[文章编号] 1003－4684(2025)01-0055-06

基于 RoBERTa 的中医药专利命名实体识别

邓　娜，何昕洋，熊才权，宗泽华
（湖北工业大学计算机学院，湖北 武汉 430068）

[摘　要] 中医药发明专利的成分及功能实体具有种类复杂、歧义繁多等特点。针对传统命名实体识别方法无法充分获取其中的语义特征表示，上下文信息及一词多义等问题，提出一种中医药发明专利命名实体识别模型，该模型将 RoBERTa-WWM 预训练模型、双向长短期记忆(BiLSTM)网络、条件随机场(CRF)三个模块串联结合，将专利摘要依次通过 RoBERTa-WWM 进行语义提取生成含有先验知识的语义词嵌入；BiLSTM 网络增强词嵌入中的上下文特征信息；CRF 解码序列，输出概率最大结果。实验结果表明，该模型在真实中医药发明专利文本语料库上，F1 值(F-Measure)相较其他主流方法在成分与功能的识别上分别提升了 5.80％和 6.63％，能有效提升中医药发明专利摘要中药物成分及功能识别的准确率。

[关键词] 中医药发明专利；命名实体识别；RoBERTa-WWM；BiLSTM
[中图分类号] TP312　　[文献标识码] A

中医药发明专利作为中药知识产权的核心要素，不仅是企业争相掌握的重要资源，更是能促进中医药学成果普及、推动中医药科学创新的重要动力。摘要是中医药发明专利的重要组成部分。从撰写内容看，它简要概括了中医药发明专利的成分、功效和目的等要素，读者能够根据摘要快速了解专利内所介绍的发明信息。从组织结构看，它是中医药发明专利中不可或缺的部分。读者在进行专利检索时，一方面，大多数中医药发明专利资源库面向整段摘要提供检索时，无法对摘要中所提到的“成分”、“功能”等结构进行单独检索；另一方面，读者获取中医药发明专利中成分及功能信息通常通过阅读整段摘要实现，这样的做法费时费力。随着大数据时代的到来以及人工智能技术的兴起，如何从中医药发明专利的摘要中识别出需求信息，成为了当前专利研究关注的一大热点问题。而其中关键一环则是摘要结构中成分及功能的命名实体识别。样例识别如图 1 所示。

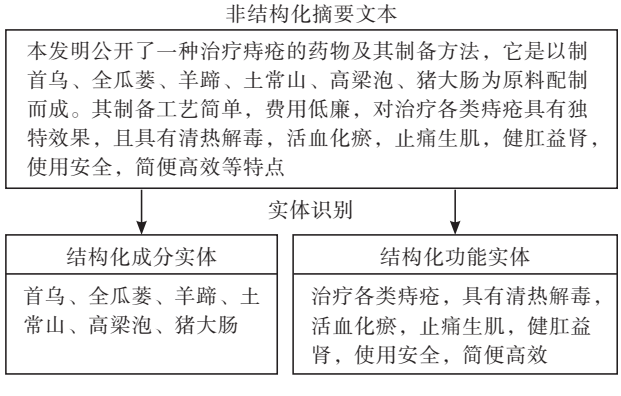


图 1　中医药专利实体识别样例

命名实体识别作为自然语言处理领域经典问题之一，解决问题的目标是从非结构化文本中识别出结构化的目标命名实体。中药发明专利的命名实体识别任务对比于传统医药领域的命名实体识别任务面临更大的挑战，主要原因在于：第一，相比于传统领域，中医药领域的文本晦涩难懂，拥有更多不同类目的未登录核心词，传统的命名实体识别方法难以对其进行有效识别。第二，传统领域用词规范严谨，歧义切分现象相对较少，而中药领域的专利发明人为了达到专利审查所要求的创新性，同时出于对自身技术保护的考虑，会使用更加抽象的地方口语词甚至自定义词。因此，中药发明专利中包含大量未登录词且基本都为专业术语，严重影响命名实体识别的准确性。鉴于此，本文提出新的模型来解决以上问题，同时引入几种典型的深度学习模型，设计了一系列摘要功能和成分的识别实验，并对这些模型的性能表现进行对比分析。

