多特征融合的中医古籍医案命名实体识别研究*

张璐瑶[1] 束建华[1] 王 鹏[2,3] 阚红星[1] 徐永祥[1] 周 洁[2] 唐书宣[1]

([1] 安徽中医药大学医药信息工程学院 合肥 230012 [2] 安徽中医药大学中医学院 合肥 230012
[3] 安徽中医药大学新安医学与中医药现代化研究所 合肥 230012)

〔摘要〕 **目的/意义** 构建中医古籍医案命名实体语料库，提升通用领域命名实体识别模型在中医古籍医案领域的识别精度与适用性。**方法/过程** 制定中医古籍医案命名实体标注规范，并据此对 2 384 则新安医案进行人工标注。构建 RoBERTa – BiLSTM – CRF 中医古籍医案命名实体识别模型，利用 RoBERTa 预训练语言模型生成具有语义特征的字向量，利用 BiLSTM – CRF 模型学习序列全局语义特征并解码输出最佳标签序列。引入词典和规则特征，增强模型对实体边界和类别的感知能力。**结果/结论** 模型在所建立的新安医案命名实体语料库上展现了良好的识别效果。融合领域术语词典与规则特征后，模型的综合 *F*1 值提升至 72.8%。

〔关键词〕 中医古籍医案；命名实体识别；语料库；词典；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 – 6036.2024.11.008

Named Entity Recognition of Traditional Chinese Medicine Ancient Records Based on Multi – feature Fusion

ZHANG Luyao[1], *SHU Jianhua*[1], *WANG Peng*[2,3], *KAN Hongxing*[1], *XU Yongxiang*[1], *ZHOU Jie*[2], *TANG Shuxuan*[1]
[1] *School of Medical Informatics Engineering, Anhui University of Chinese Medicine, Hefei 230012, China*; [2] *School of Chinese Medicine, Anhui University of Chinese Medicine, Hefei 230012, China*; [3] *Institute of Xin'an Medicine and Modernization of Traditional Chinese Medicine, Anhui University of Chinese Medicine, Hefei 230012, China*

〔**Abstract**〕 **Purpose/Significance** To construct a named entity corpus of traditional Chinese medicine (TCM) ancient records, and to improve the recognition accuracy and applicability of the general domain named entity recognition (NER) model in the field of TCM ancient records. **Method/Process** Annotation standards for entities in TCM ancient records are formulated, and 2 384 Xin'an medical records are annotated. A RoBERTa – BiLSTM – CRF model is developed, and word vectors with semantic features are generated using the RoBERTa pre – trained language model. The BiLSTM – CRF model is used to learn the global semantic features of sequences and decode and output the optimal label sequence. Dictionary and rule features are incorporated to enhance the model's capability to recognize entity boundaries and categories. **Result/Conclusion** The model shows a good recognition effect on the named entity corpus of Xin'an medical cases. Integration of domain terminology dictionaries and rule – based features improves the overall *F*1 score to 72.8%.

〔**Keywords**〕 traditional Chinese medicine (TCM) ancient records; named entity recognition (NER); corpus; dictionary; natural language processing (NLP)

〔修回日期〕 2024 – 07 – 27

〔作者简介〕 张璐瑶，实验师，发表论文 3 篇；通信作者：束建华，教授，硕士生导师。

〔基金项目〕 中央财政中医药事业传承与发展专项经费资助基金项目（项目编号：RZ2200001383）；安徽省高校协同创新项目（项目编号：GXXT – 2023 – 071）；安徽省高等学校科学研究重大项目（项目编号：2024AH040143）。

1 引言

中医古籍医案蕴藏着丰富的中医药临床知识，将这些临床知识显式化以辅助临床医生决策，是一项十分有意义、有价值的研究任务。近年来，临床医学文本中命名实体抽取受到了广泛关注[1]。但中医古籍医案的特殊性如专业深度、结构复杂与表达非标准化为命名实体识别带来诸多挑战。首要难题是高质量标注数据稀缺。其次是医案文本规范化不足，无标点、简写、错字情况频现，阻碍通用模型理解与泛化。再者，嵌套实体识别研究不足，如中药“黄芩”有“酒炒黄芩”“炒黄芩”“芩”等多种表述，增加了识别难度。已有研究[2-3]利用双向长短期记忆网络（bidirectional long short - term memory，BiLSTM）和条件随机场（conditional random field，CRF）搭建通用命名实体识别模型 BiLSTM - CRF，并在中医医案数据集上进行尝试。然而通用框架命名实体识别类别不全面，在处理多义词及嵌套实体等问题时仍有不足[4-5]。

