融合汉字部首的 BERT – BiLSTM – CRF 中医医案命名实体识别模型*

刘 彬 肖晓霞 邹北骥 周 展 郑立瑞 谭建聪

（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** 研究提取中医医案中医疗术语的方法，实现医案自动结构化，为医案知识发现提供结构化数据。**方法/过程** 提出一种 BERT 结合长短期记忆人工神经网络、条件随机场和部首特征的深度学习命名实体识别模型，在 BERT 词向量中嵌入汉字部首，采用双向长短期记忆人工神经网络提取实体特征，使用条件随机场进行序列预测。将人工标注的 400 份共计 5 万余字的医案按照 3∶1 划分为训练集和测试集，使用该模型识别中医医案中的身体部位、药物、症状、疾病 4 类命名实体。**结果/结论** 该模型在测试集 *F*1 值为 84.81%，优于其他未嵌入部首的模型，表明该模型能够更有效地识别中医医案中的命名实体，更好地结构化医案。

〔关键词〕 实体识别；部首特征；BERT 模型；双向长短期记忆模型；条件随机场；自然语言处理

〔中图分类号〕R – 058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673 – 6036.2023.06.008

Study on Named Entity Recognition of Chinese Medical Records Based on BERT – BiLSTM – CRF with Chinese Radicals

LIU Bin, *XIAO Xiaoxia*, *ZOU Beiji*, *ZHOU Zhan*, *ZHENG Lirui*, *TAN Jiancong*

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

〔**Abstract**〕 **Purpose/Significance** To study the method of extracting medical terms from Chinese medical records, to realize the automatic structure of medical records, and to provide structured data for knowledge discovery of medical records. **Method/Process** The paper proposes a deep learning named entity recognition (NER) model based on BERT combining long short – term memory (LSTM), conditional random fields (CRF) and radical features. This model embeds Chinese radicals in BERT word vector, extracts entity features with BiLSTM, and uses CRF for sequence prediction. 400 medical cases with a total of more than 50 000 words manually marked are divided into training set and test set according to 3∶1, the model is used to identify four types of named entities in Chinese medical records, namely body, medicine, symptom, and disease. **Result/Conclusion** The *F*1 value of this model on the test set is 84.81%, which is superior to other models without embedded radicals, indicating that the model can more effectively identify named entities in Chinese medical records and better structured medical records.

〔**Keywords**〕 entity recognition; radical features; BERT; bi – directional long short – term memory (BiLSTM); conditional random fields (CRF); natural language processing (NLP)

〔修回日期〕 2022 – 12 – 19

〔作者简介〕 刘彬，硕士研究生；通信作者：肖晓霞，副教授，硕士生导师。

〔基金项目〕 国家重点研发计划中医药现代化研究重点专项（项目编号：2017YFC1703306）；科技创新 2030——“新一代人工智能”重大项目课题（项目编号：2018AAA0102102）；湖南省中医药管理局重点课题（项目编号：A2023048）。

1 引言

中医医案是中医名家临床过程记录，蕴含丰富诊疗经验、中医诊疗理论和学术创新。因此中医医案成为学习和传承中医药的重要文献，有效提取医案中记录的症状、体征、证、处方、剂量、治法等命名实体，可以将医案结构化后利用机器学习、深度学习等人工智能方法探索中医诊疗规律、构建智能诊断模型，对中医传承与发展具有重要意义。

命名实体识别[1]是指从特定文本中识别和提取具有特定含义的实体，是文本结构化的第 1 步，也是机器翻译、问答系统、句法分析等自然语言处理（natural language processing，NLP）任务的重要基础工具。相对于英文每个词都有空格分隔，中文词与词之间没有明显的分隔符，导致中文命名实体识别更加困难。提取医案中记录的实体信息可以视为一种中文命名实体识别任务，因此构建高准确率、实用的中医命名实体提取模型迫在眉睫。

