基于Bert的中医方剂文本命名实体识别

徐丽娜,李 燕,钟昕妤,陈月月,帅亚琦

(甘肃中医药大学信息工程学院,甘肃 兰州 730000)

摘要:针对中医药领域常用命名实体识别模型存在的边界模糊和歧义性等问题,本文提出基于大规模预处理中文语言模型(Bert)的中医方剂文本命名实体识别方法。通过Bert预训练模型接受其相对应的词向量,将预处理完成的词向量输入到长短期记忆(Bi-LSTM)模块中,完成对文本上下文语义信息的捕获,最后使用条件随机场(CRF)模块解码输出得到的预测标签排序,依次检索和排序各类中医方剂文本实体,从而完成整个实体识别步骤,结果显示出Bert对中医方剂各类实体识别具有较高的适用性,中医方剂各类实体识别的准确率得到显著提升。

关键词:深度学习;中医方剂;命名实体识别模型

中图分类号:TP391.1;R289 **文献标识码**:A **DOI**:10.3969/j.issn.1006-1959.2023.04.006

文章编号:1006-1959(2023)04-0032-06

Named Entity Recognition of Traditional Chinese Medicine Prescription Based on Bert

XU Li-na,LI Yan,ZHONG Xin-yu,CHEN Yue-yue,SHUAI Ya-qi

(Information Engineering Institution of Gansu University of Chinese Medicine,Lanzhou 730000,Gansu,China)

Abstract:Aiming at the boundary ambiguity and ambiguity of named entity recognition models commonly used in the field of traditional Chinese medicine, a named entity recognition method of TCM prescription text based on large-scale preprocessed Chinese language model (Bert) is proposed. The corresponding word vector is accepted by the Bert pre-training model, and the preprocessed word vector is input into the long-term and short-term memory (Bi-LSTM) module to capture the semantic information of the text context. Finally, the conditional random field (CRF) module is used to sort the predictive tags obtained from the output, and the text entities of all kinds of TCM prescriptions are retrieved and sorted in turn, so as to complete the whole entity recognition step. The results show that Bert has high applicability to all kinds of entity recognition of traditional Chinese medicine prescription, and the accuracy of entity recognition of traditional Chinese medicine prescription has been significantly improved.

Key words:Deep learning;Traditional Chinese medicine prescriptions;Named entity recognition models

中医方剂(traditional Chinese medicine prescriptions)是千百年来名医大家临床实践总结得出的成果,是集中医、方法、方剂、中药四大理论为一体的综合体系,不仅可以运用中医的主要手段即辩证论治的理论用于指导临床防治疾病,也是药性理论的具体表现[1]。因中医方剂有其独到的组合方式,蕴含着丰富的信息,但事实上这些复杂的方剂变换中就已包含了大量规律,故提取中医药信息便成为了促进中医药信息化结构化的重要手段方法[2]。中医药文本数据挖掘从最初的在海量历史数据中进行简单查询,发展成为不仅从大量数据中获取知识,还能够揭示事物的内在发展规律,预测事物发展趋势的潜在关联或模型的一项技术[3]。利用数据挖掘可以分析和讨论中医治疗的疗效和基本理论、中医辨证论治的各种方法以及方剂的分类理论,从而了解中医的用药规律[4]。命名实体识别作为文本数据挖掘中重要的一环,其主要任务为在非结构化数据中提取出一组相关性较高的名词。目前文献研究集中于中医医案及电子病历,或是单一研究药品命名实体识别,对中医方剂文本领域关注度较少,尤其在中医方剂领域中基于种类多样且复杂的实体类型,使得此方法在实际应用中进展较慢。而将命名实体识别技术运用在中药学中,从海量的中医方剂文本数据中识别出不同类型的且较为准确的实体信息将是中医药学者未来的重点研究方向。本文主要基于Bert的中医方剂文本命名实体识别的数据获取与标注、实验模型与框架、实验结果进行分析。

1 数据获取与标注

1.1 数据获取 从以PDF格式存放的《中医方剂大词典》[5]中获取数据,整理分散在各个文献中的所有方剂,包括各种适应证、试验用例及实验研究资料等。通过文本识别技术,将《中医方剂大词典》进行文本识别,并转换成TXT等命名实体识别模型可输入的格式。

