DOI:10.13193/j.issn.1673-7717.2026.02.004

融合多维度信息的中医医案命名实体识别算法研究

陈隽柏[1],郭凤英[1],赵步天[1],杜松星[2],田甜[3],闫占峰[4],王丽[1]
(1.北京中医药大学管理学院,北京 102488;2.北京中医药大学东方医院,北京 100078;
3.北京中医药大学中医学院,北京 102488;4.北京中医药大学东直门医院,北京 100700)

摘要:目的 解决当前中医医案命名实体识别模型识别效果欠佳的问题,提升中医医案结构化的效率与精准度。方法 构建融合多维度信息的BDM-CRF命名实体识别神经网络模型,融合双向长短期记忆网络模型(Bidirectional long short-term memory,BiLSTM)、门控空洞卷积神经网络模型(Gated-dilated convolutional neural networks,DGCNN)和多头注意力机制模型(Multi-head attention,MHA)进行多维度特征抽取,通过条件随机场(Conditional random field,CRF)模型对特征抽取结果进行约束输出;通过对比实验、消融实验、预训练模型对比实验及神经网络结构对比实验验证该模型的有效性,通过个案分析验证并展示模型的提取效果。结果 本研究提出的模型比对照模型精确率平均提升2.30%、召回率平均提升4.73%、F1值平均提升3.53%。个案分析结果表明,本研究模型能正确识别出中医医案文本中的大部分实体内容。结论 本研究提出的融合多维度信息的中医医案命名实体识别模型能取得更佳的识别效果,可有效提升中医医案结构化的精准度及效率,促进中医医案处理智能化。

关键词:中医医案;命名实体识别;神经网络;模型融合
中图分类号:R2-03;R249 **文献标志码**:A **文章编号**:1673-7717(2026)02-0017-08

Research on Multi-dimensional Named Entity Recognition in Traditional Chinese Medicine Cases

CHEN Junbai[1],GUO Fengying[1],ZHAO Butian[1],DU Songxing[2],TIAN Tian[3],YAN Zhanfeng[4],WANG Li[1]
(1. Management school of Beijing University of Chinese Medicine,Beijing 102488,China;
2. Dongfang Hospital,Beijing University of Chinese Medicine,Beijing 100078,China;
3. School of Traditional Chinese Medicine,Beijing University of Chinese Medicine,Beijing 102488,China;
4. Dongzhimen Hospital,Beijing University of Chinese Medicine,Beijing 100700,China)

Abstract:*Objective* To solve the problem of poor recognition performance of current traditional Chinese medicine case naming entity recognition models,and to improve the efficiency and accuracy of the structuring of traditional Chinese medicine cases. *Methods* A BDM-CRF named entity recognition neural network model was constructed to integrate multi-dimensional information. The model integrated Bidirectional Long Short-Term Memory(BiLSTM),Gated-Dilated Convolutional Neural Networks(DGCNN) and Multi-Head Attention(MHA) models for multi-dimensional feature extraction. The feature extraction results were constrained by the Conditional Random Field(CRF) model. The effectiveness of the model was verified through comparative experiments,ablation experiments,pre-training model comparison experiments and neural network structure comparison experiments,validating and demonstrating the extraction performance of the model through case analysis. *Results* The model proposed in this study achieved an average improvement of 2.30% in precision,4.73% in recall and 3.53% in F1 score compared to those of the control model. The case analysis results indicated that the model in this study can correctly identify most of the entity content in the text of traditional Chinese medicine medical cases. *Conclusion* The proposed model for the named entity recognition in traditional Chinese medicine (TCM) medical records,which integrates multi-dimensional information,can achieve better recognition results,effectively improve the accuracy and efficiency of TCM medical record structuring,and promote the intelligent processing of TCM medical records.

Keywords:TCM case records;named entity recognition;neural network;model fusion

---

基金项目:国家自然科学基金青年科学基金项目(82204963);中央高校基本科研业务费专项(2024-JYB-PY-005);北京中医药大学揭榜挂帅教改项目(JX-JBGS-2501005);北京中医药大学基本科研业务费揭榜挂帅项目(2025-JYB-JBGS-009);北京中医药大学教育教学改革专项(xjyz25010);北京中医药大学自主选题青年教师项目(040201004002011);北京中医药大学研究生自主科研项目(ZJKT2024033)
作者简介:陈隽柏(2000-),男,广东佛山人,硕士,研究方向:健康大数据管理。
通讯作者:郭凤英(1979-),女,河南开封人,副教授,硕士,研究方向:智慧医疗、自然语言处理、中医药大数据。E-mail:guofy@bucm.edu.cn。

2019年10月20日中共中央、国务院发布《关于促进中医药传承创新发展的意见》指出要加强中医医案典籍研究利用，结合数字化、影像化技术实现中医传承现代化。中医医案是中医医家临床活动的记录文本，大多以非结构化形式存在[1]，是中医知识理论和实践经验的重要承载[2]。通过对中医医案进行结构化处理，提取出症状、疾病、治疗方法等关键实体信息，从而可以构建起一个清晰的中医医学知识体系，有助于系统化总结中医名家的实践经验，挖掘疾病用药治疗规律[3]。