1　相关工作

当前中医药专利的命名识别方法主要分为 3 种：基于中医药关键词的词频统计方法、基于自定义

[收稿日期] 2022－08－22
[第一作者] 邓　娜(1985－)，女，湖北武汉人，湖北工业大学副教授，研究方向为专利分析和机器学习。
[通信作者] 熊才权(1966－)，男，湖北鄂州人，工学博士，湖北工业大学教授，研究方向为人工智能、辩论模型、智能决策。

识别规则的机器学习方法与基于预训练模型的深度学习方法。早期医药命名实体识别研究大多基于规则词频统计方法,制定特定规则对信息源内的目标关键词出现次数进行统计。如 Song 等[1]通过构建医药字典对生物医药名称进行识别,但这种针对特定任务设计的字典或准则的针对性过强,无法满足其他实体识别任务的需求,因而普遍缺乏泛化性且可移植性较差。传统的机器学习方法通常需要专业人员选择合适的特征来训练模型,Leaman 等[2]提出 tmchem 模型,该模型将两个 CRF 模型以并联形式进行组合,对多个人工提取的特征进行融合,从而对生物医药中的专有名词进行命名实体识别。Li 等[3]将词频信息和共现词表信息融入 CRF 模型,进一步提升模型对医学专有名词识别的准确性。相比于早期命名实体识别研究,上述基于机器学习的方法在命名实体识别上的泛化性和可移植性有了一定改善,但对识别对象的特征选取过于依赖专家的专业知识,且识别方法单一,因而方法与运用范围有限。近年来,以 Word2vec 词嵌入方法为基础的深度学习受到学者们的广泛关注,并将其运用于各个命名实体识别任务。Kim Y 等[4]提出的 TextCNN (text convolutional neural networks)模型,众多学者[5-8]提出 BiLSTM-CRF 模型等,相比于传统机器学习方法,深度学习方法通过不同的神经网络自动学习并更新隐藏在目标源中的特征信息,在不需要依赖专家对特征进行人工构建的前提下,在各个命名实体识别任务中也能取得不错的效果。但由于缺乏专家对实体特征提取的监督与筛选,上述深度学习方法的效果很大程度上受限于训练集的质量,因而人们提出从大量无标签文本中获取先验语义知识来增强语义表示的 BERT(bidirectionalencoder representations fromtransformer)预训练模型。

预训练模型可以从自定义语料库中学习到含有先验语义的词嵌入表示,从而有效缓解因训练集质量不足导致词嵌入表示信息获取不充分问题。例如 2018 年谷歌发布的论文[9]中提到的 BERT 模型,通过海量语料进行预训练,得到包含目标序列充分的局部和全局特征词嵌入表示。沈同平[10]、王子牛等[11]将 BERT 与 BiLSTM-CRF 相结合应用于中文命名实体识别,对比实验证明:BERT 预训练模型相比于静态词嵌入在命名实体识别任务效果更好。BERT 基于 12 层 Transformer 中的 Encoder 部分连接组成[12],通过两种预训练任务 MLM(Masked Language Model)和 NSP(Next Sentence Prediction,)从无标签的非结构化语料内学习先验的中医药语义信息。BERT 预训练模型在进行 MLM 任务时会像处理英文单词一样随机掩盖训练样本的字,但中文语料不同于英文语料,中文语料序列的语义单元组成是词而非字,即 BERT 的 MLM 任务无法在预训练过程中获取词的信息。因此,为使 BERT 更好适应并提取中文语料的特征信息,哈工大讯飞联合实验室发布中文 RoBERTa-WWM 训练模型。该模型在进行先验语义的预训练时会以句子为单位进行并训练,这样更能保留训练目标句法层次的先验知识。在进行分词任务时,以词组为单位进行切分,大幅度降低了没登录词所带来的特征缺失问题。

2 相关模型

2.1 RoBERTa-WWM 预训练模型

RoBERTa-WWM 模型在保留 BERT 模型优点的基础上,进一步对预训练语料与模型的结构层次进行优化[13]。在大规模中文语料的基础上,使用更大的单次训练样本数和更多的训练次数。同时移除 NSP 任务,以减少训练成本。相比于传统 BERT 模型,RoBERTa-WWM 模型把输入文本转化为语义更为丰富的词嵌入表示,并为其添加句子特征与位置特征,以便更好地区分上下文信息,用更长的序列长度进行训练,最大程度保留上下文信息。

