计算机科学
COMPUTER SCIENCE

# 融合BERT模型与词汇增强的中医命名实体识别模型

李旻哲, 殷继彬

引用本文

李旻哲, 殷继彬. 融合BERT模型与词汇增强的中医命名实体识别模型[J]. 计算机科学, 2024, 51(6A): 230900030-6.

LI Minzhe, YIN Jibin. TCM Named Entity Recognition Model Combining BERT Model and Lexical Enhancement [J]. Computer Science, 2024, 51(6A): 230900030-6.

---

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.230900030

# 融合 BERT 模型与词汇增强的中医命名实体识别模型

李旻哲　殷继彬
昆明理工大学信息工程与自动化学院　昆明 650500
(597085899@qq.com)

**摘　要**　现有的中医命名实体识别相关研究较少,基本都是基于中文病例做相关研究,在传统中医编写的病例文本中表现不佳。针对中医案例中命名实体密集且边界模糊难以划分的特点,提出了一种融合词汇增强和预训练模型的中医命名实体识别方法 LEBERT-BILSTM-CRF。该方法从词汇增强和预训练模型融合的角度进行优化,将词汇信息输入到 BERT 模型中进行特征学习,达到划分词类边界和区分词类属性的目的,提高中医医案命名实体识别的精度。实验结果表明,在文中构建的中医病例数据集上针对 10 个实体进行命名实体识别时,提出的基于 LEBERT-BILSTM-CRF 的中医案例命名实体识别模型综合准确率、召回率、F1 分别为 88.69%,87.4%,88.1%,高于 BERT-CRF,LEBERT-CRF 等常用命名实体识别模型。

**关键词**:自然语言处理;中医案例;词汇增强;BERT;BLSTM-CRF
**中图分类号**　TP391

## TCM Named Entity Recognition Model Combining BERT Model and Lexical Enhancement

LI Minzhe and YIN Jibin
Faculty of Information Engineering and Automation,Kunming University of Science and Technology,Kunming 650500,China

**Abstract**　There are few researches on TCM named entity recognition,and most of them are based on Chinese medical cases,and they do not perform well in TCM case texts. Aiming at the characteristics of dense named entities and fuzzy boundary in TCM cases,this paper proposes a method of TCM named entity recognition,LEBERT-BILSTM-CRF,which combines lexical enhancement and pre-training model. This method is optimized from the perspective of the fusion of vocabulary enhancement and pre-training model,and the vocabulary information is input into the BERT model for feature learning,so as to achieve the purpose of dividing word class boundaries and distinguishing word class attributes,and improve the accuracy of TCM medical case named entity recognition. Experiments show that when ten entities are identified on the TCM case data set constructed in this paper,the comprehensive accuracy rate,recall rate and F1 of the TCM case named entity recognition model based on LEBERT-BILSTM-CRF is 88.69%,87.4% and 88.1%,respectively. It is higher than common named entity recognition models such as BERT-CRF and LEBERT-CRF.

**Keywords**　Natural language processing,Chinese medicine case,Vocabulary enhancement,BERT,BiLSTM-CRF

## 1　引言

中医医案是中医传承与创新的主要载体之一,记录了中医诊疗的全过程,拥有层次丰富的医疗知识,具有人工智能服务的研究空间,如自然语言处理、智能问答等。但在如今医案的应用研究中,仍存在着数据利用率低、文本信息特征提取困难等问题。此外,中医医案是以长段且无序的非结构化文本为数据载体,其中包含了病人的疾病、症状、证候、处方及个人信息,且没有统一的标注语料和标注规范,导致传统的数据挖掘方法精度较低。

中医病历中包含了大量医疗信息,对中医案例进行实体识别是构建中医信息库的第一步。目前,针对实体识别的研究已经有很多,但是在中医案例方面的相关研究较少,中医实体识别方面仍存在一些亟待解决的问题,如实体边界模糊、实体歧义性等。针对中医医案中命名实体密集且边界模糊难以划分的特点,提出了一种融合词汇增强和预训练模型的中医命名实体识别方法。该方法从词汇增强和预训练模型融合的角度进行优化,通过优化后的预训练模型与 BiLSTM 准确地划分词类边界。条件随机场(Conditional Random Fields,CRF)用于区分不同词类属性标签,提高中医医案命名实体识别的精度。

## 2　相关工作

命名实体识别方法的发展整体上经历了基于规则与词典的方法,以条件随机场(CRF)为代表的统计机器学习和以循环神经网络(RNN)、预训练模型(BERT)为代表的深度学习 3 个阶段。基于规则与词典的方法和统计机器学习模型都需要依靠逻辑设计和训练语料中的统计信息来进行手工设计得到大量特征。这些统计学习方法的识别性能很大程度上依赖于特征的准确度,所以要求团队中要有语言学专家,十分费时费力[1-3]。