命名实体识别（Named entity recognition，NER）是一种用于自动识别文本中的特定实体并分类为预定义类别的自然语言处理（Natural language processing，NLP）技术[4-6]。在面向中医医案的命名实体识别算法模型中，需要同时考虑文本中字符的基本排列顺序和组合方式等字符序列特征、前后文的语义联系等时序特征、字词在句子中的相对位置及其组合方式等空间特征及远距离段落间潜在的语义关系等中长距离相互依赖特征[7]。研究发现，由于中医数据量大而繁杂，特别是中医药古文献，多以非结构化的文言形式进行记录，常存在一词多义及实体边界模糊的问题，且在语言表达、文字风格和句式形式等方面与现代文本差异较大，人工提取步骤繁杂、工作量大，而传统基于词典及机器学习算法的特征提取效果不理想[8]。基于大规模预训练模型等深度学习方法的命名实体识别技术是目前研究的前沿和热点，凭借优秀的泛化能力，可以主动学习句子中隐藏的语义信息，提升中医医案结构化的效率与精准度[9]。然而，当前的深度学习模型通常仅使用单种特征提取模块，会导致忽视其他维度信息，造成识别效果欠佳[10]。基于此，本研究提出一种融合多维度信息的中医医案命名实体识别算法，通过融合双向长短期记忆网络模型（Bi-directional long short-term memory，BiLSTM）、门控空洞卷积神经网络模型（Gated-dilated convolutional neural networks，DGCNN）以及多头注意力机制（Multi-head attention，MHA）3种神经网络模型，分别抽取中医医案中的字符序列特征、前后文的语义联系的时序特征和字词在句子中的相对位置及其组合方式等空间特征，以及远距离段落间潜在的语义关系等中长距离相互依赖特征，将3种维度信息并连进行融合输出，最大程度提升上下文信息获取能力，提高中医医案命名实体识别的精度。

1 研究现状

目前针对中医医案的命名实体识别算法包括基于词典和规则的方法、以条件随机场（Conditional random field，CRF）为代表的机器学习方法和基于神经网络及大规模预训练模型的深度学习方法。

1.1 基于规则、机器学习的方法

基于规则及机器学习方法的中医医案命名实体识别模型较为成熟，如肖晓霞等[11]使用结合词典改进的N-gram模型识别医案文本中的症状、体征，实现了对4754份医案文本的结构化；高佳奕等[12]基于CRF实现从中医临床医案抽取症状命名实体，通过验证发现抽取结果符合中医辨证理论。但是，以上研究存在依赖分词效果、对较长实体识别效果较差等局限性，因此目前机器学习模型多与深度学习模型相结合用于实体输出约束[13]。

1.2 基于深度学习的方法

深度学习方法是目前研究的前沿和热点，具有无监督、自动学习等优势，能有效减少人工制定特征的工作量。李明浩等[14]基于长短期记忆递归神经网络（Long short-term memory，LSTM）-CRF实现了对中医医案症状术语的识别，与条件随机场方法相比，能更有效地识别字符较多的长症状术语；肖瑞等[15]基于BiLSTM-CRF模型实现中医医案命名实体识别，经人工测试后，模型实体识别效果良好；LIU Q P等[16]基于可动态更新分词策略的BiLSTM-CRF模型实现了中医胸部不适病例医案的命名实体识别，结果表明引入词典可有效提升命名实体识别模型的性能。以上研究较基于规则和机器学习方法在识别效果上获得了一定提升，但由于中医医案相关训练语料较为稀少，单纯的LSTM或BiLSTM模型主要依赖于模型自身的参数学习来捕获文本中的特征，在识别效率、特征表示能力上会受到较大限制。

1.3 基于大规模预训练模型的方法

大规模预训练模型通过对大规模无标注数据的预训练，通过遮蔽语言模型、双向Transformer等技术，充分捕捉字符、词、句子间关系，有效提升模型训练效果[17]。BERT是第一个采用无监督、深度双向机制的大规模预训练模型，应用于各种下游任务，可提升神经网络模型在命名实体识别任务的性能[4]。吴佳泽等[18]分别将BERT、RoBERTa、ALBERT 3种预训练模型与CRF结合，提升了中医医案命名实体识别效率；胡为等[19]提出基于中医医案训练的TcmYiAnBERT预训练模型，并在中医医案命名实体识别任务上实验证明其优越性。研究表明，运用大规模预训练模型能进一步提升中医医案命名实体识别效果。然而，以上研究仅使用了基于卷积神经网络（Convolutional neural network，CNN）卷积结构或循环神经网络（Recurrent neural network，RNN），特征抽取方式较为单一，上下文获取信息能力有限，不利于解决中医医案常见的一词多义及实体边界模糊的问题。

1.4 研究述评

目前基于大规模预训练模型的中医医案命名实体识别模型效果最佳，常用的模型结构是：利用BERT等预训练模型对中医医案语料进行向量化，利用BiLSTM模型进行特征抽取，利用CRF模型学习标签间的关系进行输出约束。该类结构无须依赖先验的分词工具，能够直接基于字符级别进行识别；通过大规模预训练模型的引入，避免了从头学习特征表示，在数据量不足或多样性较低的情况下，仍能捕捉到足够的语言特征[20]。以上结构能取得较为良好的效果，但仍有一定的提升空间。首先，大部分预训练模型使用通用语料进行训练，而中医医案文本具有大量专业术语，在表述方式和语义结构上可能与通用语料存在差异，采用基于中医医案训练的TcmYiAnBERT预训练模型能更为适应；其次，中医医案数据量一般较小，且存在一定的冗余信息，引入注意力机制可以将模型的信息处理能力更多关注于对识别模型有效的实体信息上，实现识别效率及效果的提升；最后，当前模型结构特征抽取层多使用基于RNN循环网络的BiLSTM模型，特征捕捉方式单一，无法对空间特征建模[21]，而融合基于卷积结构的神经网络可解决上述问题。基于此，本研究提出一种融合BiLSTM、DGCNN、MHA的神经网络架构，并结合CRF模型进行约束输出，构建