作者简介:徐丽娜(1996.8-),女,甘肃定西人,硕士研究生,主要从事中医药数据挖掘研究

通讯作者:李燕(1976.5-),女,甘肃兰州人,硕士,副教授,硕士生导师,主要从事中医药数据挖掘研究

1.2 数据标注 从获取到的数据中进行筛选，基于内容完整的方剂数据，选择在主治中明确提到“心悸”“心痛”“怔忡”等关键词的方剂，删除主治功效中虽含有关键字，但不是主要症状的方剂，最后得到治疗心血管疾病的方剂 567 首。并遵照中医药学语言系统（TCMLS）中的标准，选择方剂（prescription）、药物（medicine）、疾病（disease）、功效（efficacy）、炮制方法（processing method）五类实体进行研究，命名实体定义及标注规则见表 1。

中医方剂数据语料标注采用方法为“BIO+命名实体”的序列标注方法，序列标注是指对序列中的每个字符分配一个特定的标签，当给定一个序列时，对序列中的每个元素进行标注。通常一个序列即为一个句子，而一个元素即为句子中的一个单词。序列标注中最常用的方法是 BIO 标注法，B 标记为实体的开始，I 标记为实体的中间部分，O 标记为非实体字符的部分[6]。命名实体标注信息见表 2。

为了保持标注的一致性，由一人单独完成标注任务，根据选取好的语料对句子进行切分，使用 BIO 标注方法对筛选完成的 567 首方剂数据进行人工标注，并将现有数据划分为 70%的训练集和 30%的测试集，随后即可进行后续研究，中医方剂数据标注示例见表 3。

表 1 命名实体定义表

实体类别	实体类别定义	示例
方剂	方剂学名或通用名称	稳心 1 号
药物	方剂所包含中草药名称	生地，熟地
疾病	方剂主治的疾病名称	冠心病心绞痛
功效	方剂在防治各类疾病方面的作用	养阴疏肝，理气通络
炮制方法	中草药原料制成方剂方法过程	水煎

表 2 命名实体标注信息表

实体类别	头实体标注	中间实体标注	尾实体标注
方剂	B-Prescription	I-Prescription	O-Prescription
药物	B-Medicine	I-Medicine	O-Medicine
疾病	B-Disease	I-Disease	O-Disease
功效	B-Efficacy	I-Efficacy	O-Efficacy
炮制方法	B-Processing method	I-Processing method	O-Processing method
非实体	O	O	O

表 3 中医方剂标注示例

原句	截水丸【组成】缩砂
序列标注	B-PrescriptionI-PrescriptionO-Prescription O O O O B-MedicineI-Medicine
原句	仁 30 g，蓬术 2
序列标注	I-MedicineI-MedicineI-MedicineO-Medicine O B-Medicine I-Medicine I-Medicine
原句	3 g，汉椒 15 g，
序列标注	I-Medicine O-Medicine O B-Medicine I-Medicine I-Medicine I-Medicine O-Medicine O
原句	桂 15 g，苍术 1
序列标注	B-Medicine I-Medicine I-Medicine O-Medicine O B-Medicine I-Medicine I-Medicine
原句	5 g，茱萸 15 g
序列标注	I-Medicine O-Medicine O B-Medicine I-Medicine I-Medicine I-Medicine O-Medicine

2 实验模型与框架

近年来随着中医文本数据量的增多，对于从大量数据中挖掘有用知识的技术也在快速发展，命名实体识别任务作为重要的一环，模型的质量将影响整个识别任务的准确性。预训练模型因其可获得高质量的词向量被广泛关注，将预训练模型运用到中医方剂文本识别任务中，从一定程度上可以减轻临床人员录入数据的负担，还能够有效提高识别准确率，为挖掘更多中医方剂的知识奠定基础，更能促进中医药相关事业的发展及传承[7]。

本文模型框架由 Bert 模块、BiLSTM 模块和 CRF 模块组合构建而成，模型结构见图 1。Bert-BiLSTM-CRF 模型命名实体识别的主要步骤如下：首先，将已标注完成的中医方剂文本标记词，如“心血康饮”通过 Bert 预训练模型并接受其相对应的词向量，然后将预处理完成的词向量输入到 BiLSTM 模块中，更好地完成对文本上下文语义信息的捕获，最后使用 CRF 模块解码 BiLSTM 模块中输出得到的预测标签排序，然后依次检索和排序各类中医方剂文本实体，从而完成整个实体识别步骤。