随着人工智能技术的快速发展,深度学习取得了长足的

通信作者:殷继彬(41868028@qq.com)

进步。深度学习的主要优势在于可以克服传统机器学习技术中需要人工提取特征的缺陷，无需复杂的特征工程，就能够从海量的数据中自动获取更加精确、抽象的信息，具备出色的泛化能力。

Zhao 等[4]提出将实体识别看作分类任务，使用多标签卷积神经网络(Convolutional Neural Networks，CNN)达到实体识别的目的，得到比传统识别方法更优秀的效果。Cao 等[5]将 E-CNN 和 BiLSTM-CRF 进行结合，解决实体边界划分不准确造成复合实体识别困难的问题。Guo 等[6]提出将 Transformer-XL(Transformer-extra long)与 BiLSTM 模型进行融合使用，解决在仅使用 Transformer 时在方向上信息较少的问题。Derlin 等[7]提出了一款基于 Transform 模型训练出的 BERT 模型，通过对 BERT 模型进行微调，能够适应很多不同类型的命名实体任务。Yan 等[8]对 transform 进行进一步调整，使用于命名实体识别。Lasri 等[9]对 BERT 在句法运用上进一步分析。Ma 等[10]在 Lattice-LSTM 上进行修改，将词典特征的上下文直接与 BERT 训练后的结果进行融合，实现词典信息与预训练模型相结合。Liu[11]提出通过将词典信息与 BERT 中的 Transform 层结合，使 BERT 能学习到更多特征。Ren 等[12]在对渔业标准划分中使用了融合注意力机制和 BERT-BiLSTM-CRF，召回率、准确率等均超过使用其中单一几个网络。Liu 等[13]将轻量级 BERT(A Lite BERT，ALBERT)与两个 BiLSTM 结合，实现了序列标注的高精度化，在实验中达到了 91.56%的准确率。为同时获得全局语义信息和方向信息，Liao 等[14]提出使用注意力机制动态融合 Transformer 编码器和 BiLSTM 的模型。

## 3 数据集构建

中医案例与其他领域的文本不同，它是一种具有客观性、程式性的特殊文本。它含有各种专有名词，如治方、病名、治法等多种常见实体属性以及证候等文言文中医实体。

中医的医疗案例有多种形式，优秀的医疗案例应当将理、法、方、药结合起来，展现出辩证施治的完整过程。医案的记录应当包括患者的病史、症状、脉象、舌象等，以便深入探究疾病的发病机制，并据此制定治疗方案和药物。本文旨在为中医命名实体识别研究提供一个基础数据资源，促进中医命名实体识别领域的发展。

### 3.1 中医命名实体类别选择

中医案例基本由患者信息、患者病史、患者医案、中医诊断、西医诊断、治法、方药、处方这 8 个部分组成。

根据以上信息构建一条完整中医医案的需求，本文选取消化科的中医案例作为研究主体，并依据医案写作特点和数据分布定义了病名、性别、年龄、诱因、症状、舌象、脉象、证候、治法以及治方共 10 个重要的实体类别。

### 3.2 中医命名实体类别选择

为了收集中医案例文本数据，本文使用网络爬虫工具批量爬取公开的中医平台病例数据。为了提高数据集的质量和准确性，本研究对网络爬虫抓取的数据进行了人工筛选及实体标注，最终保留了 800 份真实有效的中医确诊病例。这些病例中实体总数为 53360 个。各实体类型的数量及分布情况如表 1 所列。“治方”类实体占 38%，“症状”类实体占 26%，“治法”类实体占 10%，“证候”类实体占 9%，这 4 类实体占实体总数的 83%；其余实体占总数的 17%。

表 1 命名实体数量

Table 1 Number of named entities

| 命名实体 | 实体数量 |
|---|---|
| 病名 | 2852 |
| 性别 | 800 |
| 年龄 | 800 |
| 证候 | 4903 |
| 症状 | 13653 |
| 治法 | 5321 |
| 治方 | 20412 |
| 舌象 | 2614 |
| 脉象 | 1336 |
| 诱因 | 669 |

## 4 模型框架

模型通过融合字、词级别的语义特征，增强模型对于中医案例中实体潜在特征的提取能力，提高模型对于命名实体的识别能力，达到提升模型性能的目的。如图 1 所示，该模型由 3 个部分组成：输入表示层 LEBERT、特征提取层 BiLSTM 以及标签解码层 CRF。首先将字符序列传入 BERT 预处理模型中，BERT 会依据字符的位置向量、字向量和文本向量加和得到最终的输入向量。编码后获取上下文语义信息，将向量输入 BILSTM 网络模型。经过 BILSTM 的双向编码后，最终输出给 CRF 选择最合理的标签序列。