一种 BDM－CRF 的中医医案命名实体识别模型，以实现实体识别效果的提升。

2 命名实体识别模型构建

本研究提出的模型架构如插页Ⅶ图1所示，包括特征抽取层和输出约束层2个部分，模型的输入使用 TcmYiAnBERT 模型将中医医案文本序列转换为富含语义信息的嵌入向量。在特征抽取层，首先分别将嵌入后的向量输入到 BiLSTM 模型抽取字符序列特征，输入到 DGCNN 模型抽取出时序特征和空间特征；随后将嵌入后的向量、BiLSTM 模型的输出向量及 DGCNN 模型的输出向量输入到 MHA 模型，提取字符序列中长距离的相互依赖特征；在输出约束层，首先对 BiLSTM 模型、DGCNN 模型及 MHA 模型的输出向量进行向量拼接，接着通过多层感知机（Multi－layer perceptron，MLP）进行融合，然后经过全连接层和 Softmax 层后获得命名实体识别概率结果，将脉象、症状、药方等复杂实体识别和分离。最后通过 CRF 模型约束输出，得到最终识别结果。

2.1 特征抽取层

2.1.1 字符序列特征抽取模块 本研究使用 BiLSTM 模型抽取输入向量的字符序列特征，捕捉中医医案中语言表述中的序列依赖关系，有效处理文言文中的顺序敏感内容。LSTM 是一种特殊形式的 RNN，可以获取更大更长的距离信息，通过使用一组门控制器能有效解决了传统 RNN 的梯度消失和爆炸问题。BiLSTM 是由前向 LSTM 与后向 LSTM 组合而成，分别获取文本序列的前向信息和后向信息进行拼接得到最终的隐藏层特征表示，相较于单纯的 LSTM 模型能更好地捕捉双向语义依赖，对于语义的判别更为细致[22]。

本研究在特征抽取的过程中，利用公式（1）～（6），将嵌入后的向量 $X=[x_1,x_2,\cdots,x_n]$ 输入到字符序列特征抽取模块得到输出向量 $H_1=[H_1^1,H_1^2,\cdots,H_1^n]$。其中，输入门 i_t 控制当前单元的信息，遗忘门 f_t 控制信息的遗忘，输出门 o_t 控制信息的输出，短期记忆 h_t 记录了当前时刻的输出，长期记忆 c_t 把当前单元的状态保存并作为下一单元的输入，σ 为 Sigmoid 激活函数，W 为权重矩阵，b 为偏置项，x_t 为当前单元的输入，s_t 为当前时刻更新的状态，tanh 为双曲正切激活函数。

$$i_t=\sigma(W_i[h_{t-1},x_t]+b_i) \tag{1}$$

$$o_t=\sigma(W_f[h_{t-1},x_t]+b_f) \tag{2}$$

$$o_t=\sigma(W_f[h_{t-1},x_t]+b_o) \tag{3}$$

$$s_t=\tanh(W_s[h_{t-1},x_t]+b_s) \tag{4}$$

$$c=f_t*c_{t-1}+i_t*s_t \tag{5}$$

$$h_t=ot*\tanh(c_t) \tag{6}$$

2.1.2 时序特征和空间特征抽取模块 本研究使用 DGCNN 模型抽取输入向量的时序特征和空间特征，识别中医医案中的非线性结构，如药方和病情描述中的组合和相对位置。利用公式（7）～（9），将嵌入后的向量 $X=[x_1,x_2,\cdots,x_n]$ 输入到字符序列特征抽取模块得到输出向量 $H_2=[H_2^1,H_2^2,\cdots,H_2^n]$。DGCNN 是一种带有残差连接的门控空洞卷积神经网络模型，通过使用卷积层堆叠获取更多信息，卷积的操作如公式（7）所示。

$$c_t=W_c\oplus_{k=0}^{r}x_{t\pm k} \tag{7}$$

其中，W_c 为窗口大小 r 的滤波器，c_t 表示经过卷积变换后的输出，⊕表示向量拼接。在此基础上引入空洞实现对数据的间隔采样，实现对更长句子的覆盖，空洞卷积的计算方式如公式（8）所示。

$$c_t=W_c\oplus_{k=0}^{r}x_{t\pm k\delta} \tag{8}$$

其中 δ 为空洞率，W_c 为窗口大小 r 的滤波器，c_t 表示输出，⊕表示向量拼接。同时增加门控机制控制数据流通降低梯度消失的风险，具体结构如图所示。每一个卷积门控单元内的操作如公式（9）所示。

$$Y=X+\mathrm{Conv1D_1}(X)\otimes\sigma(\mathrm{Conv2D_2}(X)) \tag{9}$$

其中，X 代表输入，$\mathrm{Conv1D_1}$ 和 $\mathrm{Conv2D_2}$ 为权值不共享的两个参数结构相同的空洞卷积，σ 为 sigmoid 函数，⊗为向量 Hardamard 乘积。