中医古籍医案命名实体识别任务复杂，需要根据领域数据构建最佳识别模型。鉴于此，本研究提出一种融合领域术语词典和规则等多特征的中医古籍医案命名实体识别框架。以新安医案为数据源，建立新安医案命名实体语料库，基于鲁棒优化版双向编码器表征法（robustly optimized BERT pretraining approach，RoBERTa）联合 BiLSTM - CRF，构建 RoBERTa - BiLSTM - CRF 命名实体识别模型，融合领域术语词典和规则特征，对疾病、症状、证候、病因病机、治则治法、处方和药物 7 类命名实体进行识别，综合 $F1$ 值达 72.8%。

2 相关工作

中医药领域语料库构建标准尚未形成统一共识，公开标注的语料较少，大多数研究基于私有标注数据集[6-7]。刘丽红等[8]围绕中医药古籍文献数据，开展命名实体标注规范制定研究，为后续中医药文献语料标注工作提供了重要参考。中医古籍医案与中医药古籍文献的规范程度不同，其标注规范的制定需结合古籍医案语料自身特点。

命名实体识别是指从非结构化文本中提取指定类别的实体词。随着深度学习技术的发展，命名实体识别方法开始以神经网络模型为主[9-10]。在中医药领域，命名实体识别研究主要分为基于字切分和基于词切分两种方法[11]。羊艳玲等[12]采用词切分方法对名老中医治疗高血压医案进行实体识别，然而词切分方法易出现错误切分，会造成后续模型无法识别的问题。高佳奕等[13]改变序列切分方法，采用基于字切分的方法处理中医医案，有效缓解了词切分错误问题。中医药专业术语多为四字词组，基于字切分方法在表示中医药术语特征方面尚有不足[14]。为提高模型对中医药专业术语的识别能力，有研究[15-16]提出引入领域词典方法。如将中医领域词典融入 BiLSTM - CRF 模型，识别准确率较不引入词典方法提高 2.54%[16]。还有研究[17]提出利用词向量表示中文字词方法。肖瑞等[18]采用 Word2Vec 方法生成医案词向量矩阵，再利用 BiLSTM - CRF 模型提取命名实体。但是 Word2Vec 方法生成的词向量具有唯一性，无法应对多义词表征问题。为此，有研究[19]提出利用预训练语言模型生成动态词向量。BERT 预训练语言模型能够基于上下文语境实时生成词向量，有效解决相同词语不同语义的表示问题[20]。屈丹丹等[21]对比 BERT 模型与 Word2Vec 方法对肺癌医案症状命名实体识别的影响，结果表明 BERT 模型表示效果优于 Word2Vec 模型。预训练语言模型因出色的词向量表示能力，常作为序列字符特征提取器与 BiLSTM - CRF 模型联合处理命名实体识别任务[22-23]。

3 实验方法

中医古籍医案命名实体识别框架包含构建命名实体语料库和构建命名实体识别模型两部分，见图 1。命名实体语料库构建采用人工标注；命名实体识别模型构建则基于命名实体识别框架 BERT - BiLSTM - CRF[1]，将命名实体识别分为 3 个层次：数据表示层、语句建模层和标签标注层。