2.2 BiLSTM 层

该层由双向 LSTM 层前后拼接组成,能通过不同方向的梯度传递学习到更加全面的中医药专利语句的上下文特征信息。将 RoBERTa-WWM 层输出的词向量同时输入到双向 LSTM 模型进行学习,以此来强化词嵌入表示的上下文信息。输出的特征信息,使用 Softmax 函数将所获得的特征值转为对应中医药实体的概率分布。LSTM 层的模型结构如式(1)~(5)所示:

$$i_t = \sigma(x_t \cdot w_i + h_{t-1} \cdot w_i' + b_i) \tag{1}$$

$$f_t = \sigma(x_t \cdot w_f + h_{t-1} \cdot w_f' + b_f) \tag{2}$$

$$o_t = \sigma(x_t \cdot w_o + h_{t-1} \cdot w_o' + b_o) \tag{3}$$

$$\tilde{c}_t = \tanh(x_t \cdot w_c + h_{t-1} \cdot w_c' + b_o) \tag{4}$$

$$c_t = i \otimes \tilde{c}_t + f_t \otimes c_{t-1} \tag{5}$$

$$h_t = o_t \otimes \tanh(c_t) \tag{6}$$

其中,σ 为 sigmoid 函数,$\otimes$ 是点积运算,$i_t = \sigma(x_t \cdot w_i + h_{t-1} \cdot w_i' + b_i)$、$f_t = \sigma(x_t \cdot w_f + h_{t-1} \cdot w_f' + b_f)$、$o_t = \sigma(x_t \cdot w_o + h_{t-1} \cdot w_o' + b_o)$ 和 $\tilde{c}_t$ 分别表示 LSTM 模型的输入门、遗忘门、输出门和记忆单元;w_i,h_{i-1},b_i 代表输入门、遗忘门和输出门的权重矩阵和偏置项;h_t 为 t 时刻的网络输入也对应于词表示层的输出。

2.3 CRF 层

Softmax 函数虽然可以输出每个中医药实体对

应的概率标签,但无法保证标签之间的逻辑关系,CRF 层则可以对标签的结构进行规范控制。例如在 I-function 后才能出现 B-function。通过 BiLSTM 层输出的文本特征向量标签无法对其逻辑依赖进行约束。因此,本文添加 CRF 层对 BiLSTM 网络层的输出进行修正,将放入 BiLSTM 层的输出放入语言模型中使其得到包含上下文逻辑依赖的合理最优序列。BiLSTM 层的输出作为 CRF 层的输入,输出得分 score (X,y) 如式(7)所示:

$$\text{score}(X,y)=\sum_{i=0}^{n}A_{y_i,y_{i+1}}+\sum_{i=1}^{n}P_{i,y_i} \tag{7}$$

其中,转移矩阵元素 $A_{y_i,y_{i+1}}$ 表示标签从 y_i 转移到 y_{i+1} 的概率,y_i 为 y 中的元素。根据相同的 X,输出预测结果 y 的概率和训练过程中的似然函数公式以及输出最高得分的结果公式如式(8)～(10)所示。

$$p(y\mid X)=\frac{e^{\text{score}(X,y)}}{\sum_{\tilde{y}\in Y_x}\text{score}(X,y\sim)} \tag{8}$$

$$\log(p(y\mid X))=\text{score}(X,y)-\log(\sum_{\tilde{y}\in Y_x}e^{\text{score}(X,y\sim)}) \tag{9}$$

$$y^{*}=\arg_{y\in Y_x}\max\text{score}(X,y\sim) \tag{10}$$

3　本文方法

本文提出一种基于 RoBERTa-WWM 的中医药发明专利命名实体识别模型 RoBERTa-WWM-BiLSTM-CRF,该模型采取 RoBERTa-WWM 为预训练语言模型,利用 BiLSTM 网络增强词嵌入中的上下文特征信息,使用 CRF 解码序列,对中医药发明专利的命名实体进行识别。RoBERTa-WWM-BiLSTM-CRF 模型整体结构如图 2 所示。