2.1.3 中长距离的相互依赖特征抽取模块 本研究使用 MHA 抽取输入向量的中长距离的相互依赖特征，处理跨越多个词、句子甚至段落的复杂实体，从而缓解中医医案中的实体边界模糊问题[23]。注意力机制（Attention mechanism）通过模拟人脑注意力，对重要内容打分加权，以突出重要特征[24]。利用公式（10）～（12），将 X、H_1 及 H_2 共同输入到中长距离的相互依赖特征抽取模块得到输出向量 $H_3=[H_3^1,H_3^2,\cdots,H_3^n]$。

$$\mathrm{Attention}(Q,K,V)=\mathrm{Softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{10}$$

其中，Attention（Q，K，V）为注意力值，Q 为查询向量矩阵，K 为键向量矩阵，V 为值向量矩阵；d_k 为词向量维度，通过 QK_T 计算 Q 和 K 间相似度，除 d_k 以防结果过大，添加 Softmax 归一化处理并乘以值向量得最终注意力值。当向量的维度较高时，采取多头注意力机制确保获取单词的全部特征，将 Q、K、V 线性映射，形成多个子空间，拼接计算结果如公式（11）和公式（12）所示。

$$\mathrm{head_i}=\mathrm{Attention}(QW_i^Q,KW_i^K,VW_i^V) \tag{11}$$

$$H_3=\mathrm{MultiHead}(Q,K,V)=\mathrm{Concat}(\mathrm{head_X},\mathrm{head_{H1}},\mathrm{head_{H2}})W^O \tag{12}$$

2.2 输出约束层

本层主要由模型融合及输出约束两部分，模型融合主要通过向量拼接完成，输出约束使用 CRF 模型完成。首先将 2.1 节中的向量 H_1，向量 H_2，向量 H_3 进行拼接得到向量 $H_4=[H_4^1,H_4^2,\cdots,H_4^n]$，如公式（13）所示。

$$H_4=\mathrm{Concat}(H_1,H_2,H_3) \tag{13}$$

接着将向量 H_4 输入到多层感知机得到融合了字符序列特征、时序特征和空间特征及中长距离的相互依赖特征的（n×k）维向量 $H_5=[H_5^1,H_5^2,\cdots,H_5^n]$，其中，n 代表字符数，k 代表标签个数。通过融合多维度特征，充分考虑了中医医案中语言表达和风格差异带来的多样化特征；随后，将向量 H_5 输入到全连接层和 Softmax 层得到每个字符对应的最大概率标序列向量 $H_6(H_1,H_2,\cdots,H_k)$；最后，将向量 H_6 输入到 CRF 模型中，利用公式（14）～（16）进行输出约束，得到结果序列 Y。CRF 模型是一种序列建模框架，通过在标签间添加约束，赋予合理标签序列更大的概率值，能有效解决标签偏差的问题。对于输入序列 $A(a_1,a_2,\cdots,a_k)$ 得到其对应的 $Y=(y_1,y_2,\cdots,y_n)$

预测标签序列的概率公式如公式(14)所示。

$$P(Y \mid A) = \frac{e^{T(A,Y)}}{\sum_{Y' \in Y_H} T(A,Y')} \tag{14}$$

其中,Y'为真实标签值,Y_H 为可能的标签值,得到预测序列 Y 的似然函数如公式(15)所示。

$$\ln[P(Y \mid A)] = T(A,Y) - \ln\left(\sum_{Y' \in Y_A} T(A,Y')\right) \tag{15}$$

通过似然函数求出合理的标签序列并输出概率最大的标签序列,具体公式如公式(16)所示。

$$Y^* = \arg_{\tilde{y} \in Y_A} \max s(A,\tilde{y}) \tag{16}$$

3 实验设计

3.1 数据收集与清洗

本研究的中医医案数据来源于6册《中国现代名中医医案精粹》,其中收集了由国家中医药管理局、卫生部及人事部于2010年选出的434位名老中医的5172则医案,涵盖了中医内科、中医外科、妇科等多学科。

从以上医案文本中随机抽取1000则作为本研究实验数据集,通过 Python 3.7 编写光学字符识别(Optical character recognition, OCR)程序,将抽取医案的 PDF 文本转换为 TXT 文本格式,对照原始文本由两位北京中医药大学学生进行人工校正,将遗漏、乱码等错误之处进行修改,将两份校正结果进行比对,差异部分由课题组全体成员共同讨论得出最终结果;最后将修改校正后的医案数据按诊次分条存放至excel文件中,以患者ID为患者00001的患者为例,存放至 Excel 文件中的医案文本数据示例见图2。

3.2 数据标注

本研究在借鉴现有研究成果[18-19]的基础上设计了适用于中医医案的标注规范,定义了6种实体类型:舌像、脉象、症状、证候、治法、中药及用量,各实体标注样例如图3(a)所示。采用"BIO三元标注法"进行命名实体标注,其中 B-X 为命名实体起始,I~X 为命名实体中间或结束,O 为非命名实体,具体标注示例如图3(b)所示。