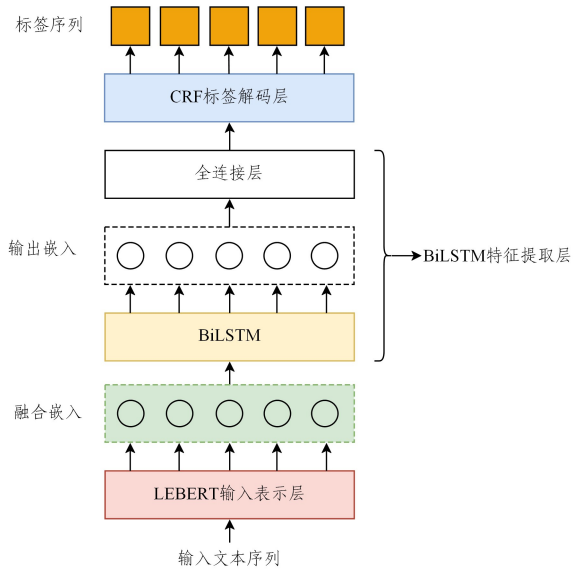



图 1 模型结构图

Fig. 1 Model structure diagram

### 4.1 输入表示层 LEBERT

为了解决中医病例中各个标签中命名实体密集且边界模糊的特点，在 BERT 模型的基础上采用 LEBERT 模型，获取每个字相关的词，利用词汇信息对命名实体进行边界划分，达到划分命名实体边界的目的。

与 BERT 相比，融合词汇增强的预训练模型 LEBERT 将汉语句子转换为字词对序列，将字符和词汇特征作为输入，输出字词对序列。Transform 层之间连接一个词典适配器，使得词汇信息能够有效地融合到 BERT 中。本文所使用的 LEBERT 的实现原理如图 2 所示，其中$C_n$表示每个字符，$W_n$

表示每个$C_n$匹配的词语，$h_n$表示经过 Transform 层后的数据。

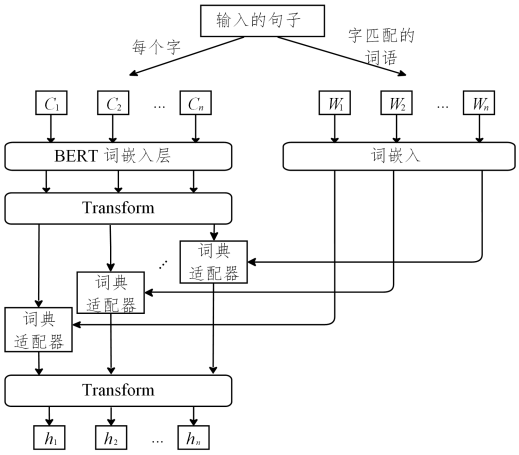



图 2　LEBERT 模型

Fig. 2　LEBERT model

正文内容使用一个词典 $D$ 和一组含有 $n$ 个字符组成的汉语句子，以词典作为基础，将字符序列 $S$ 中所有可能包含的词语找出来。如图 3 所示，以“参苓白术散”为例，共获取 3 个隐藏词，“参苓”“参苓白术散”“白术”(〈pad〉为空填充)。每个字获得对应的隐藏词，隐藏词即获得配对的字符序列，计算式如式(1)所示。

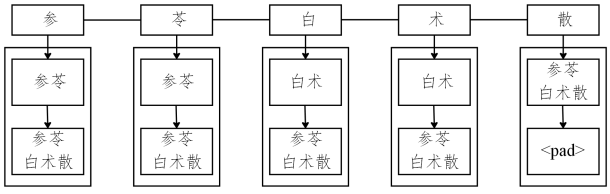



图 3　输入“参苓白术散”

Fig. 3　When the input is “Ginseng and Atractylodes Formula”

$$S_{CW}=\{[C_1,w_1],[C_2,w_2],[C_3,w_3],\cdots,[C_n,w_n]\} \tag{1}$$

通过对 BERT 适配器进行研究，实现将外部词典信息特征与句子信息的融合。将字符序列嵌入信息$x_{ij}^{w}$与 Transform 层输出$h_i^c$融合，$x_{ij}^{w}$表示如式(2)所示：