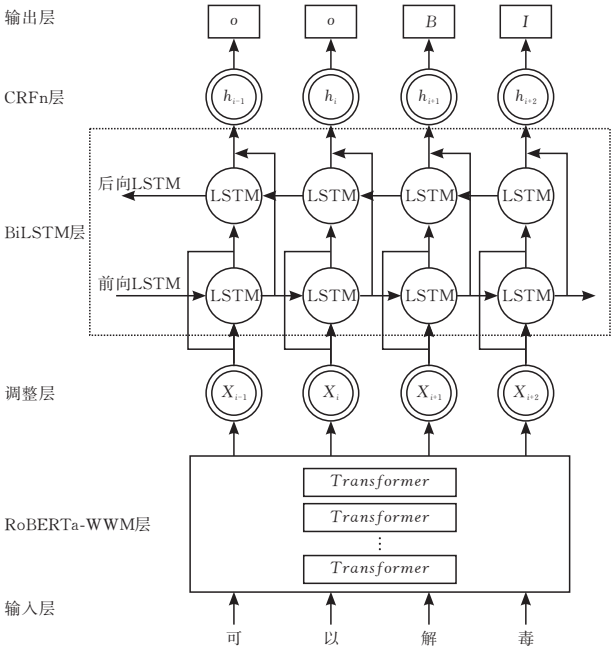


图 2　RoBERTa-WWM-BiLSTM-CRF 模型的总体结构

由图 2 可见,该模型由 3 部分组成,分别为 RoBERTa-WWM 预训练层、BiLSTM 神经网络层和 CRF 分类层。首先,将中医药专利摘要文本输入 RoBERTa-WWM 预训练模型,以词为单位进行切分,并将其转换成词组序列输入 Transformer 的 Encoder 层,通过多头注意力机制,提取摘要文本中的语义信息,达到一词多义的效果;其次,将最后一层 Transformer 的 Encoder 层输出词嵌入输入 BiLSTM 模型记录文本序列上的依赖关系,并强化其中的上下文特征信息;最后,传递到 CRF 层获取概率最大的标签信息,完成对中药专利中功能与成分的抽取。本文使用该模型对中医药发明专利中的成分及功能的进行识别提取,步骤如下:

1)对原始中医药发明专利的摘要进行数据预处理,将摘要文本表示为 $Q=\{q_1,q_2,q_3,\cdots,q_n\}$,其中,$n$ 为中医药专利摘要文本数量,第 i 个中医药专利摘要文本表示为 $q_i=\langle w_{i1},w_{i2},w_{i3},\cdots,w_{in}\rangle$,该向量中每个元素为中医药专利摘要中的词语,预先定义的中医药成分与功能类别表示为 $C_{\text{Componet}}=\{c_1,c_2,c_3,\cdots,c_x\}$ 和 $F_{\text{function}}=\{f_1,f_2,f_3,\cdots,f_y\}$,其中 x,y 为预先定义的中医药成分数及功能数。该类别通过词组级别的划分对中医药专利摘要中的成分及功能文本进行标注,预定义类别与普通文本类别之间使用空格分割。

2)使用哈工大 LTP 作为分词工具,输入的中医药专利摘要文本 $Q=\{q_1,q_2,q_3,\cdots,q_n\}$ 切分并输出为词组集合 $T=\{t_1,t_2,t_3,\cdots,t_n\}$,结合预训练中每个词组的先验知识,RoBERTa-WWM 模型根据学习到的先验知识自动进行微调,并更新模型参数,最终输出目标单词的最终词向量 $\boldsymbol{H}_C=\{h_1,h_2,h_3,\cdots,h_x\}$ 和 $\boldsymbol{H}_F=\{h_1,h_2,h_3,\cdots,h_y\}$。

3)将 RoBERTa-WWM 模型输出的最终词向量 $\boldsymbol{H}=\{h_1,h_2,h_3,\cdots,h_{x||y}\}$ 输入 BiLSTM 模型中,根据公式(1)－(6)对部分无用特征进行过滤清洗,并通过前后向的梯度传播获取增强上下文语义信息的隐藏状态序列 $S_c=\{s_1,s_2,s_3,\cdots,s_x\}$ 和 $S_F=\{s_1,s_2,s_3,\cdots,s_y\}$。

4)将 BiLSTM 层输出的隐藏序列 $S=\{s_1,s_2,s_3,\cdots,s_{x||y}\}$,根据上下文信息对每个特征打上标签并输入 CRF 层,通过 CRF 层增加标签间的约束关系,根据 CRF 模型中的回归逻辑层输出最优的标注序列 $y_c=\{y_1,y_2,y_3,\cdots,y_x\}$ 与 $y_f=\{y_1,y_2,y_3,\cdots,y_y\}$ 最后根据公式(7)－(10)获得每个序列的最终预测分数,并将其解码为对应的最大概率标签。