3.3 数据划分

为评估本研究提出模型的性能,将标注好的中医医案数据按照6∶2∶2的比例划分为训练集、验证集和测试集,具体实体个数统计见表1。

3.4 评价指标

本研究使用宏观精确率(P)、召回率(R)和F1值(F1)进行模型评价,具体计算公式(17)~(19)所示。

$$P = \frac{TP}{TP + FP} \tag{17}$$

$$R = \frac{TP}{TP + FN} \tag{18}$$

$$F1 = \frac{2PR}{P + R} \tag{19}$$

TP为真正例数量,FP为假正例数量,FN为假反例数量。

就诊ID	患者ID	医案内容
就诊000001	患者00001	王某,女,2岁半。初诊:1983年6月16日。主诉及病史(其父代诉):午后发热,右侧肢体活动障碍,左眼外斜视40 d。病始于1983年1月,持续高热1周,体温39~40 ℃,纳呆,消瘦,经某医院照胸片诊为右侧支气管淋巴结结核。住院接受链素、雷米封等治疗。2个月后体温逐渐恢复正常,胸片复查好转。但于1983年5月再次高烧,体温40 ℃以上,伴有嗜睡、喷射性呕吐、神志不清、阵阵抽搐。经查:瞳孔等大等圆,对光反射迟钝,项强,克尼格征阳性,右侧巴宾斯基征阳性。化验白细胞数11 200/mm²,中性30%,淋巴69%,嗜酸1%,腰穿脑脊液压力升高,化验脑脊液细胞数1150/mm²,白细胞32%,蛋白微量,糖5管试验弱阳性。诊为结核性脑膜炎。抢救治疗两天,神志苏醒,体温较前下降,但发现右侧肢体瘫软无力,右下肢步履拖拉,行走需人搀扶,右上肢上举受限,右手呈握拳状难伸,左眼向外斜视,精神迟钝,言语不清,夜睡不安、盗汗,虽仍继续以抗结核药物及"脉通液"等治疗;但每日午后体温仍持续在37~38 ℃,现已40 d余。诊查:体温37.5 ℃,神志尚清,答话含糊,面色不荣,精神不振,形体瘦弱,左眼向外斜视,右上肢上举,高不过头,右手指屈伸不能自如,呈握拳状拘挛,右下肢瘫软,舌红无苔,脉象细数。辨证:阴虚生热,伤及经络。治则:滋阴清热,通经活络。处方:鳖甲10 g,青蒿5 g,白薇5 g,丹皮10 g,地骨皮10 g,生地10 g,玉竹10 g,地龙10 g。
就诊000002	患者00001	6月27日。上方药服3剂,体温降至37.2 ℃以下。继续服用原方药7剂,潮热已解,精神好转,但便秘、腹胀、纳呆,舌质淡红、苔白厚,脉沉细而缓。诊为脾胃虚弱,食积停滞,治以健脾和胃,消积导滞。处方:陈皮10 g,炒神曲10 g,鸡内金10 g,炒麦芽10 g,代代花5 g,香橼5 g,丹参10 g,荷梗10 g。
就诊000003	患者00001	6月30日。进3剂,纳食稍好,大便未通。以原方加生大黄3 g(后下)、黄芩6 g、佩兰10 g、厚朴花5 g,以泄胃火、清肠热。

图2 医案文本数据样例

舌像:"舌红""无苔" 脉象:"细数" 症状:"午后发热""右侧肢体活动障碍""左眼外斜视40天""持续高热1周""体温39~40 ℃""纳呆""消瘦""右下肢瘫软""右上肢上举""右手指屈伸不能自如,呈握拳状拘挛""面色不荣""精神不振""形体瘦弱""左眼向外斜视""神志尚清""答话含糊""夜睡不安、盗汗""每日午后体温仍持续在37~38 ℃,现已40 d余" 证候:"阴虚生热""伤及经络" 治法:"滋阴清热""通经活络" 中药及用量:"鳖甲10 g""青蒿5 g""白薇5 g""丹皮10 g""地骨皮10 g""生地10 g""玉竹10 g""地龙10 g"	午 B–Symptom 后 I–Symptom 发 I–Symptom 热 I–Symptom , O 右 B–Symptom 侧 I–Symptom 肢 I–Symptom 体 I–Symptom 活 I–Symptom 动 I–Symptom 障 I–Symptom 碍 I–Symptom
a.实体标注样例	b. BIO标注样例

图3 命名实体标注示例

表1 中医医案数据集实体类型 单位:个

数据集	舌像	脉象	症状	证候	治法	中药及用量	总计
训练集	429	328	2729	560	493	4116	8655
验证集	165	139	999	251	186	1763	3503
测试集	153	126	1108	163	153	1600	3303

3.5 实验环境及超参数设置

本研究基于Python 3.7.13,使用PyTorch 1.6深度学习框架进行实验,优化器选用Adam模型,模型超参数见表2。

表2 模型超参数

超参数	取值
batch_size	8
dropout	0.5
learning_rate	0.000 05
epoch	64

3.6 基线模型

本研究实验主要从4个方面验证本文提出模型的有效性,①对比实验,为了验证本研究模型的优越性,本研究对比现有基于大规模预训练模型的中医医案命名实体识别模型。②消融实验,为了验证本研究提出的BDM神经网络架构的有效性,在本研究提出模型的基础上,依次删除DGCNN模型及多头注意力机制模型。③词嵌入方法对比实验,为了验证TcmYiAnBERT在中医医案命名实体识别任务中的性能,本研究对比测试TcmYiAnBERT及其他常用预训练模型的效果。④模型结构对比实验,以验证本研究提出的BDM并联模型结构与其他常用模型结构的效果以验证其有效性及优越性。以上所有实验均进行十折交叉验证以降低偶然性,提高结果可信度。