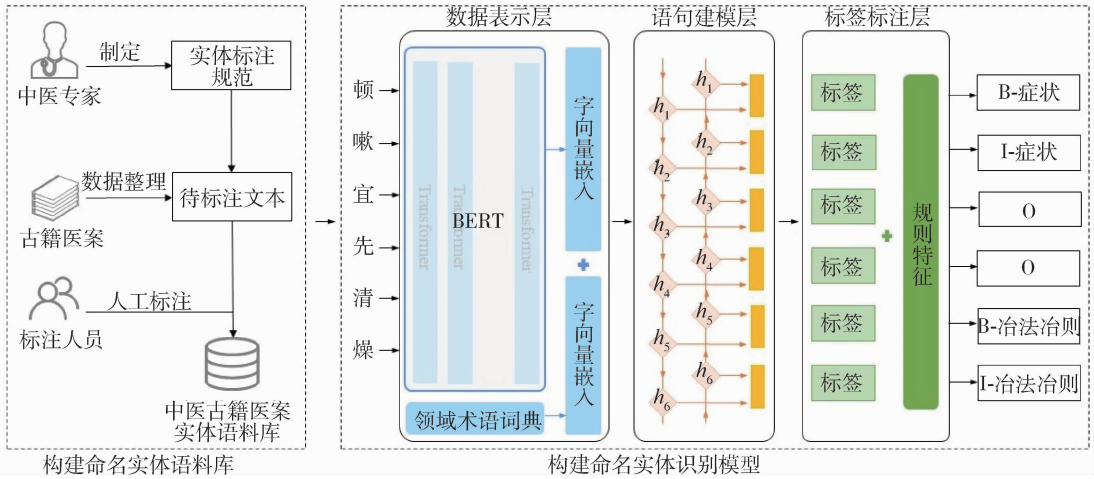

图 1 中医古籍医案命名实体识别框架

3.1 构建中医古籍医案命名实体语料库

选取 2 384 则新安医案作为数据源，内容涵盖内科、外科、妇科及儿科等多科病种。整个标注过程分为 3 个阶段，见图 2。

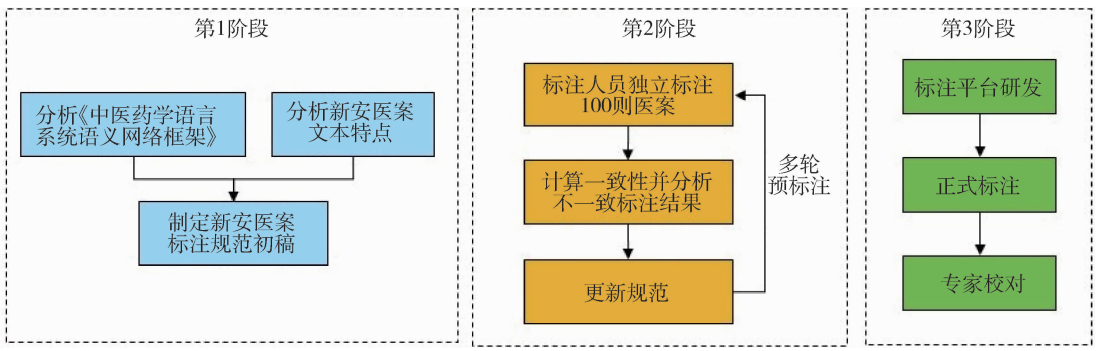

图 2 新安医案命名实体标注过程

3.1.1 第 1 阶段：制定新安医案标注规范 借鉴中文电子病历命名实体标注方法，参考《健康信息学 中医药学语言系统语义网络框架（GB/T 38324—2019）》分类体系，结合新安医案语料特点，遵照最小化原则，制定以中医疾病、症状、证候、病因病机、治法治则、处方、药物为核心要素的 7 类命名实体标注规范[24]。

3.1.2 第 2 阶段：实体预标注及标注规范更新 对标注人员统一培训，充分熟悉标注规范后，每位标注人员独立标注同一批新安医案，计算标注结果一致性情况，分析不一致标注原因，进而不断更新命名实体标注规范。经过多轮预标注后，标注规范趋于稳定，当标注一致性达到 80%，视为可以开展第 3 阶段工作。

3.1.3 第 3 阶段：正式标注 利用自主研发的“新安医案智能解析系统”正式标注新安医案。标注完成后，由 1 名中医学领域专家和 1 名中医药信息学领域教师对命名实体进行校对。

3.2 构建命名实体识别模型

3.2.1 数据表示层 负责将医案文本转变为包含字义、词义和句义特征的向量矩阵，并与领域术语词典的汉字编码向量拼接，组成包含术语语义特征的医案文本字向量。基于 BERT 预训练语言模型，利用多向度注意力层理解每个字在语句中的上下文关系，通过叠加归一组件连接注意力层

