发现了专注于古汉语分词的 NLP 工具包 jiayan[21]（https://github.com/jiaeyan/Jiayan），未来可将 jiayan 与 jieba 对比，检验其能否有效改善中医药古籍的分词质量。

（3）实体关系抽取和数据审核不够智能化。本文在实体关系抽取时主要采取人工匹配，没有使用自然语言处理技术，耗费了一定时间和人力，未来可以考虑引入关系抽取模型提高抽取效率；同时在知识图谱的数据审核方面也主要采用人工核对，工作量较大，未来可以通过属性矫正、实体对齐等减少相应成本。

参 考 文 献

[1] LEE L H, LU Y. Multiple Embeddings Enhanced Multi-Graph Neural Networks for Chinese Healthcare Named Entity Recognition[J]. IEEE Journal of Biomedical and Health Informatics, 2021, 25(7): 2801-2810.

[2] GLIGIC L, KORMILITZIN A, GOLDBERG P, et al. Named entity recognition in electronic health

doi:10.3772/j.issn.2095-915x.2023.05.009

records using transfer learning bootstrapped Neural Networks[J]. Neural Networks, 2020(121): 132-139.

[3] MULYAR A, UZUNER O, MCINNES B. MT-clinical BERT: scaling clinical information extraction with multitask learning[J]. Journal of the American Medical Informatics Association, 2021, 28(10): 2108-2115.

[4] KANG T, PEROTTE A, TANG Y, et al. UMLS-based data augmentation for natural language processing of clinical research literature[J]. Journal of the American Medical Informatics Association, 2021, 28(4): 812-823.

[5] KORMILITZIN A, VACI N, LIU Q, et al. Med7: a transferable clinical natural language processing model for electronic health records[J]. Artificial Intelligence in Medicine, 2021, 118: 102086.

[6] WEBER L, SÄNGER M, MÜNCHMEYER J, et al. HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition[J]. Bioinformatics, 2021, 37(17): 2792-2794.

[7] MA L L, YANG J, AN B, et al. Medical Named Entity Recognition Using Weakly Supervised Learning[J]. Cognitive Computation, 2022, 14(3): 1068-1079.

[8] 陈美杉 , 夏晨曦 . 肝癌患者在线提问的命名实体识别研究：一种基于迁移学习的方法 [J]. 数据分析与知识发现 , 2019, 3(12): 61-69.

[9] 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(7): 2504-2510.

[10] 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.

[11] 屈倩倩 . 基于自然语言处理的《伤寒论》研究 [D]. 合肥：安徽中医药大学 , 2021.

[12] 卢克治 . 基于中医古籍的知识图谱构建与应用 [D]. 北京：北京交通大学 , 2020.

[13] ZHANG Y, YANG J. Chinese NER using lattice LSTM[J]. arXiv preprint arXiv: 1805. 02023, 2018.

[14] 崔丹丹 , 刘秀磊 , 陈若愚 , 等 . 基于 Lattice LSTM 的古汉语命名实体识别 [J]. 计算机科学 , 2020, 47(S2): 18-22.

[15] 张笑天 . 基于 Lattice LSTM 的医学文本中文命名实体识别研究与实现 [D]. 成都：电子科技大学 , 2019.

[16] 许乾坤 , 刘耀 . 科技政策隐性扩散路径自组织研究 [J]. 情报资料工作 , 2022, 43(1): 61-70.

[17] 马捷 , 吴文玲 , 孙恒宇 , 等 . 中医诊疗知识库术语规范化研究 [J]. 图书情报工作 , 2021, 65(2): 17-26.

[18] 杨波 , 廖怡茗 . 面向企业动态风险的知识图谱构建与应用研究 [J]. 现代情报 , 2021, 41(3): 110-120.

[19] 李叶叶 , 李贺 , 沈旺 , 等 . 基于多源异构数据挖掘的在线评论知识图谱构建 [J]. 情报科学 , 2022, 40(2): 65-73, 98.

[20] 李旭晖 , 曾逸权 , 刘洋 . 一种面向语义的多粒度时间数据建模方法 [J]. 情报科学 , 2020, 38(6): 116-125.

[21] 邢付贵 , 朱廷劭 . 基于大规模语料库的古文词典构建及分词技术研究 [J]. 中文信息学报 , 2021, 35(7): 41-46.