2.1 Bert 模块 Bert 预处理模型源于谷歌的一个开源项目 word2vec[8]。本文的 Bert 预训练模型使用 12 层的 Transformer 架构，由编码器与解码器组成，它们使用分层堆叠来实现自然语言处理中的最佳特征提取器，在训练时，只有 Transformer 的编码器被使用，且每层编码器由自注意力机制与前馈神经网络构成，通过从输入的中医方剂长文本信息中提取特征，最终实现文本表示。

2.2 BiLSTM 模块 BiLSTM 模块包括前向长短时记忆神经网络（LSTM）和后向长短时记忆神经网络（LSTM）[9]，本文所使用的 LSTM 结构首先通过遗忘和记忆状态门中的新信息，因为其可以传递对后续计算有价值的信息，舍弃无价值信息，再在每一步输出隐藏层的状态，其中遗忘、记忆和输出由遗忘门、记忆门和通过前一时刻隐藏层的状态和当前输入计算的输出控制[10]，基本单元结构见图 2。

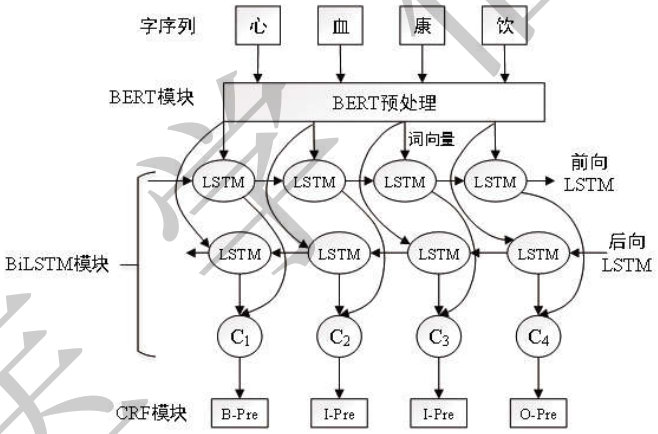

图 1 Bert-BiLSTM-CRF 模型框架图

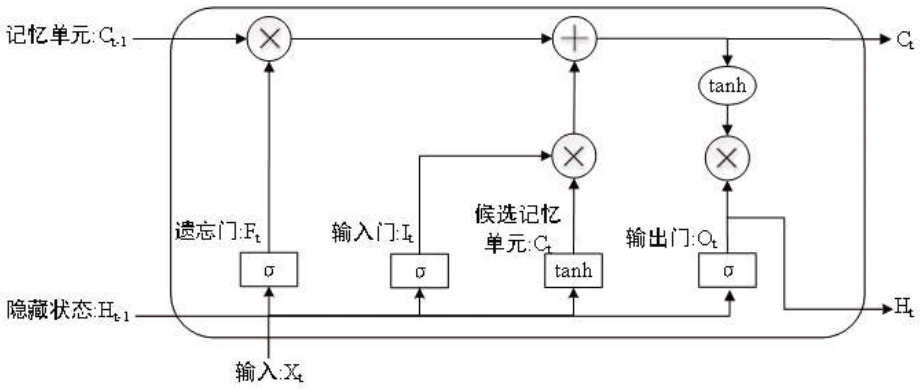

图 2 LSTM 基本单元结构图

双向长短时记忆神经网络模型(BiLSTM)可克服传统循环神经网络的梯度消失问题,允许网络选择性地保留以前的信息,且双向传输模型在传播过程中利用时间来处理前后文的信息,最终输出结果也为事前向传播和后向传播相结合而成的[11]。在传统中医方剂文本中,很多情况是多个词构成一个实体,即长序列实体,利用双向长短期记忆神经网络,可以快速获取长距离所属特征,有效优化模型对长序列实体的识别性能。