与前馈网络层的输入输出，生成医案字符动态编码向量，再与领域术语词典的汉字编码向量进行拼接，最终得到融合词典特征的字向量特征，具体过程，见图 3。

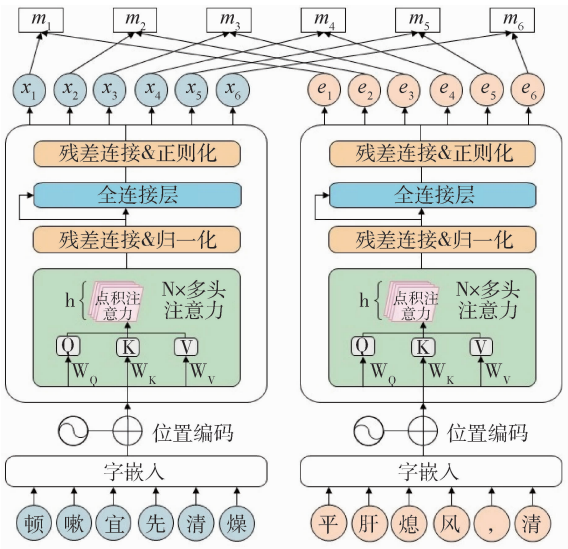

图 3 融合词典特征的字向量生成过程

语句 $S = [x_1, x_2, \cdots, x_n]$ 包含 n 个汉字，经过嵌入层，与位置编码矩阵相加得到大小为 $n \times m$ 的字向量矩阵 I。I 分别与权值矩阵 W_q、W_k、W_v 相乘得到查询矩阵 Q、键矩阵 K 和值矩阵 V。语句 S 中每个汉字 x_i 的动态编码向量根据公式（1）—（3）计算得出。其中 Z_i 表示单个注意力机制单元的自注意力得分。多个自注意力单元分值矩阵拼接后与权值矩阵 W 进行线性映射，获得多头注意力得分矩阵 Z，再与字向量 I 相加和正则化，经过全连接层后再与下一个编码器叠加归一化处理，最终得到每个汉字的动态编码向量 X。同样，将术语词典文本按医案文本处理后，得到每个词典的汉字编码向量 E。最后将医案文本字向量 x_i 与词典字向量 e_i 拼接，得到嵌入字向量 m_i。

$$Z_i = \text{Softmax}\left(\frac{Q \cdot K^T}{\sqrt{d_k}}\right)V \tag{1}$$

$$Z = \text{Concatenate}(Z_1, Z_2, \cdots, Z_h)W \tag{2}$$

$$X = \text{LayerNorm}(I + Z) \tag{3}$$

$$m_i = x_i \oplus e_i \tag{4}$$

3.2.2 *语句建模层* 语句建模层通过学习表示层的字向量特征构建医案语句模型，计算每个字向量所属实体类别的分值。其接收数据表示层传来的字向量，利用 BiLSTM 序列模型双向学习序列特征，构建语句模型，判断每个汉字对应的类别并输出分值，见图 4。

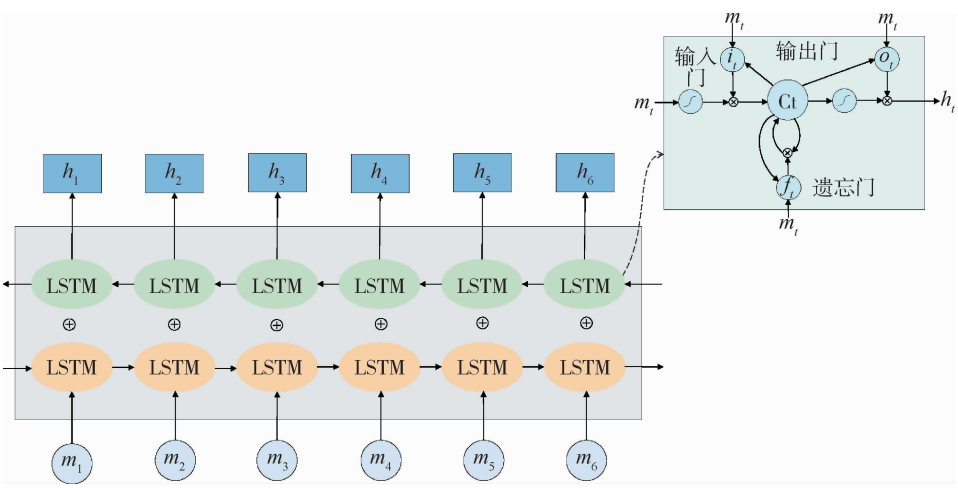

图 4 双向长短期记忆模型

BiLSTM 包含一层前向长短期记忆单元（long short - term memory，LSTM）和一层后向 LSTM。LSTM 利用输入门 i_t、输出门 o_t 和遗忘门 f_t 计算输入数据的语义特征。

$$i_t = \sigma(W_{xi}m_t + W_{hi}h_{t-1} + W_{ci}c_{t-1} + b_i) \tag{5}$$

$$f_t = \sigma(W_{xf}m_t + W_{hf}h_{t-1} + W_{cf}c_{t-1} + b_f) \tag{6}$$

$$o_t = \sigma(W_{xo}m_t + W_{ho}h_{t-1} + W_{co}c_t + b_o) \tag{7}$$

$$c_t = f_t c_{t-1} + i_t \tanh(W_{xc}m_t + W_{hc}h_{t-1} + b_c) \tag{8}$$

$$h_t = o_t \tanh(c_t) \tag{9}$$

在 t 时刻，数据表示层传入的字向量 m_t 作为

LSTM 层的输入，与 $t-1$ 时刻 LSTM 层的隐状态 h_{t-1} 结合，得到当前时刻隐状态 h_t。拼接前向层 LSTM 隐状态 $\overrightarrow{h_t}$ 和后向层 LSTM 隐状态 $\overleftarrow{h_t}$，得到 BiLSTM 的输出隐状态 $h_t = (\overrightarrow{h_t} \oplus \overleftarrow{h_t})$。