从统计学习方法到神经网络，研究者使用不同方法提取医疗文本中的实体。肖晓霞等[2]提出一种基于自然语言处理的中医医案文本快速结构化方法，通过构建症状词典，采用结合词典的改进 N - gram 模型提取医案文本中的体征、症状等实体，并在结构化过程中不断对词典进行更新，模型 *F*1 值达到 82.99%。高佳奕等[3]提出一种条件随机场（conditional random field，CRF）与长短期记忆人工神经网络（long short - term memory，LSTM）混合模型，基于已标注的肺癌医案进行实验，*F*1 值达到 84%。许力等[4]提出一种双向长短期记忆人工神经网络（bi - directional long short - term memory，BiLSTM）模型与预训练语言模型双向编码器表征（bi-directional encoder representations from transformers，BERT）[5]相结合的模型，并将其应用于生物医学领域进行实体识别，该模型在 3 个数据集上的平均 *F*1 值达到 89.45%，取得良好效果。

中医医案是描述疾病诊疗的文本，症状、身体部位、药物、疾病等术语的汉字部首结构具有明显特点。汉字的部首有不同相对位置，如左右结构汉字“肝”“淡”，上下结构汉字“苔”“薄”，包围结构汉字“困”“闷”等。部首还具有表意作用，如“疒”部首[6]就和疾病方面具有密切关系，大部分包含该部首的汉字与疾病相关，如“病”“瘫”“疯”等；部首“月”出现在与肉体相关的字上，如“肢”“肠”等。汉字偏旁部首特征明显，目前常被广泛用于增强不同汉语自然语言处理任务。

为了更好地识别医案中的命名实体，本文以预训练模型 BERT[7]作为基础模型，并在输入的文本向量中嵌入偏旁部首特征进行训练，构建更加精准的融合汉字部首的 BERT - BiLSTM - CRF 命名实体提取模型。

2 模型与方法

BERT - BiLSTM - CRF 模型框架，见图 1[7]。首先，将文本数据经由 BERT 的嵌入层映射成字向量，然后输入到 Transformer 的编码层学习上下文特征；其次，文本向量与部首特征结合，经过双向长短期记忆人工神经网络进行语义编码处理；最后由条件随机场计算得到模型预测的标签序列。例如输入是“肢体活动障碍”，输出则是对应的预测标签。

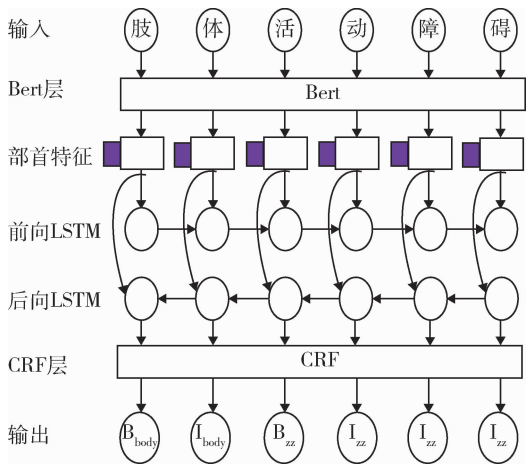

图 1 BERT - BiLSTM - CRF 模型

2.1 基于 BERT 的词向量生成

BERT 预训练模型是谷歌人工智能研究院于 2018 年 10 月提出的一种预训练语言模型，其在机

器阅读理解顶级水平测试中表现突出。预训练过程依据上下文语义信息对随机掩盖的词进行预测，可以更好地学习上下文内容特征；句间关系预测则为每个句子的句首和句尾分别插入［CLS］和［SEP］标签，通过学习句子间的关系特征预测两个句子的位置是否相邻。

BERT 模型采用 12 层或 24 层双向 Transformer 编码结构，其中输入文本被转换为词向量，并为每个句子加上开始及结束标志，由此得到输入向量 E_1，E_2，…，E_N，经过多层 Transformer 编码器后输出向量 T_1，T_2，…，T_N，见图 2。