$$x_{ij}^{w}=e^{w}(w_{ij}) \tag{2}$$

其中，$e^{w}$是一个预先训练的单词嵌入查找表，$w_{ij}$是$w_i$中的第 $j$ 个单词。由于两者的向量长度存在差异，通过式(3)使之对齐。

$$v_{ij}^{w}=W_2(\tanh(W_1x_{ij}^{w}+b_1))+b_2 \tag{3}$$

其中，$W_1$是$d_c-d_w$的矩阵，$W_2$是$d_c-d_c$的矩阵，$b_1$和$b_2$是偏置值，$d_w$和$d_c$分别为单词嵌入的维数和 BERT 的隐藏大小。这里计算每个词的最大可能性，如上图所示，若“参苓”两字后没有“白术散”，则会分配“参苓”词组作为“参”和“苓”的最匹配词组。$v_{ij}^{w}$分配的给第 $i$ 个字符，表示为$V_i=(v_{i1}^{w},v_{i2}^{w},v_{i3}^{w},\cdots,v_{im}^{w})$，大小为 $m-d_c$的矩阵，其中 $m$ 为分配的词总数。每个单词间相关性计算式如式(4)所示：

$$a_i=\mathrm{softmax}(h_i^{c}\boldsymbol{W}_{\mathrm{atten}}V_i^{\mathrm{T}}) \tag{4}$$

其中，$\boldsymbol{W}_{\mathrm{attn}}$是双线性注意力的权重矩阵。式(5)为获得所有单词的加权和：

$$z_i^{w}=\sum_{i=1}^{m}a_{ij}v_{ij}^{w} \tag{5}$$

最后将经过加权的词汇信息输入字符向量中。

$$\overset{\wedge}{h}_i=h_i^{c}+z_i^{w} \tag{6}$$

## 4.2　特征提取层 BiLSTM

中医命名实体中常有病名与症状混合出现的情况，判断实体所属标签取决于前后文中的描述信息。但引入词汇信息后各个命名实体之间趋向于板块化，前后文信息减少，导致实体识别的准确率有所降低。故采用 BiLSTM 网络提取前后文中的特征信息，提高命名实体识别准确率。

如图 4 所示，LSTM 的神经元中引入了门函数，通过 3 个门结构(遗忘门重置数据、输入门读取数据、输出门输出数据)实现对前后文信息的取舍。

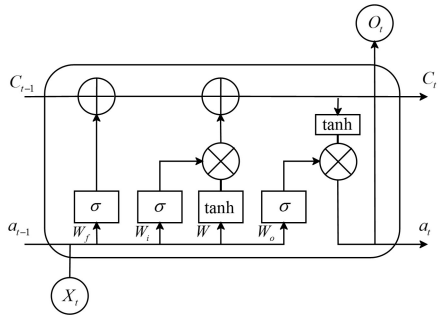



图 4　LSTM 的神经元

Fig. 4　Neurons of LSTM

每个神经元都有 3 个输入与 3 个输出，$X_t$是该时刻新加入的信息，$a_{t-1}$与$c_{t-1}$是上文信息的表示。$X_t$与$a_{t-1}$合并后克隆为 4 份，3 份分别与遗忘门、输入门和输出门的权重$W_f$，$W_i$，$W_o$进行计算，通过函数的激活获得遗忘权重、输入权重和输出权重；最后一份放入 RNN 中计算。但由于一个 LSTM 只能处理单向，为了更加准确地分析语句，将两个输入方向相反的 LSTM 模型组合为 BiLSTM 进行使用，如图 5 所示。

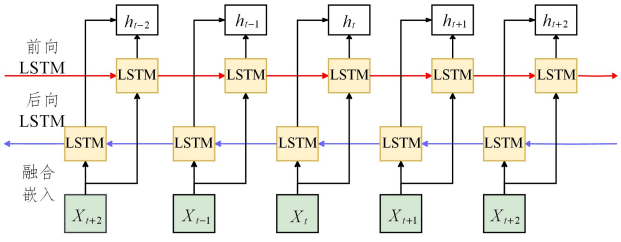



图 5　将 LSTM 组成 BiLSTM

Fig. 5　LSTM models are combined into BiLSTM

## 4.3　标签解码层 CRF

CRF 是一种基于统计的序列标注算法。CRF 将序列标注看作是一个$k^n$分类问题，把上下文的信息以及相邻标签的序列关系纳入考虑，获取全局信息进而获得全局最优的标记序列。

CRF 计算条件概率、得分函数、标签概率如下：

$$P(y_1\cdots y_n\mid x_1\cdots x_n)=P(y_1\cdots y_n\mid x),x=(x_1\cdots x_n) \tag{7}$$

$$s(X,y)=\sum_{i=0}^{n}A_{y_i,y_{i+i}}+\sum_{i=1}^{n}P_{i,y_i} \tag{8}$$

$$P(y\mid X)=\frac{\mathrm{e}^{s(X,y)}}{\sum\limits_{\tilde{y}\in Y_x}\mathrm{e}^{s(X,\tilde{y})}} \tag{9}$$