3.2.3 标签标注层 标签标注层通过学习实体类别标签之间约束关系，选择最佳实体类别标签并输出。其利用 CRF 模型学习类别标签约束特性，解决实体标签偏置问题，见图 5。

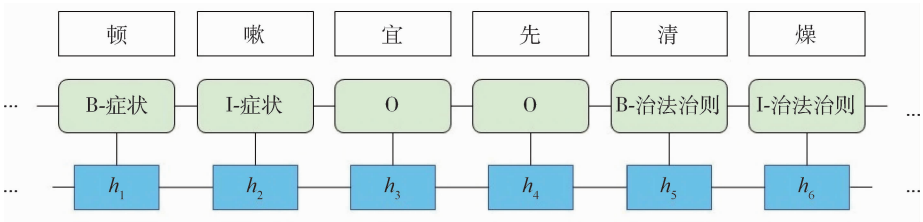

图 5 CRF 模型预测类别标签

CRF 模型是在给定观测序列 X 的条件下，计算标记序列 Y 的条件概率 $P(Y \mid X)$。观测矩阵 B 是 BiLSTM 层输出的 h_i 对应到实体类别 y_i 的概率，转移矩阵 A 是输出 y_{i-1} 转移到 y_i 的概率。模型对所有标签归一化计算，得到当前给定输入 h_i 条件下类别标签 y_i 的分值 $S(h_i, y_i)$。分值 S 再经过正则化处理得到输出标签 y_i 的概率 $P(y_i \mid h_i)$。

$$S(h_i, y_i) = \sum_{i=1}^{n} (A(y_{i-1}, y_i) + B(y_i, h_i)) \tag{10}$$

$$P(y_i \mid h_i) = \frac{exp(S(h_i, y_i))}{\sum_{y} exp(S(h_i, y_i))} \tag{11}$$

CRF 模型在训练时采用极大似然估计方法得到模型参数，预测时采用维特比算法得出待标注序列的最佳标签路径 Y。

$$Y = (y_1^*, \cdots, y_n^*) = \text{argmax}(P(y_i \mid h_i)) \tag{12}$$

3.3 评价指标

实验采用准确率（precision，P）、召回率（recall，R）和综合 $F1$ 值（$F1$ - measure，$F1$）作为衡量指标。将标注数据集中标注为非 O 实体的标签视为正样本 S，标注为 O 实体的标签视为负样本 N。TS 表示正确识别正样本的样本数量，FS 表示错误识别正样本的样本数量，TN 表示正确识别负样本的样本数量，FN 表示错误识别负样本的样本数量。准确率为正确识别的正样本数量占所有识别为正样本总样本量的比值，该指标衡量模型正确识别命名实体的准确程度。召回率为正确识别的正样本数量占实际正样本总样本量的比值，该指标衡量模型识别整个语料库中所有实体的能力。$F1$ 值是宏观意义上准确率与召回率的平均值，代表模型综合识别性能。

$$P = \frac{TS}{TS + FS} \tag{13}$$

$$R = \frac{TS}{TS + FN} \tag{14}$$

$$F1 = \frac{2 \times P \times R}{P + R} \tag{15}$$

4 实验结果与分析

4.1 实验准备

原始医案数据形式为自由文本，实验先对其进行规范化整理，包括删除重复医案和多余标点符号，修正错别字，规范药物别名，补全指代词和简写词等，见表 1。实体标注类别包括疾病、症状、证候、病因病机、治法治则、处方和药物 7 种。标注采用 BIO 命名实体标注方案，B - X 代表命名实体起始，I - X 代表命名实体中间或结束，O 代表非命名实体。经多轮人工标注，最终构建包含76 515 个命名实体的新安医案命名实体语料库。

表 1 医案预处理示例

新安医案	预处理前	预处理后
医案名称	痰阻气逆致咳	痰阻气逆致咳
作者	——清末民国·王仲奇《王仲奇医案》	清末民国·王仲奇
出处	（案 16）左 初诊（佚）	《王仲奇医案》
医案原文	二诊 呼吸不畅，咳痰不利，胸膈间如束缚。年老气怯，气不运痰，痰濡滞于络中，呼吸之道未能畅通，故见证如此。但心胸中仍时难过，大便溏泻，舌腭间糜腐复起，此则如腐草生菌、朽木生耳，有诸内然后形诸外也，久病如此，生气索然矣。脉濡滑，神疲欲眠，精神元气之不振可知矣。远志肉（炙）一钱、茯苓三钱、石斛二钱、冬虫夏草一钱二分、川贝母（去心）钱半、赖橘红八分、生苡仁三钱、冬瓜子四钱、鹅管石（煅透）一钱、十大功劳叶三钱、丝瓜络三钱、伽楠香（剉研细末冲）一分。	呼吸不畅，咳痰不利，胸膈间如束缚。年老气怯，气不运痰，痰濡滞于络中，呼吸之道未能畅通，故见证如此。但心胸中仍时难过，大便溏泻，舌腭间糜腐复起，此则如腐草生菌、朽木生耳，有诸内然后形诸外也，久病如此，生气索然矣。脉濡滑，神疲欲眠，精神元气之不振可知矣。远志肉一钱、茯苓三钱、石斛二钱、冬虫夏草一钱二分、川贝母钱半、赖橘红八分、生苡仁三钱、冬瓜子四钱、鹅管石一钱、十大功劳叶三钱、丝瓜络三钱、伽楠香一分。