5)对预处理的中医药专利摘要数据集采用 8∶2 的比例进行训练集、测试集的划分。

6)对训练集进行模型训练,将训练好的模型保持参数不变,对测试集数据直接使用训练好的参数进行验证,依据不同的评测标准获得相应的识别结果。具体流程可见图 3。

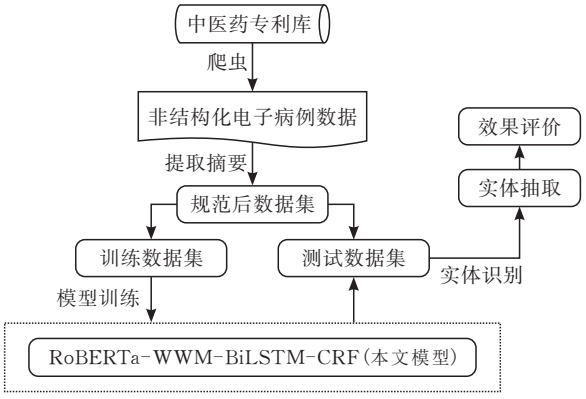


图 3 中药发明专利成分与功能的实体识别流程

4 实验结果与分析

4.1 数据获取与实验流程

本文通过爬虫技术获取了中国专利信息网上的 7000 条中医药发明专利数据。每条专利数据包含实验所需的专利名称、申请号以及摘要部分。对所爬取的非结构化摘要部分进行数据预处理、成分及功能的数据标注以及标注一致性检验的操作,构建出中药发明专利数据集。由于部分企业及个人考虑到中药发明专利的核心成分被泄露,因而对其在成分上进行一定的隐藏;所以数据集中的成分样本数量少于功效样本数量。将上述实验数据进行预处理后将训练集与测试集按照 8∶2 进行划分。同时采用 BIO 标注方式,一共标注出 5 类:O,B-Component,I-Component,B-Function,I-Function。其中 O 代表中医药非实体,B 代表中医药实体的词首,I 代表中医药实体的词身。

4.2 实验设置

本文的实验模型使用 Pytorch 深度学习框架构建。同时本实验使用试错法寻求到最优的超参数设置。表 1 为本文模型所涉及的超参数值。

表 1 实验可调参数设置

Super Parameter	Description	Select Value	Value
Droppout	Neuron ratio to prevent overfitting	[0.3,0.5]	0.5
Epoch	Number of iterations	[30,50,80,100]	80
Batch_Size	Number of samples selected for one training	[32,64,128]	128
Learning Rate	Learning rate	[2E-5,4E-5,5E-5]	4E-5
MaxLen	Maximum length of unit statement	[64,128]	128

4.3 模型评估

本文的模型评测指标采用准确率(Precision)、召回率(Recall)和 F1 值。其中,精确率指的是预测为正同时实际也为正的中医药实体概率。召回率是针对原中医药样本中,实际为正的中医药实体样本占被预测为正的中医药实体样本中的概率。

F1 值是召回率与精确率的调和平均值。计算公式为:

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}},R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}},F1=\frac{2P*R}{P+R}$$

其中,TP 指的是原中医药实体标签为正,实际预测也为正确的样本;TP+FP 指所有预测正确的中医药实体样本;TP+FN 指实际为正的中医药实体样本总和。

4.4 对比实验分析

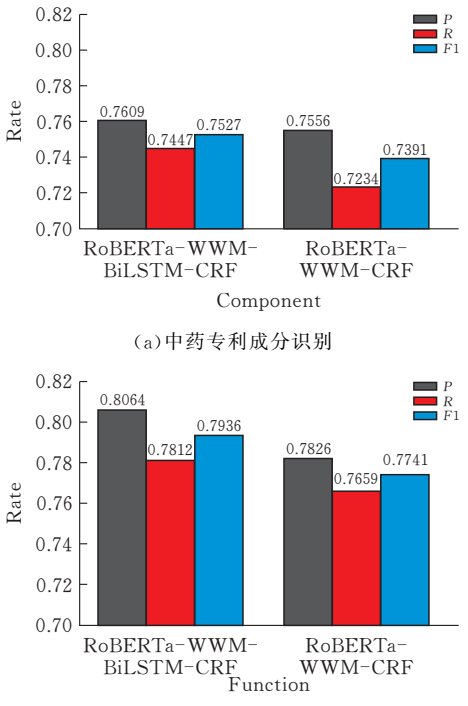
为了验证本模型相较于其他主流模型在中药发明专利中有更好的成分及功能的识别效果,本文设计了两类对比试验:第一类实验采用当前主流的不同的预训练模型与本模型进行对比实验设计;第二类实验用同一预训练模型的不同下游模型进行对比试验设计。具体结果如下所示。