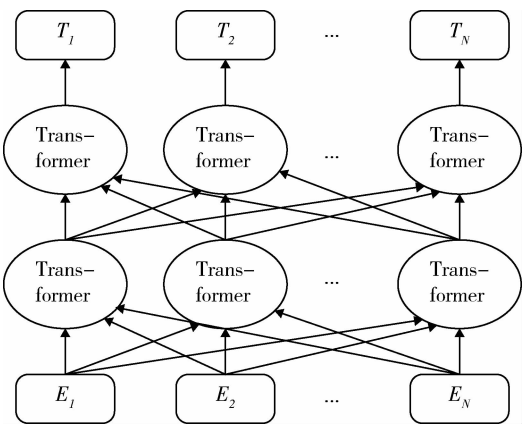

图 2 BERT 模型结构

在本文中，首先将训练数据转换为词向量，并为每个句子加上［SEP］和［CLS］标志，由此得到 BERT 的输入向量 E_1，E_2，…，E_N，经过多层 Transformer 编码器后得到输出向量 T_1，T_2，…，T_N，将输出向量 T_1，T_2，…，T_N 记为 A_1。然后将输入的训练数据与部首词典进行匹配，找到对应部首，将部首数据进行嵌入后得到向量 A_2，最后将向量 A_1 和 A_2 拼接起来作为双向长短期记忆人工神经网络的输入向量，以完成下游任务。

2.2 双向长短期记忆人工神经网络

近年来，长短期记忆人工神经网络[8]被广泛用于语音识别、自然语言处理等领域并获得巨大成功。一个基本的长短期记忆人工神经网络结构，见图 3。其具体参数如下所示：

$$i_t = \sigma(W_{hi}h_{t-1} + W_{xi}x_t + b_{hi} + b_{ii}) \quad (1)$$

$$f_t = \sigma(W_{hf}h_{t-1} + W_{xf}x_t + b_{hf} + b_{if}) \quad (2)$$

$$o_t = \sigma(W_{ho}h_{t-1} + W_{xo}x_t + b_{io} + b_{ho}) \quad (3)$$

$$\tilde{c}_t = tanh(W_{xc}x_t + b_{ic} + W_{hc}h_{t-1} + b_{hc}) \quad (4)$$

$$c_t = f_t c_{t-1} + i_t \tilde{c}_t \quad (5)$$

$$h_t = o_t tanh(c_t) \quad (6)$$

其中 x_t 为 t 时刻的输入，W 和 b 分别表示权重矩阵和偏置向量，h_{t-1}是 $t-1$ 时刻的隐藏层状态，h_t 是 t 时刻的隐藏层状态，c_t 是 t 时刻的细胞状态，i_t、f_t、o_t、$\tilde{c}_t$分别是输入门、遗忘门、输出门和细胞门的激活向量，σ 是 sigmoid 函数。对某时刻的输入，长短期记忆人工神经网络首先对细胞门中存储的历史信息 c_{t-1}进行处理，通过遗忘门层 f_t 根据上一时刻隐藏层输出 h_{t-1}与当前时刻的输入 x_t，控制细胞门状态中哪些信息将会得到保留，随后再将更新后的细胞门状态的激活值与输出门层 o_t 结合得到当前时刻的隐藏层状态并将其传递下去。

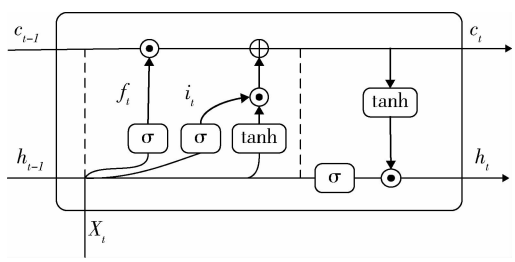

图 3 LSTM 单元结构

双向长短期记忆人工神经网络由前向长短期记忆人工神经网络和后向长短期记忆人工神经网络组成，其分别从两个不同的方向对输入序列进行处理，然后将得到的两个结果拼接起来，作为最终输出，从而更好捕捉双向的语义依赖。