将医案按 3∶1 比例随机划分为训练集数据和测试集数据。新安医案命名实体类别及数量，见表 2。

表 2 新安医案命名实体类别及数量（个）

数据集	疾病	症状	病因病机	证候	治法治则	处方	药物
训练集	3 499	14 185	3 291	425	2 240	1 583	32 166
测试集	1 165	4 728	1 097	141	746	527	10 722
合计	4 664	18 913	4 388	566	2 986	2 110	42 888

实验在 PyTorch 2.0.1 框架下进行，采用 NVIDIA GeForce RTX 4090 GPU 进行加速计算。字嵌入维度设为 728，BERT 模型最长文本序列长度设为 512，BiLSTM 模型隐藏层数设为 256。为了缓解 BiLSTM 模型在训练过程中的过拟合现象，引入随机失活策略（dropout），将 dropout 比率设定为 0.5。此外，为了进一步增强模型的学习能力与收敛速度，采用适应性矩估计（adaptive moment estimation，Adam）梯度下降算法，学习率调至 0.001。

4.2 字向量方法对模型性能提升的影响

为分析字向量表示方法作为中医古籍医案命名实体识别框架数据表示层的可行性及有效性，实验以 BiLSTM－CRF 为基准模型，分别对比直接字符嵌入方式、Word2Vec 静态字向量表示方法及 BERT 动态字向量表示方法对新安医案的语义表征能力，同时以 2019 年全国知识图谱与语义计算大会（China conference on knowledge graph and semantic computing，CCKS 2019）任务一面向中文电子病历的命名实体识别测评数据作为佐证，见表 3。

表 3 模型识别效果对比（%）

数据集与模型	准确率	召回率	*F*1
新安医案			
BiLSTM－CRF	54.2	55.4	54.8
Word2Vec－BiLSTM－CRF	57.1	55.6	56.3
BERT－BiLSTM－CRF	74.9	65.8	70.1
CCKS 2019			
BiLSTM－CRF[25]	73.6	67.5	70.4
Word2Vec－BiLSTM－CRF[26]	79.6	80.0	79.8
BERT－BiLSTM－CRF[27]	83.8	84.1	83.9

字向量表示方法效果优于直接字符嵌入方式。BERT 动态字向量表示方法通过学习上下文语境实时生成字向量，同一个汉字在不同语句中可以生成不同向量矩阵。这较 Word2Vec 生成的唯一字向量方法语义更丰富，更适合表示中医古籍医案多义词

的语义特征。同样的结论在 CCKS 2019 电子病历数据集上得到验证，进一步验证了在自然语言处理领域中以 BERT – BiLSTM – CRF 联合模型作为命名实体识别任务主流框架的做法。

4.3 5 种中文 BERT 模型识别性能对比

实验进一步对比 5 种中文 BERT 预训练语言模型对中医古籍医案文本表征能力，见表 4。5 种模型分别是中文 BERT 模型（BERT – base）[28]，中文 BERT 全词掩码模型（BERT – wwm）[29]，中文 RoBERTa 全词掩码模型（RoBERTa – wwm）[30]，纠错型掩码语言模型（masked language model as correction BERT，MacBERT）[31]，以及基于中文古籍语料训练的 Siku – RoBERTa 模型[32]。

表 4 不同预训练模型命名实体识别结果（%）

模型	准确率	召回率	*F*1
BERT – base – BiLSTM – CRF	71.1	69.1	70.1
BERT – wwm – BiLSTM – CRF	74.1	67.2	70.5
RoBERTa – wwm – BiLSTM – CRF	78.1	65.4	71.1
MacBERT – BiLSTM – CRF	73.3	67.7	70.4
Siku – RoBERTa – BiLSTM – CRF	62.2	49.4	55.1