4.4.1 与其他主流模型对比与分析实验 表 2 展示了第一类实验在中药发明专利数据集上的对比结果。由于非专业人工标注的不稳定性与上述中药专利数据集自身的缺陷性,所有模型的各项评价指标数据相比较于公开数据集偏低,同理中药专利内成分的识别效果也比功效的识别效果低。

表 2 不同模型实验结果

Model	Function			Component		
	P	*R*	*F*	*P*	*R*	*F*
TextCNN-CRF	0.6486	0.6153	0.6316	0.6222	0.6021	0.5833
BiLSTM-CRF	0.6757	0.6410	0.6576	0.6304	0.6041	0.6170
TextCNN-BiLSTM-CRF	0.7027	0.6667	0.6842	0.6596	0.6458	0.6526
BERT-BiLSTM-CRF	0.7368	0.7179	0.7272	0.7021	0.6875	0.6947
RoBERTa-WWM-BiLSTM-CRF	0.8064	0.7812	0.7936	0.7609	0.7447	0.7527

分析表 2 可以看出,本模型在中药专利数据集的成分与功能识别上对比于其他基线模型取得了最优的识别效果。对于基于 Word2vec 词嵌入的基线模型 TextCNN-CRF 模型、BiLSTM-CRF 模型以及 TextCNN-BiLSTM-CRF 模型由于其词嵌入无法准确的获取中医药专利中的多语义特征,对比基于 BERT 的中文预训练模型作为词嵌入层,识别结果并不理想。参照表中的准确率、召回率和 F1 值可知,BERT-BiLSTM-CRF 模型相比于 TextCNN-BiLSTM-CRF 模型在功能和成分的识别方面分别提升了 3.41%,5.12%,4.3% 和 4.25%,4.17%,4.21%。分析原因是由于 word2vec 产生是静态的,不考虑上下文的词表示。无法体现词的复杂特性,包括一词多义、语法、语义等信息。而本文模型相比 BERT-BiLSTM-CRF 模型,在功能和成分的识别方面,准确率、召回率和 F1 值分别提升了 6.96%,6.64%,6.33% 和 5.88%,5.72%,5.80%。分析原因主要是由于相较于 BERT 模型采用的语义提取单一是字而非词,对于中医药文本的语义信息提取大打折扣。本文使用的 RoBERTa-wwm 预训练模型可根据全词掩码任务获取词级别的语义表示,提取到的语义特征能更好的融合上下文的信息,模型更加适用于包含不规则语句的中医药文本。综上所述,本模型相较于其他不同预训练模型,更能适用于中药发明专利中所包含的功能和成分的命名实体识别任务。



(a)中药专利成分识别

(b)中药专利功能识别

图 4　不同下游模型对比实验结果

4.4.2　同一预训练模型的单变量性能对比实验

为了对模型进行更细微的对比,将本实验模型与其作为基准的 RoBERTa-WWM-CRF 模型分别在中药发明专利中的功能和成分的识别上进行对比。评价指标同样使用准确率、召回率和 F1 值,结果如图 4 所示。

分析图 4 可知,本模型与 RoBERTa-WWM-CRF 模型都在中药专利的功能识别方面取得了较好的效果,但在成分识别方面相对较差,分析原因是由于中药专利的功能表述样本数量相对较大,模型对功能的语义表征学习较为全面。而成分表述样本数量相对较少,且包含有些未登录中药专有名词,这使一部分成分特征不是很明显,导致识别错误。本模型对比于 RoBERTa-WWM-CRF 模型的准确率、召回率和 F1 值,在功能和成分的识别方面分别提高了 2.38%,1.53%,1.95% 和 0.53%,2.13%,1.36%。说明引入 BiLSTM 层可以对中药发明专利中不规则长度的成分和功能特征进行更好的捕获。总的来说,本模型比 RoBERTa-WWM-CRF 模型在中药发明专利的两个类别识别上都具有更好的效果。

4.4.3　消融实验　为了验证本文模型中,各个模块之间的有效性,本文通过删除完整框架的单个模块设计了三个辩题实验。控制训练集与测试集与基线模型对比实验相同,性能指标采用准确率、召回率和 F1 值作为评价标准,实验结果如表 3 所示。