2.3 条件随机场

长短期记忆人工神经网络输出的标签之间具有很强的依赖性。例如在命名实体识别中，一个实体的起始字 B 后不应再紧接着出现另一种实体的中间字 I，实体类型应该保持一致，或实体中间字 I 不能独立出现等。因此，最终预测标签序列中的各个标签并非独立出现，而是在特定约束条件下产生的观测值序列。条件随机场[8]能够利用长短期记忆人工

神经网络的输出序列作为特征。

对于给定的输入序列 $x = [x_1, x_2, \cdots, x_n]$，其输出预测序列为 $y = (y_1, y_2, \cdots, y_n)$，那么该预测序列分数可计算为：

$$s(x,y) = \sum_{i=0}^{n} A_{y_i,y_{i+1}} + \sum_{i=1}^{n} P_{i,y_i} \quad (7)$$

其中 A 被称为转移矩阵，条件随机场中需要训练的参数即为转移矩阵，它将对训练文本中出现的一些重要语法约束进行学习。P 被称为发射矩阵，在矩阵 P 中按照时间顺序选择的一条路径即为序列 y，则其出现概率可以由 Softmax 函数计算：

$$P(y \mid x) = \frac{e^{s(x,y)}}{\sum_{\tilde{y} \in y_x} e^{s(x,\tilde{y})}} \quad (8)$$

其中 y_x 是对输入序列 x 所有可能出现的预测序列观测值集合，在训练过程中需要对 y 的对数概率最大化：

$$\log(p(y \mid x)) = s(x,y) - \log\left(\sum_{\tilde{y} \in y_x} e^{s(x,\tilde{y})}\right) \quad (9)$$

预测过程中，选择预测结果 y^* 使其序列评分最高：

$$y^* = \underset{\tilde{y} \in y_x}{argmax} s(x,\tilde{y}) \quad (10)$$

3 语料来源与标注

3.1 语料来源

实验所用语料来自于中医典籍《中国现代名中医医案精粹丛书》中的第 1 本，选取其中 400 例医案共计 5 万余字进行标注。其中包含性别、姓名、年龄、症状、疾病、药物、部位、处方等信息。实验所用部首由在线新华字典爬取而来[6]，使用在线新华字典将中医医案数据集中的所有汉字全部查询一次，构建一个包含汉字 - 部首对的词典，并精心选取 12 个出现频率较高的部首（“疒”“艹”“口”“木”“氵”“扌”“辶”“亻”“纟”“月”“日”“忄”）作为最终部首字典，共计 740 个部首对。以部分部首词典为例：“肝，月”“咳，口”“瘦，疒”“胸，月”“药，艹”“病，疒”，其中每一条汉字 - 部首对都作为单独的一行。

根据医案文本特征并结合已有研究[2]中的方法使用正则表达式对其包含的症状和药物进行提取，通过人工校正最终得到症状字典（157 词）和药物字典（1 693 词）用于医案语料标注。

3.2 语料标注

结合双向最大匹配法[9]对语料进行标注。同时运用正向最大匹配法和逆向最大匹配法，比较两者结果，取分词数少的结果作为最终结果。将医案按照症状（symptom）、药物（medicine）、部位（body）、疾病（disease）4 种实体进行标注。随机从医案中选出 400 条，总计 53 000 余字进行标注，训练集和测试集为 3∶1。

使用“BIO”方式对语料进行实体标注。将实体起始字标记为“B”，若实体为单个字，也标记为“B”。将实体非起始字标记为“I”，余下非实体字全部标记为“O”。此外，采用半自动方式对语料进行标注。基于症状字典和药物字典使用双向最大匹配法，对匹配到的实体进行自动标注。没有匹配到的字则标为“O”，见图 4。