其中，$y_i$代表第 $i$ 个字符的标签，$x_i$代表第 $i$ 个字符；CRF 的转移矩阵为 $A$，$A_{ij}$代表标签$y_i$到$y_j$的转移概率；$y_0$和$y_n$分别代表句子的开始和结束；$Y_X$为所有标签概率。

# 5　实验内容

实验所使用的数据集为第 3 章构建的中医案例命名实体

数据集，共包含 800 个消化科的确诊病例数据。采用 5 折交叉验证对数据集进行划分，将 800 个数据分为 5 份，其中每份 160 个数据。将 3 份数据作为训练集，一份数据作为测试集，最后一份作为验证集，共测试 5 次取平均值。将原始训练集和测试集合并，与验证集混合，将数据集划分为 5 堆，从而减少数据集划分的随机性对模型性能的影响。

### 5.1 实验环境

实验环境为：Intel(R) Core(TM) i9-10900K CPU @ 3.70GHz 处理器，内存 RAM 为 32 GB，显卡 GPU 选用 NVIDIA RTX3090。本文编程语言版本为 Python3.8.1，深度学习框架为 Pytorch。

### 5.2 实验准备

1)模型构建来源：

我们使用的 BERT 模型是基于 Devlin 等[8]构建的，BERT 中共有 12 层 Transform，并使用 huggingface4 中的 bert-base-chinese 中相关模块进行初始化，使用 200dimension 预训练词嵌入，利用该词嵌入中的定向跳转图模型进行训练。在 BERT 模型的第一个 Transform 和第二个 Transform 之间插入词典适配器，将预训练得到的词嵌入信息连接词典适配器，在训练期间微调以实现 BERT 与词典适配器的融合。

2)超参数：

本文选择的是 Adam 优化器，在 BERT 中的 learning-rate 设置为 $1\times10^{-5}$，在 LEBERT 中由于有词典适配器传入数据，因此将 learning-rate 设置为 $1\times10^{-4}$。对于所有模型，设置的最大 epoch 为 20，最大序列长度为 256。

3)基准评估线：

为了证明 LEBERT-BILSTM-CRF(LBC)的优越性，引入其他模型进行比较。

4)用于对比的模型：

(1)BERT-CRF(BC)：BERT 系列最常见的训练模型。基于中文序列标记任务直接对预训练模型 BERT 进行微调，随后加入 CRF 层对结果进行筛选分类。

(2)BERT-BILSTM-CRF(BBC)：基于上述模型在 BERT 训练结果的基础上将 CRF 改为 BILSTM-CRF 进行对比。

(3)BERT-softmax(BS)：使用较常用的分类器 softmax，用于与 BC 和 BBC 号模型进行实验对照。

(4)BERT-BiLSTM-softmax(BBS)：使用 BiLSTM 与分类器 softmax 组合，用于与前述模型进行对比。

(5) ALBERT-CRF：ALBERT 常被看作是轻量级的 BERT[15]，对小型数据集可能有较好性能，同时迭代速度比 BERT 更快。

(6)ALBERT-BILSTM-CRF：ALBERT-CRF 号模型的扩展。

(7)LEBERT-CRF(LC)：与提出的模型作对比，验证加入 BILSTM-CRF 后的效果。

(8)LEBERT-softmax(LS)：同 LC 号模型。

(9)LEBERT-BiLSTM-softmax(LBS)：将 BiLSTM 与分类器 softmax 组合，观察效果。

5)评价标准：

本文使用 3 项重要的指标衡量模型的表现，分别是精确率 $P$、召回率 $R$ 和 $F1$ 值。

$$P=\frac{T_P}{T_P+F_P}\times100\% \tag{10}$$

$$R=\frac{T_P}{T_P+F_N}\times100\% \tag{11}$$

$$F1=\frac{2PR}{P+R}\times100\% \tag{12}$$

其中，$T_P$ 表示正确识别的实体数量；$F_P$ 表示错误识别的实体数量；$F_N$ 表示未识别出的实体数量。精确率 $P$ 表示所有预测为正样本的集合中预测正确的比例；召回率 $R$ 表示所有正样本中预测正确的比例；$F1$ 值综合精确率和召回率，是评估模型性能的综合指标。

### 5.3 对比实验

为了验证本文中 LEBERT-BiLSTM-CRF 模型的性能，将提出的模型与多个相似的同类型模型进行对比实验。对比实验中包含 BERT-Softmax，BERT-CRF，BERT-BILSTM-CRF，BERT-BILSTM-Softmax，ALBERT-CRF，ALBERT-BiLSTM-CRF 这些常见的命名实体识别模型。