4 实验结果与讨论

4.1 对比实验

对本研究模型与前人模型进行对比实验。选取文献研究中常用的AlBERT-CRF[18]、BERT-CRF[18]、BERT-BiLSTM-CRF及TcmYiAnBERT-BiLSTM-CRF[19]进行对比,具体结果见表3。由表3可知,本研究模型F1值为73.67%,相比于AlBERT-CRF、BERT-CRF、BERT-BiLSTM-CRF及TcmYiAnBERT-BiLSTM-CRF模型F1值分别提升了5.02%、4.13%、2.59%和2.36%;相对于通用模型AlBERT-CRF、BERT-CRF及BERT-BiLSTM-CRF,本研究模型精确率及召回率均取得了提升,表明本研究模型相比于通用模型,在面向中医医案命名实体识别任务时能识别出更多正确的信息;相对于TcmYiAnBERT-BiLSTM-CRF模型,本研究模型尽管在精确率下降了1.87%,但能获得召回率更大幅度的提升(6.6%),表明本研究模型相比于单纯的BiLSTM模型尽管会带来更多的噪声,影响识别精度,但能获取更多的信息,识别出更多的实体,最终实现更佳的识别效果。

4.2 消融实验

在本研究模型的基础上进行消融实验,实验结果见表4。其中-MHA为删除了多头注意力机制,-DGCNN为删除了门控空洞卷积神经网络,-BDM为删除了整个特征抽取层。结果显示,消融BiLSTM模型使本模型的F1值降低了2.3%,消融MHA模型使本模型的F1值降低了1.61%,消融DGCNN模型使本模型的F1值降低了2.08%。以上结果表明,消融BDM模型3个特征抽取模块中的任一部分,模型的性能均呈现下降,且性能下降的程度相对均衡,尤其是消融BiLSTM及消融DGCNN后,两者F1值仅相差0.22%,表明不同模型抽取的特征信息对于中医医案命名实体识别任务均十分重要,进一步凸显了融合多维度信息对于提升中医医案命名实体识别准确度的必要性;同时,消融整个特征抽取层,即去除BMD模型的所有3个模块,仅使用TcmYiAnBERT进行向量嵌入后通过CRF识别输出,模型的F1值降低了3.87%,召回率降低了8.76%,表明本研究所提出BDM模型能显著提升中医医案信息抽取能力,在精确度相差不大的情况下能大幅增加实体的识出率,进一步表明本研究提出模型结构的有效性。

表3 对比实验结果 单位:%

模型	P	R	F1
AlBERT-CRF	68.34	68.95	68.65
BERT-CRF	66.58	72.76	69.54
BERT-BiLSTM-CRF	68.41	73.97	71.08
TcmYiAnBERT-BiLSTM-CRF	73.34	69.40	71.31
本研究模型	71.47	76.00	73.67

表4 消融实验结果 单位:%

消融模型	P	R	F1
本研究模型	71.47	76.00	73.67
-BDM	72.56	67.24	69.80
-BiLSTM	70.39	72.38	71.37
-MHA	71.13	73.02	72.06
-DGCNN	68.17	75.37	71.59

4.3 词嵌入方法对比实验

在本研究提出模型的基础上,将TcmYiAnBERT模型替换为其他常用的预训练模型,探究输入向量预训练模型的选取对实验效果的影响,具体结果见表5。结果显示,将TcmYiAnBERT模型替换为BERT模型,使实验的F1值降低了1.14%;将TcmYiAnBERT模型替换为AlBERT模型,使实验的F1值降低了1.78%。以上结果表明,中医领域专有TcmYiAnBERT模型已经提前学到大量中医医案语义信息,针对中医医案数据集的向量化比通用领域预训练模型更具优势,这与胡为等[19]研究结论一致。同时,与4.1中对比实验结果对比发现,本研究提出的模型在使用BERT通用预训练模型及中医领域专有TcmYiAnBERT预训练模型进行字符级嵌入后,与当前主流的BiLSTM模型相比,分别获得了1.45%、2.36%的F1值提升,表明本研究提出的模型针对中医医案数据集的文本向量化,能够通过融合多特征信息表现出更强的语义表征能力,进一步体现本研究模型的优越性及有效性。

4.4 模型结构对比实验

为探究多神经网络协同方式对实验效果的影响,对比集成结构、串联结果和本研究使用的模型结构(并联结构)。

(1)集成结构:通过将向量表示层得到的文本向量分别输

入字符序列特征抽取模块、时序特征和空间特征抽取模块、中长距离的相互依赖特征抽取模块，对 3 个模型的输出进行投票确定识别结果[25]。如图 4(a)所示。

表 5 词嵌入方法对比实验结果 单位：%

模型	P	R	F1
BERT	64.27	83.17	72.53
AlBERT	69.25	74.73	71.89
TcmYianBERT(本研究模型输入)	71.47	76.00	73.67

(2)串联结构：首先将文本向量输入字符序列特征抽取模块获取字符序列特征，然后将其输入到时序特征和空间特征抽取模块获取时序特征和空间特征，最后将前面获取的特征输入到中长距离相互依赖特征抽取模块获取中长距离相互依赖特征[26]。如图 4(b)所示。