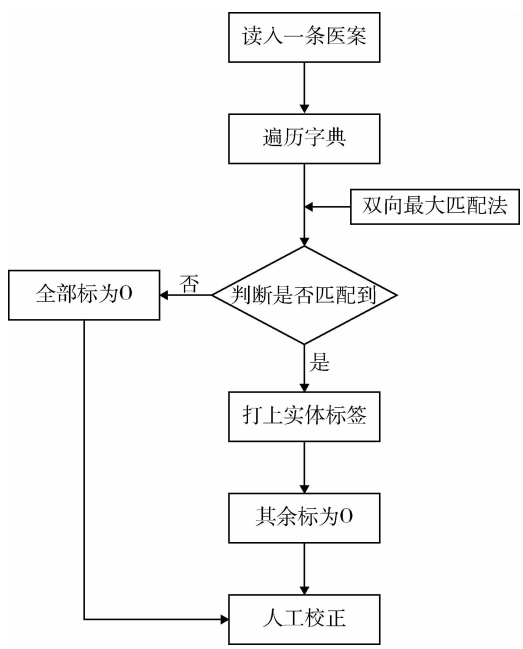

图 4 医案标注流程

标注完成后训练集与测试集中实体类型、实体数量，见表 1。

表 1 实体类型与实体数量（个）

实体类型	训练集	测试集
症状	3 424	1 113
部位	690	183
药物	150	76
疾病	219	69

4 实验与讨论

4.1 超参数设置

实验参数采用已有研究[7]中的默认参数。训练语料句子的最大长度设置为 480，双向长短期记忆人工神经网络的隐藏层单元数量设置为 128，训练轮数设置为 15。初始学习率设置为 0.000 05，实验批量大小设置为 4。

采用 Pycharm 作为主要程序开发平台，Python 版本为 3.6。操作系统为 Win 10，内存 16GB，硬件配置为 Intel(R)Core(TM)i5 - 10300H@2.50GHz。

4.2 实验结果

实验结果采用准确率 P、召回率 R、$F1$ 值 3 个指标进行评价，见表 2。其中第 1 行是基于医疗文本进行预训练的基线 BERT 模型。第 2 行加入条件随机场层，第 3 行在第 2 行基础上加入双向长短期记忆人工神经网络层，最后一行是 BERT - BiLSTM - CRF 模型融合部首特征。可以看到，在加入部首特征后，模型 $F1$ 值提升到了 84.81%。这表明，BERT - BiLSTM - CRF 模型的性能在嵌入部首特征后得到进一步提升，表现优异，更适合用于中医医案命名实体识别。

表 2 模型实验结果对比（%）

模型	P	R	$F1$
BERT	78.72	83.63	81.10
BERT - CRF	82.13	84.47	83.28
BERT - BiLSTM - CRF	84.49	84.60	84.55
BERT - BiLSTM - CRF - radical	84.26	85.37	84.81

BERT - BiLSTM - CRF - radical 模型在不同类型临床实体上的表现，见表 3。可以看出该模型对症状实体和药物实体的识别效果优于身体部位实体和疾病实体。其中模型对药物词的识别效果最好，$F1$ 值为 88.74%；对身体部位的识别效果最差，$F1$ 值为 83.29%。

表 3 不同实体识别结果对比（%）

实体类型	P	R	$F1$
部位	80.93	85.79	83.29
疾病	86.15	81.16	83.58
药物	89.33	88.16	88.74
症状	84.38	85.37	84.87

4.3 讨论

从以上结果可以看出，在 BERT 模型的基础上增加不同人工神经网络层以及部首特征，能够明显提高命名实体识别效果。条件随机场层是序列标注的重要组成部分，其中包含转移特征，可以为最后预测的标签添加一些约束以保证预测结果合理，特别是当标签之间存在依赖关系时。双向长短期记忆人工神经网络在长短期记忆人工神经网络的基础上对输入序列进行了额外的反向学习，使其对输入序列的部分语法规则学习更加准确，加强了对上下文所携带信息的记忆。由于医学命名实体的部首具有很强的实体类别特点，嵌入部首特征有效提高了模型性能。例如部首“疒”与疾病实体密切相关，将其添加到模型中能够强化对疾病实体识别的训练，从而提高模型性能。

本文采用由本研究团队从医案及中医诊断教材中抽取获得的词典对语料进行标注，提高了语料标注效率，随着医案命名实体提取算法的应用，术语词典中词的质量和数量也会改善，将促进语料标注效率提高。