如表 2 所列，本章所提的 LEBERT-BiLSTM-CRF 以 88.69%的准确率、87.4%召回率、88.1%的 F1 位列第一，其中准确率超过第二的 BERT-softmax 1.73%，F1 超过第二的 BERT-softmax 1.1%。

表 2 不同预训练模型的准确率、召回率和 F1

Table 2 Accuracy, recall and F1 of different pre-trained models

(%)

| 模型 | 准确率 | 召回率 | F1 |
|---|---|---|---|
| LEBERT-BiLSTM-CRF | **88.69** | **87.4** | **88.1** |
| BERT-BiLSTM-CRF | 85.78 | 87.2 | 86.4 |
| BERT-softmax | 86.96 | 87.1 | 87.0 |
| BERT-BILSTM-Softmax | 85.93 | 86.7 | 86.3 |
| BERT-CRF | 84.86 | **87.4** | 86.1 |
| ALBERT-CRF | 80.90 | 74.8 | 77.9 |
| ALBERT-BiLSTM-CRF | 80.68 | 83.7 | 82.1 |

表 2 中以 ALBERT 为基础的两个模型准确率明显低于其余几个模型，推测原因为中医的词汇大多较为偏僻晦涩，如“夜寐难安”这种文言文形式的症状标签；另一方面，这两个模型在十分依赖前后文信息的“诱因”标签上准确率非常低。ALBERT 作为小型的 BERT，在只有少量训练数据的情况下并未完全收集到相关的词汇信息特征，导致准确率较其余模型偏低。

对比 BERT-softmax，BERT-BiLSTM-CRF，BERT-CRF，BERT-BiLSTM-softmax 这 4 个模型在 10 个标签上的准确率。如图 6 所示，softmax 的准确率均略高于 CRF 和 BILSTM-CRF。本次用于训练的数据较少，在仅使用 BERT 模型的情况下，模型吸收的前后信息较为单一和固定，即使增加了 BiLSTM 层纳入前后文的信息也难以使得后续 CRF 分类器做出对数据更合理的分类。相反，较少考虑前后文信息的 softmax 能在这种情况下做出更好的判断。例如在性别上，softmax 和 BiLSTM-softmax 正确率达到百分百，而 BILSMT-CRF 和 CRF 达不到百分百。

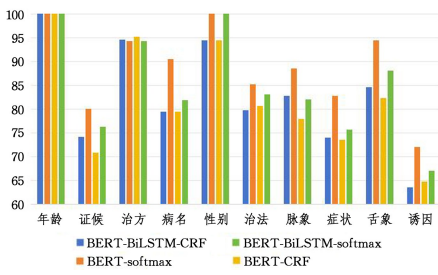



图 6 BERT 系列模型在 10 个命名实体的准确率

Fig. 6 Recognition accuracy of BERT series model on ten named entities

### 5.4 消融实验

为证明融合后的模型具有更好的性能，将 LEBERT-BiLSTM-CRF 拆分组合为 LEBERT-CRF，LEBERT-softmax，LEBERT-BiLSTM-softmax，将 4 个模型的结果进行对比。如表 3 所列，LEBERT-BiLSTM-CRF 获得了最高的准确率和 F1。

表 3 4 个模型准确率、召回率以及 F1

Table 3 Accuracy rate, recall rate and F1 of four models

（%）

| 模型 | 准确率 | 召回率 | F1 |
|---|---|---|---|
| LEBERT-BiLSTM-CRF | **88.69** | 87.4 | **88.1** |
| LEBERT-CRF | 87.31 | **87.7** | 87.5 |
| LEBERT-softmax | 87.91 | 86.8 | 87.3 |
| LEBERT-BiLSTM-softmax | 88.12 | 87.32 | 87.7 |

如图 7 所示，LEBERT-BiLSTM-CRF 模型在 10 个命名实体中都有优秀的表现。纳入了词汇增强能力后的 BERT 在面对上一个模型出现的问题时，例如“性别”标签，使用 CRF 能获得更高的准确率。在性别、年龄、症状、病名、证候，以及综合准确率上，相比另外两个模型，LEBERT-BiLSTM-CRF 有更高的准确率。

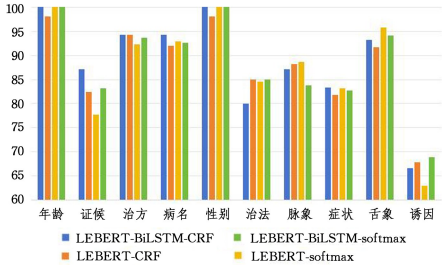



图 7 4 个模型在 10 个命名实体识别上的准确率比较

Fig. 7 Comparison of recognition accuracy of four models on ten named entities

模型总体对比结果如表 4 所列。

（1）在年龄和性别的标注序列中，参与统计的模型都有较好的效果。对于这些单一且易分辨的命名实体，在缺少 BiLSTM 提取的前后文信息或是不使用词典适配器融入词汇信息时获取的信息较为单一，CRF 的判断能力会略微弱于 softmax 分类器，CRF 在这种情况下表现较差。