(3)并联结构(本研究结构)：具体介绍可见第 2 节，结构如图 4(c)所示。

对比实验结果见表 6。结果显示，3 种不同模型结构获得的 F1 值均高于章节 4.1 中使用单一神经网络的模型，其中，本研究模型采用的并联结构获得的 F1 值为 73.67 最高，集成结构获得的 F1 值为 73.37，比本模型结构的 F1 值降低了 0.3%，串联结构获得的 F1 值为 72.07，比本模型结构的 F1 值降低了 1.6%。以上结果表明，在 3 种神经网络协同方式中，本研究选择的并联结构能获得了更丰富、准确的语义表示，一定程度提升模型的性能。主要原因是多神经网络模块融合的模型能通过整合多个基础神经网络实现降低误差，因此性能优于基础神经网络模型[7]。

表 6 模型网络结构对比实验结果 单位：%

神经网络结构	P	R	F1
集成结构	71.76	75.05	73.37
串联结构	67.24	77.65	72.07
并联结构(本研究模型)	71.47	76.00	73.67

4.5 个案分析

为了探究识别结果及其误差原因，本研究进行个案分析。随机选取一则中医医案文本输入到本研究提出模型中进行实体识别，并将模型的识别结果与手工标注结果进行对比，结果如图 5 所示。

结果显示，本研究模型能正确识别出中医医案文本中的舌

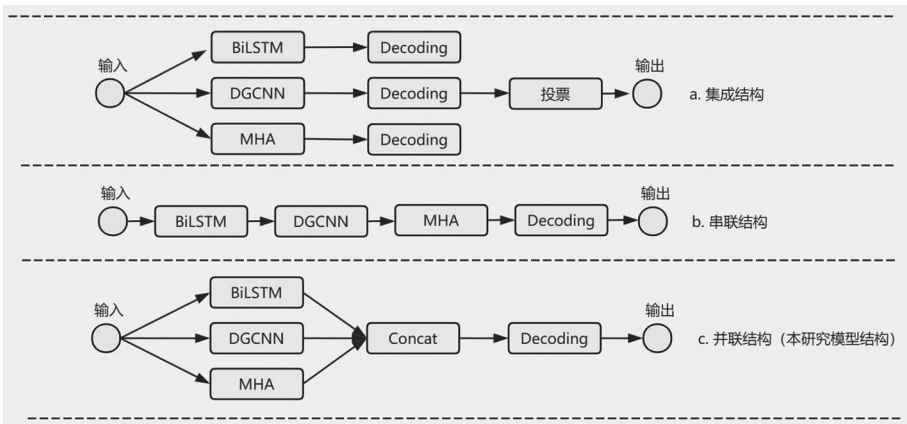


图 4 模型结构

医案原文

李某，女，28岁，已婚。初诊：1972年8月6日。1 d往往痛如劈裂，苦不可耐，常须注射止痛剂，并口服镇痛、镇静药，以求缓解痛苦。经行第2天后辄痛势递减，经净渐止。发作时伴头晕失眠，泛恶不食，烦躁易怒，目不欲睁，腰酸肢楚，口干咽燥，乳房作胀。平素月经周期或提前或错后，经量中等，色红间块。末次月经在7月10日。诊查：就诊时经期将届，正值头痛发作，舌边尖红，苔薄黄少津，脉细弦而数。辨证：为肝肾阴虚，水不涵木，肝阳上亢。治法：拟平肝潜阳，滋水涵木，疏风定痛之法。处方：钩藤9 g，菊花（后下)9 g，白蒺藜9 g，生石决24 g，杭白芍15 g，厚元参15 g，细生地15 g，女贞子9 g，香白芷1.8 g，北细辛1.8 g，生蔓荆6 g，香附米6 g，紫苏梗6 g，藁本6 g，川芎6 g，2剂。

实体类别	手工标注结果	本研究模型识别结果
舌像	舌边尖红、苔薄黄少津	舌边尖红、苔薄黄少津
脉象	脉细弦而数	脉细弦而数
症状	头痛、头晕、失眠、泛恶、不食、烦躁、易怒、目不欲睁、腰酸、肢楚、口干、咽燥、乳房作胀	头晕、失眠、泛恶、不食、烦躁、易怒、目不欲睁、腰酸、肢楚、口干、咽燥、乳房作胀、经量中等、色红间块、末次月经在、正值头痛、头痛发作
证候	肝肾阴虚，水不涵木，肝阳上亢	肝肾阴虚，水不涵木，肝阳上亢
治法	平肝潜阳，滋水涵木，疏风定痛	平肝潜阳，滋水涵木，疏风定痛
中药及用量	钩藤9 g，菊花（后下)9 g，白蒺藜9 g，生石决24 g，杭白芍15 g，厚元参15 g，细生地15 g，女贞子9 g，香白芷1.8 g，北细辛1.8 g，生蔓荆6 g，香附米6 g，紫苏梗6 g，藁本6 g，川芎6 g。	钩藤9 g，菊花（后下)9 g，白蒺藜9 g，杭白芍15 g，厚元参15 g，细生地15 g，女贞子9 g，香白芷1.8 g，北细辛1.8 g，生蔓荆6 g，香附米6 g，紫苏梗6 g，藁本6 g，川芎6 g。