2.3 CRF 模块 虽然 BiLSTM 模块已经学习了输入序列的上下文信息,但其默认序列标签的状态间是相互独立不影响的,这将会出现实体标签不一致问题,所以关键步骤仍然是在获取标签数据集后处理标签之间的关系[12]。本实验解决方案是添加 CRF 模块,这是一个条件概率模型,该模型会考虑标签之间的顺序,可以处理标签之间的相互约束,有效解决标签顺序不一致的问题,从而提高中医方剂文本模型实体识别的准确率。

3 实验结果

3.1 实验设计 本实验是在 Windows 系统下使用 Python3.9 为主要编程语言,在 Pytorch1.6.0、Pytorch-crf2.10.0 以及 Transformers0.7.2 的实验环境下进行。其中,将实验分为 2 个环节,第 1 个环节使其他条件不变,分别训练 BiLSTM-CRF 和 Bert-BiLSTM-CRF 这 2 种模型,来对比 2 种模型的性能;第 2 个环节使其他条件不变,在模型训练过程中主要改变模型学习率的大小,并最终主要通过 F 值来评判模型训练效果。

3.2 模型性能 采用目前常用来衡量命名实体识别模型性能的 3 个评分指标进行分析,分别为准确率(P)、召回率(R)以及 F 值[13],指标的数学定义如公式(1)~(3)所示;

$$P=\frac{\alpha}{\beta}\times 100\% \quad (1)$$

$$R=\frac{\alpha}{A}\times 100\% \quad (2)$$

$$F=\frac{2PR}{P+R}\times 100\% \quad (3)$$

上述公式中,α 是识别正确的实体数目,A 是实体总数目,B 是被识别出的实体数目。P 是指在所有预测结果中与实际结果一致的实体数目占总实体数目的百分比,R 是指被正确识别的实体数目占总实体数目的百分比,F 值是准确率与召回率的综合值,用来对模型进行总体评估[14]。

在实验过程中,因 Bert 模型要求样本长度低于 512,最初将句子切分长度设置为 256,但参数设置太低可使模型识别率不理想,且分析数据发现文本长度大多集中在 200~600,因此将最大句子长度参数更换为 512,发现训练结果得到明显提升;对于 Bert 层学习率,由于预训练模型的参数在训练过程中已经更新若干次,所以在调整参数过程中为了保持模型的高泛化率,将预训练模型 Bert 层的参数设置为(2e-5),并在训练过程中将 CRF 层学习率从(2e-3)调整为(1e-3),取得了较好的效果;Batch-Size 选择设置为 16,Dropout 设置一般取值 0.5,Epoch 设置为 30,通过实验结果证明获得了较优的模型识别效果,见表 4、图 3。

表 4 模型总评对比表(%)

模型	R	P	F
BiLSTM-CRF	83.7	78.5	79.0
Bert-BiLSTM-CRF	88.5	82.1	85.9

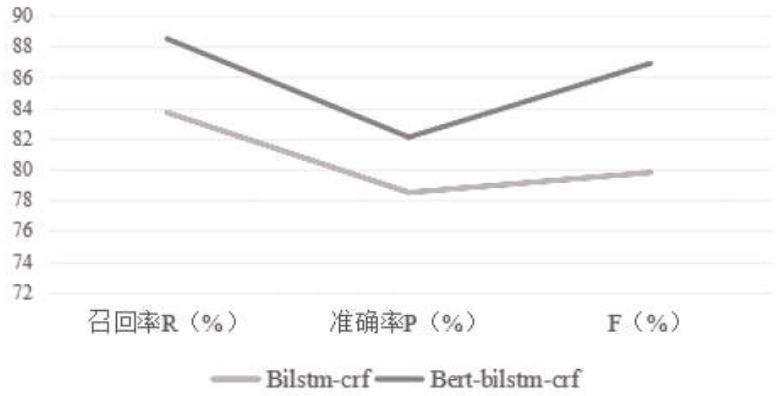

图 3 模型总评对比图

通过图 3 模型总评对比图可以看出，相比 BiLSTM-CRF 模型，加入 Bert 后，模型命名实体识别准确率整体升高 6%左右，性能明显增强。使模型性能增强的方式较多，如调整模型参数或改进算法等。本实验通过调整模型参数，不断进行模型训练及测试，将 Bert 层训练级别设置为较低，CRF 层训练级别设置较高，以此获得了更好地框架识别效果。并且本身 BiLSTM-CRF 模型由于 CRF 模块的加入，利用 CRF 模块的特性得到全局最优标签序列，在此基础上再次引入 Bert 模型进行预处理，充分提取字符级、词级和句子间的关联关系，使得预训练时得到的词向量能够更好地表达所需要的语义信息，从而提升模型的命名实体识别性能。Bert-BiLSTM-CRF 模型与 BiLSTM-CRF 模型对方剂、药物、疾病、功效和炮制方法五类实体识别的 F 值对比统计图见图 4。