（2）在中医案例最重要的证候序列标记中，LEBERT-BiLSTM-CRF 取得了最高的准确率，为 87.1%，在症状病名上同样取得了最高的准确率，在其余的几个命名实体中也有着较高准确率。与其他模型相比，LEBERT-BiLSTM-CRF 的优势在于不仅保证了前后文信息供给，而且加入了词汇信息，依靠两个信息获取途径进一步优化结果，给模型带来了正向反馈效果。

基于上述对比实验可知，加入 BiLSTM 层后 LEBERT-BiLSTM-CRF 获得了高于 BERT-CRF 和 LEBERT-CRF 的准确率。LEBERT-BiLSTM-CRF 模型在性别和年龄的识别上达到了百分百的准确率，在症状、病名、证候方面的准确率分别为 83.33%，94.23%，87.10%，综合准确率为 88.69%，与其他模型相比，LEBERT-BiLSTM-CRF 这 6 项拥有最高的准确率。

表 4 不同模型在 10 个命名实体识别上准确率对比结果

Table 4 Comparison results of recognition accuracy of different models on ten named entities

（%）

| 标签 | LBC | BBC | BS | BC | BBS | LC | LS | LBS |
|---|---|---|---|---|---|---|---|---|
| 年龄 | **100** | **100** | **100** | **100** | **100** | 98.00 | **100** | **100** |
| 证候 | **87.10** | 74.15 | 80.00 | 70.75 | 76.23 | 82.35 | 77.78 | 83.12 |
| 治方 | 94.29 | 94.61 | 94.29 | **95.24** | 94.32 | 94.24 | 92.31 | 93.62 |
| 病名 | **94.23** | 79.34 | 90.52 | 79.35 | 81.76 | 92.04 | 92.92 | 92.61 |
| 性别 | **100** | 94.37 | **100** | 94.37 | **100** | 98.00 | **100** | **100** |
| 治法 | 80.00 | 79.78 | **85.19** | 80.67 | 83.06 | 85.06 | 84.62 | 84.98 |
| 脉象 | 87.10 | 82.76 | 88.57 | 77.87 | 81.96 | 88.24 | **88.57** | 83.78 |
| 症状 | **83.33** | 73.90 | 82.78 | 73.53 | 75.63 | 81.77 | 83.19 | 82.71 |
| 舌象 | 93.15 | 84.58 | 94.37 | 82.30 | 88.12 | 91.67 | **95.71** | 94.18 |
| 诱因 | 66.67 | 63.53 | **72.00** | 64.63 | 66.96 | 67.86 | 62.96 | 68.92 |

### 5.5 实验时间对比

如表 5 所列，LEBERT-BiLSTM-CRF 与其他模型相比所耗时间虽略有增加，但所耗时间在合理范围内（比平均时间长 8%）。

表 5 不同模型的训练时间

Table 5 Training time of different models

（min）

| 模型 | 所用时间 |
|---|---|
| LEBERT-BiLSTM-CRF | 103 |
| LEBERT-BiLSTM-softmax | 95 |
| LEBERT-softmax | 92 |
| LEBERT-CRF | 94 |
| BERT-BiLSTM-CRF | 99 |
| BERT-BiLSTM-softmax | 95 |
| BERT-softmax | 87 |
| BERT-CRF | 92 |

3）部分验证集的识别结果

如表 6、表 7 所列，选取一部分验证集中的数据，利用 LEBERT-BiLSTM-CRF 模型对其进行命名实体识别。

表 6 部分验证集识别数据

Table 6 Partial validation set identification data

| 字 | 胰 | 腺 | 炎 | ， | 男 | 3 | 5 | 岁 |
|---|---|---|---|---|---|---|---|---|
| 原标签 | B-disease | I-disease | I-disease | 0 | B-gender | B-age | I-age | I-age |
| 预测标签 | B-disease | I-disease | I-disease | 0 | B-gender | B-age | I-age | I-age |

表 7 部分验证集识别数据

Table 7 Partial validation set identification data

| 字 | 饮 | 酒 | 带 | 醉 | 入 | 睡 |
|---|---|---|---|---|---|---|
| 原标签 | B-triggers | I-triggers | B-triggers | I-triggers | I-triggers | I-triggers |
| 预测标签 | B-triggers | I-triggers | B-triggers | I-triggers | I-triggers | I-triggers |