图 5 手工标注结果与命名实体识别结果对比

像、脉象、症状、证候、治法、中药及用量实体类别的大部分实体内容，但在症状实体类别中仍存在部分错误，在中药及用量实体类别中存在个别错误，主要表现为：①未能正确分辨医案文本中对既往史的描述及症状描述。在本节选取的医案文本中，“平素月经周期或提前或错后，经量中等，色红间块。”为对患者既往月经情况的描述，但由于本研究中无既往史相关的实体分类，模型将其分类为了症状描述。②个别药名未能识别全。在本节选取的医案文本中，尽管大部分中药及其用量都被正确识别，但对于中药名称为“白蒺藜”的中药，本研究模型仅识别出了“白蒺”且同时忽略了其剂量。分析原因发现，TcmYianBERT 预训练模型中的 vocab 字典中没有“蒺”这个字，导致该字在向量嵌入时被识别为“UNK”。

5 结论

本研究在融合 BiLSTM、DGCNN、MHA 3 种神经网络模型的基础上，与 CRF 模型结合构建了一个 BDM－CRF 的中医医案命名实体识别模型。实验结果表明，本研究构建的模型比对照模型精确率平均提升 2.30%、召回率平均提升 4.73%、F1 值平均提升 3.53%，表明该模型相比对照模型具有更强的中医医案语义信息理解能力，取得了较好的识别效果。本研究仍存在不足：①仅针对特征提取层的神经网络模型进行优化，未对其他层进行优化。在未来工作中，应进一步针对不同层的特点和需求，设计不同的优化策略。②目前大规模预训练模型的训练语料缺少部分中医医案常见的生僻字，造成此类汉字无法被正确地嵌入，本研究收集的中医医案语料有限，尚不足以对预训练模型进行进一步微调以解决上述问题。在未来工作中，应获取更多中医语料信息，针对预训练模型进行微调，进一步提升模型性能。

参考文献

[1] 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.
[2] 王晗，马捷，易曾，等. 面向特定病症的中医医案语料库构建——以睡眠障碍病症为例[J]. 情报科学，2024，42(2)：97－108.
[3] JIA Q，ZHANG D Z，XU H F，et al. Extraction of traditional Chinese medicine entity：design of a novel span－level named entity recognition method with distant supervision[J]. JMIR Med Inform，2021，9(6)：e28219.
[4] 王颖洁，张程烨，白凤波，等. 中文命名实体识别研究综述[J]. 计算机科学与探索，2023，17(2)：324－341.
[5] BOSE P，SRINIVASAN S，SLEEMAN WC I V，et al. A survey on recent named entity recognition and relationship extraction techniques on clinical texts[J]. Appl Sci，2021，11(18)：8319.
[6] FU L，WENG Z Q，ZHANG J H，et al. MMBERT：a unified framework for biomedical named entity recognition[J]. Med Biol Eng Comput，2024，62(1)：327－341.
[7] 段宇锋，贺国秀. 面向中文医学文本命名实体识别的神经网络模块分解分析[J]. 数据分析与知识发现，2023，7(2)：26－37.
[8] 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.
[9] 孙超，张文博. 中医古籍文本术语命名实体识别的研究进展与挑战[J]. 中华中医药杂志，2021，36(11)：6843－6845.
[10] 李雯昕，张坤丽，关同峰，等. CHIP2020 评测任务 1 概述：中文医学文本命名实体识别[J]. 中文信息学报，2022，36(4)：66－72.
[11] 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.
[12] 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(6)：1947－1954.
[13] 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.
[14] 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.
[15] 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.
[16] LIU Q P，ZHANG L L，REN G，et al. Research on named entity recognition of Traditional Chinese Medicine chest discomfort cases incorporating domain vocabulary features[J]. Comput Biol Med，2023，166：107466.
[17] 杜晋华，尹浩，冯嵩. 中文电子病历命名实体识别的研究与进展[J]. 电子学报，2022，50(12)：3030－3053.
[18] 吴佳泽，李坤宁，陈明. 基于预训练模型及条件随机场的中医医案命名实体识别[J]. 中医药信息，2023，40(9)：38－45.
[19] 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.
[20] 王欣宇，杨涛，胡孔法. 基于大语言预训练模型的中医个性化处方推荐研究[J]. 中华中医药学刊，2024，42(4)：15－18，264.
[21] 杨长沛，廖列法. 基于门控空洞卷积特征融合的中文命名实体识别[J]. 计算机工程，2023，49(8)：85－95.
[22] 施顺顺，苟震宇，张琳，等. 基于深度域适应方法的非结构化摘要功能识别研究[J]. 现代情报，2025，45(10)：3－15.
[23] 武帅，杨秀璋，何琳，等. 基于句法特征和 Bert－BiLSTM－MHA－CRF 的细粒度古籍实体识别研究[J]. 数据分析与知识发现，2024，8(12)：136－148.
[24] 张卫东，陈希鹏，李心怡，等. 基于多任务联合学习的长白山民间文学实体抽取方法研究[J]. 现代情报，2025，45(5)：15－23，98.
[25] LI Z，GAN Z，ZHANG B，et al. Noisy Label Learning for Chinese Medical Named Entity Recognition Based on Uncertainty Strategy[EB/OL]. [2022－08－08]. https://bj. bcebos. com/v1/conference/ccks2020/eval_paper/ccks2020_eval_paper_3_1_1. pdf.
[26] HU J，SHI X，LIU Z，et al. HITSZ_CNER：A Hybrid System for Entity Recognition from Chinese Clinical Text[EB/OL]. [2022 08－08]. https://ceur－ws. org/Vol－1976/paper05. pdf.