RoBERTa – wwm 模型综合表现最佳，准确率和综合 *F*1 值分别为 78.1% 和 71.1%。这是由于 RoBERTa – wwm 模型改进了 BERT – base 和 BERT – wwm 模型的训练样本生成策略，采用动态调整掩码机制，并增大训练数据量和时长，表征中文文本能力显著提升。MacBERT 模型利用相近词替代目标词的方式再次优化掩码训练机制，但在新安医案数据集中选择的替代词关联度不足，导致对新安医案语义特征捕捉不够准确。Siku – RoBERTa 模型对中医药领域术语理解有限，未能充分展现其在中医古籍方面的优势。

4.4 融合词典与规则特征对模型识别性能的影响

实验进一步分析以 RoBERTa – wwm 为数据表示层的 RoBERTa – wwm – BiLSTM – CRF 模型对新安医案 7 种命名实体的识别情况，见表 5。该模型对药物、症状和疾病 3 类实体识别效果较好，综合 *F*1 值分别为 86.3%、82.1% 和 77.5%，对证候和处方实体识别效果较差，综合 *F*1 值分别为 55.9% 和 58.5%。

表 5 新安医案命名实体识别结果（%）

实体类型	准确率	召回率	*F*1
疾病	80.1	75.1	77.5
病因病机	78.9	61.6	69.2
药物	89.5	83.4	86.3
证候	62.3	50.7	55.9
症状	87.2	77.5	82.1
治法治则	68.7	55.1	61.2
处方	69.8	50.4	58.5
总体	78.1	65.4	71.1

分析识别结果发现，模型在标注实体类别时，易将处方实体错识为治法治则实体，易将证候实体错识为病因病机实体，而病因病机和治法治则实体易出现识别不全、识别不出等问题，导致非实体标签正确识别数量降低，RoBERTa – wwm – BiLSTM – CRF 模型召回率偏低。这些问题一方面是由样本分布不均衡造成，另一方面是模型对领域术语实体边界识别模糊导致的。

鉴于新安医案包含较多中医药领域术语，采用引入领域术语词典方法提高模型对实体边界的检测能力。选取《中医临床诊疗术语》《中国医学大辞典》《黄帝内经》《伤寒论》《本草纲目》等著作作为词典数据源，提取包含证候、处方、中医术语等要素共计 47 546 个词条。利用 RoBERTa – wwm 模型生成词典字向量矩阵，再与医案文本字向量拼接，最后传入语句建模层。另外，医家表述病因病机和治则治法时习惯使用几种固定句式，如“法当 XXX”“治以 XXX”“乃 XXX 也”等。“法当”“治以”通常紧跟治则治法的表述，“乃 XXX 也”多用来描述病因病机。因此，对病因病机和治则治法实体设计相应规则模板：设语句 $S = C_l W C_r$，其中 S 表示语句，C_l 表示实体前面的词，C_r 表示实体后面的词，W 表示命名实体，‖ 表示或关系，&& 表示与关系。设函数 $f_{\text{isfaze}}(W)$ 表示将 W 识别为治法治则实体；设函数 $f_{\text{isyinji}}(W)$ 表示将 W 识别为病因病机

实体。

规则 1 C_l = '法当' ‖ C_l = '法以' ‖ C_l = '法' → $f_{\text{isfaze}}(W)$

规则 2 C_l = '治以' → $f_{\text{isfaze}}(W)$

规则 3 C_l = '此' ‖ C_l = '乃' && C_r = '也' → $f_{\text{isyinji}}(W)$

对特征融合情况进行消融实验，见表 6。引入领域术语词典特征增强了模型对实体边界的辨识能力，规则特征帮助模型辨别病因病机、治法治则等实体类别，进而提高识别效果。融合词典和规则特征后模型准确率、召回率及综合 *F*1 值分别提高了 0.4、2.5 和 1.7 个百分点。

表 6 融合词典与规则特征的新安医案命名实体识别结果（%）

模型	准确率	召回率	*F*1
RoBERTa - wwm - BiLSTM - CRF	78.1	65.4	71.1
+ 词典	78.3	66.0	71.6
+ 规则	78.3	66.9	72.1
+ 词典 + 规则	78.5	67.9	72.8

5 结论

本文对中医古籍医案命名实体识别进行实验研究，以新安医案为数据源，建立新安医案命名实体语料库，构建 RoBERTa - wwm - BiLSTM - CRF 中医古籍医案命名实体识别模型。针对病因病机、治则治法及处方等命名实体识别率不高的情况，提出融合领域术语词典和规则的多特征融合识别方案，经验证，模型识别综合 *F*1 值达 72.8%。

尽管本研究取得了初步成果，但实体识别的精准度仍有提升空间，尤其是对于处方、治则治法和证候的识别。问题根源在于医案实体的样本量相对有限，导致样本分布不均，限制了模型的学习深度。此外，模型对中医术语的深层次语义表征不足，特别是在区分同名异义词时，如“补中益气”既可以是处方名称，也可以是治则治法，这增加了识别的复杂性。尽管模型能够准确划定此类实体的边界，但在判断其具体类别时仍存在挑战。后续将继续探索多元化的学习方法，进一步提升命名实体识别的准确性和泛化能力。

作者贡献：张璐瑶负责实验设计与结果分析、图表制作、论文撰写；束建华、王鹏、阚红星负责论文修订；徐永祥负责实验实施；周洁、唐书宣负责实验数据收集。

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 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.