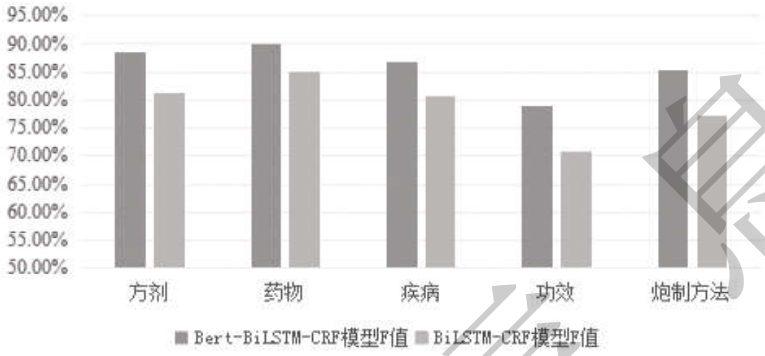

图 4 模型实体识别 F 值对比图

从图 4 模型实体识别 F 值对比图可以看出，Bert-BiLSTM-CRF 模型相较于 BiLSTM-CRF 模型，对于 5 类实体整体识别准确率都有所增长，其中药物、方剂、疾病、炮制方法的识别准确率较高，对于功效识别准确率较低，原因可能是本实验心血管疾病中医方剂数据来源于《中医方剂大词典》及相关数据库补充资源，这些数据中对于一首方剂的描述主要包括其方剂名称、药物组成、主治何种疾病以及如何炮制使用，在文本中数量较多且方剂名称大部分以丸、汤、方结尾，实体特征较明显，所以识别率较高；而方剂数据集中对于功效的描述较为稀缺，在文本数量中占据较少，且对于功效术语的规范化存在不足，容易出现用词繁复的情况，如“发汗解肌”“发汗解表”都可表述发散风寒药的功效，这种同近义词使得模型实体识别难度较大，从而导致模型识别率降低。由此可见，实体特征、实体数据量大小以及数据规范性对于模型的识别精度都存在一定程度的影响，未来可以通过扩展中医方剂文本数据量，规范化使用中医术语来提高模型整体识别精确率以及模型有效性。

4 讨论

中医药丰富的治疗手段以及灵活的方法作为中医药传统特色优势之一，在提高人们健康及生活质量方面都做出了卓越贡献。而中医方剂是中医最常用的药物治疗手段之一，通过与现代技术相结合来适应当代社会的发展需求，可使用计算机处理及分析方剂文本数据，但其需将医学数据做结构化处理，而命名实体识别技术则是结构化表示的基础。

国内对医学命名实体识别的研究最早认为是分类问题，经典的隐马尔可夫以及条件随机场等模型被广泛研究[15,16]。随着各个领域的快速发展，深度学习方法的优势逐渐显现出来，通过自动提取特征标签，在命名实体识别领域得到了较好的效果[17]。在国际领域，基于循环神经网络（RNN）与长短期记忆网络-条件随机场（LSTM-CRF）模型都具有较高的代表性。并且随着深度学习领域的发展，自注意力机制（Attention）和 Transformer 模型等在自然语言处理领域的优异表现，不断有更多优秀的改进模型在各个领域被提出[18,19]。

Bert 预训练模型是使用大规模未标注数据集训练 Bert 来提取文本特征，再使用 BiLSTM-CRF 模型进行标注。即使 Bert 在各个领域已得到广泛应用，但在中医药方剂领域的研究仍然匮乏，针对于此，本实验提出基于深度学习的 Bert 预训练模型，结合