表中，“字”表示验证集中病例的文字信息，“原标签”表示人工标注后该字所属标签，“预测标签”表示模型对文字信息进行命名实体识别后该字符的预测标签。由表中信息可得，该模型已经可以较为准确地预测一段文字中

每个文字对应的标签。

**结束语** 针对中医案例领域的命名实体识别任务中命名实体密集、杂乱无章、词义相近但又分属不同的问题，提出了一种融合词汇增强与预模型的中医命名实体识别模型，该模型在注意力机制中融入词汇信息，再融入 BiLSTM 进一步获取前后文信息后，能得到更准确的分类结果。CRF 经测试数据验证，取得了 88.69%的识别准确率、85.8%的识别召回率及 87.1%的 F1 值。该模型在我们创建的小型医疗病例数据集上获得了较好的结果，年龄、性别、症状、病名、证候的识别率以及综合准确率均略高于使用 LEBERT 搭配其他分类器和使用 BERT 搭配其他分类器的准确率，说明了该模型在中医案例命名实体识别方面具有一定的实用性。

但与上万条数据量相比，部分实体类别的训练数量不足、数据覆盖面小，从而对整体识别水平有一定的影响，如“诱因”这一标签，数据量较少，部分病例中甚至没有，导致所有模型对于该标签的预测准确率都偏低，后续研究中，将获取更多的医案数据，扩大可识别的实体范围，从而提高准确率和可用性。

## 参 考 文 献


[1] JI T, SU S L, SHANG E X, et al. Determining the rules of traditional Chinese medicine on treatment of consumptive thirst based on association rules mining[J]. China Journal of Traditional Chinese Medicine and Pharmacy, 2016, 31(12): 4982-4986.

[2] XU Z H. Statistical Model based Chinese Named Entity Recognition Methods and its Application to Medical Records[D]. Beijing: Beijing University of Chemical Technology, 2017.

[3] GAO J Y, LIU Z, YANG T, et al. Research on Named Entity Extraction of TCM Clinical Medical Records Symptoms Based on Conditional Random Field[J]. Modernization of Traditional Chinese Medicine and Materia Medica-World Science and Technology, 2020, 22(6): 1947-1954.

[4] ZHAO Z H, YANG Z H, LUO L, et al. Disease named entity recognition from biomedical literature using a novel convolutional neural network [J]. BMC Medical Genomics, 2017, 10(S5): 73.

[5] CAO C P, GUAN J P. Clinical text named entity recognition based on E-CNN and BLSTM-CRF[J]. Application Research of Computers, 2019, 36(12): 3748-3751.

[6] GUO X R, LUO P, WANG W L. Chinese named entity recognition based on Transformer encoder[J]. Journal of Jilin University(Engineering and Technology Edition), 2021, 51(3): 989-995.

[7] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [J]. arXiv: 1810.04805, 2018.

[8] YAN H, DENG B, LI X, et al. TENER: Adapting Transformer Encoder for Named Entity Recognition[J]. arXiv: 1911.04474, 2019.

[9] LASRI K, LENCI A, POIBEAU T. Does BERT really agree? Fine-grained Analysis of Lexical Dependence on a Syntactic Task[J]. arXiv: 2204.06889, 2022.

[10] MA R, PENG M, ZHANG Q, et al. Simplify the Usage of Lexicon in Chinese NER[C]// Proceedings of the 58th ANNUAL Meeting of the Association for Computational Linguistics. 2020: 5951-5960.

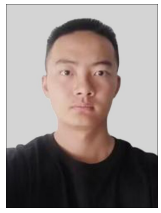
[11] LIU W, FU X, ZHANG Y, et al. Lexicon Enhanced Chinese Sequence Labeling Using BERT Adapter[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021: 5847-5858.

[12] REN Y, YU H, YANG H, et al. Recognition of quantitative indicator of fishery standard using attention mechanism and the BERT+BiLSTM+CRF model[J]. Transactions of the Chinese Society of Agricultural Engineering, 2021, 37(10): 135-141.

[13] LIU J G, XIA C H. Innovative deep neural network modeling for fine-grained Chinese entity recognition[J]. Electronics, 2020, 9(6): 1001.

[14] LIAO X F, XIE S S. Chinese Named Entity Recognition Based on Attention Mechanism Feature Fusion[J]. Computer Engineering, 2023, 49(4): 256-262.

[15] LAN Z, CHEN M, GOODMAN S, et al. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations[J]. arXiv: 1909.11942, 2019.

**LI Minzhe**, bornin 1997, postgraduate. His main research interests include deep learning and natural language processing.

**YIN Jibin**, born in 1976, Ph.D, associate professor. His main research interests include human-computer interaction and artificial intelligence.