2 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.

3 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.

4 刘华云，韩晨静，熊婕，等. 针刺临床文献自然语言处理中术语的智能化标注和抽取方法［J］. 中国针灸，2022，42（3）：327 - 331.

5 包振山，宋秉彦，张文博，等. 基于半监督学习和规则相结合的中医古籍命名实体识别研究［J］. 中文信息学报，2022，36（6）：90 - 100.

6 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.

7 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.

8 刘丽红，付璐，姚克宇，等. 中医药古籍文献实体标注规范探索［J］. 中华医学图书情报杂志，2022，31（12）：1 - 6.

9 LI I，PAN J，GOLDWASSER J，et al. Neural natural language processing for unstructured data in electronic health records：a review［J］. Computer science review，2022，46（11）：100511.

10 沈蓉蓉，夏帅帅，晏峻峰. 命名实体识别在中医药领域研究进展［J］. 医学信息学杂志，2023，44（1）：47 - 53.

11 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.

12 羊艳玲，李燕，钟昕妤，等. 基于 BiLSTM－CRF 的中医医案命名实体识别［J］. 中医药信息，2021，38（11）：15－21.

13 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.

14 李黛男，崔家鹏. 关于中医术语歧义问题的研究与思考［J］. 光明中医，2022，37（8）：1486－1490.

15 WANG Q，ZHOU Y，RUAN T，et al. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition［J］. Journal of biomedical informatics，2019，92（4）：1－9.

16 LIU Q，ZHANG L，REN G，et al. Research on named entity recognition of traditional Chinese medicine chest discomfort cases incorporating domain vocabulary features［J］. Computers in biology and medicine，2023，166（11）：107466.

17 李枫林，柯佳. 基于深度学习的文本表示方法［J］. 情报科学，2019，37（1）：156－164.

18 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.

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

20 DEVLIN J，CHANG M，LEE K，et a1. Bert：pre－training of deep bidirectional transformers for language understanding［C］. Minneapolis：The 2019 Conference of the North American Chapter of the Association for Computational Linguistics，2019.

21 屈丹丹，杨涛，朱垚，等. 基于字向量的 BiGRU－CRF 肺癌医案四诊信息实体抽取研究［J］. 世界科学技术－中医药现代化，2021，23（9）：3118－3125.

22 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.

23 SONG B，BAO Z，WANG Y. Incorporating lexicon for named entity recognition of traditional Chinese medicine books［C］. Zhengzhou：The CCF International Conference on Natural Language Processing and Chinese Computing，2020.

24 国家标准化管理委员会. 健康信息学 中医药学语言系统语义网络框架：GB/T 38324—2019［S］. 北京：中国标准化研究院，2019.

25 SONG Y，LUO L，LI N，et al. NER－PS－MS：medical attribute extraction based on medical named entity recognition［C］. Hangzhou：The Evaluation Tasks at the China Conference on Knowledge Graph and Semantic Computing，2019.

26 JI B，LI S，YU J，et al. Research on Chinese medical named entity recognition based on collaborative cooperation of multiple neural network models［J］. Journal of biomedical informatics，2020，104（4）：103395.

27 乔锐，杨笑然，黄文亢. 基于 BERT 与模型融合的医疗命名实体识别［C］. 杭州：全国知识图谱与语义计算大会，2019.

28 TURC L，CHANG M，LEE K，et al. Well－read students learn better：on the importance of pre－training compact models［EB/OL］.［2024－06－19］. https：//arxiv.org/abs/1908.08962.

29 CUI Y，CHE W，LIU T，et al. Pre－training with whole word masking for Chinese BERT［J］. IEEE/ACM transactions on audio，speech，and language processing，2021，29（11）：3504－3514.

30 OTT M，EDUNOV S，BAEVSKI A，et al. Fairseq：a fast，extensible toolkit for sequence modeling［EB/OL］.［2024－06－19］. https：//arxiv.org/abs/1904.01038.

31 CUI Y，CHE W，LIU T，et al. Revisiting pre－trained models for Chinese natural language processing［C］. online：The 2020 Conference on Empirical Methods in Natural Language Processing：Findings，2020.

32 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.