BiLSTM-CRF 模型运用于中医方剂的命名实体识别任务中，对模型进行改进优化，并与 BiLSTM-CRF 模型进行对比，发现将 Bert 预训练框架与 BiLSTM-CRF 模型进行结合，运用到中医方剂文本识别任务中，中医方剂各类实体识别的准确率得到提升，充分发挥了模型在构建字符向量时充分考虑到字符间的关联关系，解决以往模型常见的边界模糊、歧义性等问题，且得益于 Bert-BiLSTM-CRF 模型设计与优化训练，模型识别性能也得到了有效提升。

总之，命名实体识别作为自然语言处理技术中最重要的一环，在各个领域已有卓越表现，但在中医领域尚在起步阶段，仍然是未来的研究重点，并且一个性能较高的中医方剂文本命名实体识别模型能够为中医智能化、信息化的发展奠定良好基础，也能够构建更精准的中医药知识网络、中医方剂知识图谱，为实现临床辅助决策、智慧医疗、中医方剂知识推荐等智能服务提供支撑与动力。本研究不足之处在于对部分实体的识别准确率较低，原因可能是实体数据量较少、实体特征不明显等，从而对模型识别准确率造成影响。目前，仍有许多隐藏有意义的药物知识未被发现，在今后的研究中，将增加更多相关数据，或引入专业词典来提高模型的有效性。

参考文献：

[1]王丽娜,胡建鹏,范婧婧,等.方剂学理论体系形成与发展[J].中医药临床杂志,2017,29(12):2044-2047.

[2]李振岳.中药方剂数据挖掘研究[D].广州:广东药学院,2010.

[3]田瑾.基于复杂网络及关联规则的失眠用药中医临床数据挖掘研究[D].北京:北京中医药大学,2015.

[4]周荣荣,闫润红,王丽萍,等.数据挖掘技术在中医方剂科学问题研究中的应用[J].中华中医药杂志,2018,33(9):4016-4020.

[5]杨璐妍,聂付敏,金维捷,等.基于数据挖掘《中医方剂大辞典》治疗乳痈用药规律探讨[J].云南中医中药杂志,2022,43(2):24-28.

[6]贾杨春,朱定局.基于深度学习的医疗命名实体识别[J].计算机系统应用,2022,31(9):70-81.

[7]黄敏婷,赵静,于涛.基于医学大数据的预训练语言模型及其医学文本分类研究 [J]. 中华医学图书情报杂志,2020,29(11):39-46.

[8]Prottasha NJ,Sami AA,Kowsher M,et al.Transfer Learning for Sentiment Analysis Using BERT Based Supervised Fine-Tuning [J].Sensors (Basel),2022,22(11):4157.

[9]Ronran C,Lee S.Effect of Character and Word Features in Bidirectional LSTM-CRF for NER [C]//2020 IEEE International Conference on Big Data and Smart Computing (BigComp).IEEE,2020.

[10]刘苏文,邵一帆,钱龙华.基于联合学习的生物医学因果关系抽取[J].中文信息学报,2020,34(4):60-68.

[11]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.

[12]张立.基于综合性临床诊断数据分析系统的医院导诊应用[J].医学信息学杂志,2020,41(12):68-72.

[13]都丽婷,夏晨曦,赵冬,等.基于条件随机域的临床文本去识别研究[J].中国卫生信息管理杂志,2017,14(2):217-222.

[14]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):40-43,48.

[15]陈锦,常致全,许军.基于 HMM 的生物医学命名实体的识别与分类[J].计算机时代,2006(10):40-42.

[16]李彦鹏,杨志豪,林鸿飞.基于条件随机域的生物医学命名实体识别[C]//第三届学生计算语言学研讨会论文集,2006.

[17]王浩畅,赵铁军,刘延力,等.生物医学文本中命名实体识别的智能化方法[C]//2006 年首届 ICT 大会信息、知识、智能及其转换理论第一次高峰论坛会议论文集,2006.

[18]Crichton G,Pyysalo S,Chiu B,et al.A neural network multi-task learning approach to biomedical named entity recognition [J].BMC Bioinformatics,2017,18(1):368.

[19]于尤婧.面向可解释性双向编码语言模型的文本分类研究[D].长春:吉林大学,2020.

收稿日期：2022-11-10；修回日期：2022-12-14

编辑/杜帆