在中医医案数据集中，疾病实体数量较少，且在实际标注过程中，发现疾病术语常分散在症状词中。此外，在中医术语中有些词既是疾病术语又是症状术语，如“脑震荡重度昏迷”，模型将疾病实体“脑震荡”错误地预测为症状实体。另一方面，症状词表述相对来说比较丰富，例如疼痛有“酸痛”“剧痛”“胀痛”等多种表达方式，导致模型对症状词的识别效果不佳。在中医医案中部位词常与方位词连用，这种细粒度的部位描述更加精准，

但也使部位实体数增加且重复率低，如“右胁”“左胸”“左耳”等，这些表述丰富灵活且重复率低的实体影响模型性能。针对上述问题，条件随机场具有一定局限性，无法根据具体语义对实体进行灵活识别，而双向长短期记忆则对学习内容具有长时记忆，能够对表述相对灵活的实体进行识别，从而提高模型性能。

对数据中没有被正确识别的命名实体进行分析主要存在以下问题。一是程度词与部位、症状、疾病等实体结合在一起时，模型将程度词预测为实体；例如句子“浮肿尤甚”，模型将“尤甚”预测为症状实体；句子“明显胀气”，模型将“明显”预测为症状实体。这主要是因为对医案实体的划分还不够全面。二是模型对长实体的识别效果不佳，例如症状实体“色素沉着皮损呈簇集性成群分布”，模型只将“色素沉着皮损”预测为症状实体。这是因为 BERT 模型对捕获长实体中的依赖关系效果不佳。

5 结语

中医医案是古今中医传承下来的宝贵财富，对指导现代中医发展具有重要作用。本研究所涉及的基于预训练 BERT 模型的中医医案命名实体识别方法中，融入部首特征的 BERT – BiLSTM – CRF – radical 模型最优，其 *F*1 值达到最高（84.81%）。但采用的实验数据量较少，不同实体分布不均衡，实体间存在相互嵌套的问题。在今后研究中将进一步增加实验数据和实体类别，均衡不同实体分布，并尝试使用其他方法提高模型性能。

参考文献

1 赵山，罗睿，蔡志平．中文命名实体识别综述［J］．计算机科学与探索，2022，16（2）：296 – 304.

2 肖晓霞，刘明婷，杨冯天赐，等．基于 NLP 的中医医案文本快速结构化方法［J］．大数据，2022，8（3）：128 – 139.

3 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（5）：20 – 24.

4 许力，李建华．基于 BERT 和 BiLSTM – CRF 的生物医学命名实体识别［J］．计算机工程与科学，2021，43（10）：1873 – 1879.

5 DEVLIN J，CHANG M W，LEE K，et al. BERT：pre – training of deep bidirectional transformersfor language understanding［C］. Minneapolis：Conference of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies（NAACL – HLT），2019.

6 李丹，徐童，郑毅，等．部首感知的中文医疗命名实体识别［J］．中文信息学报，2020，34（12）：54 – 64.

7 LI X，ZHANG H，ZHOU X H. Chinese clinical named entity recognition with variant neural structures based on BERT methods［J］. Journal of biomedical informatics，2020，107（5）：103422.

8 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.

9 陈之彦，李晓杰，朱淑华，等．基于 Hash 结构词典的双向最大匹配分词法［J］．计算机科学，2015，42（S2）：49 – 54.

关于《医学信息学杂志》启用“科技期刊学术不端文献检测系统”的启事

为了提高编辑部对于学术不端文献的辨别能力，端正学风，维护作者权益，《医学信息学杂志》已正式启用“科技期刊学术不端文献检测系统”，对来稿进行逐篇检查。该系统以《中国学术文献网络出版总库》为全文比对数据库，可检测抄袭与剽窃、伪造、篡改、不当署名、一稿多投等学术不端文献。如查出作者所投稿件存在上述学术不端行为，本刊将立即做退稿处理并予以警告。希望广大作者在论文撰写中保持严谨、谨慎、端正的态度，自觉抵制任何有损学术声誉的行为。

《医学信息学杂志》编辑部