表 3　消融实验结果

Model	Function			Component		
	P	*R*	*F*	*P*	*R*	*F*
消融实验 1	0.6757	0.6410	0.6576	0.6304	0.6041	0.6170
消融实验 2	0.7863	0.7324	0.7633	0.7552	0.7326	0.7352
消融实验 3	0.7535	0.7223	0.7324	0.7322	0.7145	0.7232
本文模型	0.8064	0.7812	0.7936	0.7609	0.7447	0.7527

消融实验 1:移除本文模型中的预训练模型 RoBERTa-WWM 模块,使用 BiLSTM-CRF 模型进行实验。

消融实验 2:移除本文模型中的 BiLSTM 模块,使用基于预训练模型的 RoBERTa-CRF 进行实验。

消融实验 3:移除本文模型中的 CRF 模块,将

RoBERTa-BiLSTM 的输出序列直接进行标签预测。

如表 3 所示，在中医药成分和功能的识别上，消融实验 1 性能都是最差，而本文模型性能最好。以中医药专利的功能识别为例，消融实验 1 对比本文模型在准确率、召回率和 F1 值上分别减少了 13.07%、14.02%和 13.60%，分析原因在于消融实验 1 缺少中医药预训练语料的先验知识，从而导致在对中药成分识别前，无法有效获取到有效的特征词向量。而本文模型相比于消融实验 2 与消融实验 3，在准确率、召回率和 F1 值上也有相应的提升，但消融实验 2 中由于缺少 BiLSTM 模块，因此对长句中的上下文信息提取能力较弱，从而导致识别性能略低于本文模型；消融实验 3 缺乏 CRF 层对最后的标签进行语法与逻辑层次的规范校准，因而实验结果与本文模型也有差距。从消融实验的结果上看，本文模型中所用到的预训练模型 RoBERTa-WWM，BiLSTM 模型，CRF 模型能有效的提升中医药专利成分及功能命名实体识别的准确度。

5 结论

本文提出结合预训练模型 RoBERTa-WWM 和 BiLSTM-CRF 应用于中药发明专利的命名实体识别模型。该模型使用经过大规模中文语料训练后的 RoBERTa-WWM 预训练模型，相较于传统的词向量模型和仅获得字级别语义的 BERT 模型，它更能解决传统词嵌入层的一词多义问题，可以根据标签数据更好地自动学习并优化模型参数，省去了传统人工构建特征的局限性。实验表明，这种预训练方法更加适用于中医药发明专利文本，能提升本文实体识别任务的准确率。本文模型相比于已有的方法，在中药发明专利摘要中的成分和功能识别任务中的表现更好。

但是，当前并没有统一规范的中医药专利数据集使用，本文所用数据集主要通过爬虫获取，人工筛选预训练标签，因而难免会造成标注误差，影响模型特征提取的准确性。此外，目前中医药专利的数据集量并不大，模型不能充分学习标签特征，也会导致特征提取能力下降。接下来的工作将会进一步扩展数据集，使用更细粒度的标签进行标注，同时考虑使用分治算法，将数据集进一步进行划分与分类，从而更好地优化任务表现。

[参 考 文 献]

[1] SONG M, YU H, HAN W S. Developing a hybrid dictionary-based bio-entity recognition technique[J]. BMC medical informatics and decision making, 2015, 15(S1): 1-8.

[2] LEAMAN R, WEI C H, LU Z. tmChem: a high performance approach for chemical named entity recognition and normalization[J]. Journal of cheminformatics, 2015, 7(S3): 1-10.

[3] UNANUE I J, BORZESHI E Z, PICCARDI M. Recurrent neural networks with specialized word embeddings for health-domain named-entity recognition[J]. Journal of Biomedical Informatics, 2017, 76: 102-109.

[4] GUO B, ZHANG C, LIU J, et al. Improving text classification with weighted word em——beddings via a multi—channel TextCNN model[J]. Neurocomputing, 2019, 363: 366-374.

[5] 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.

[6] TOPAZ M, MURGA L, GADDIS K M, et al. Mining fall-related information in clinical notes: Comparison of rule-based and novel word embedding-based machine learning approaches[J]. Journal of biomedical informatics, 2019, 90: 1-8.

[7] WANG X, ZHANG Y, REN X, et al. Cross-type biomedical named entity recognition with deep multi-task learning[J]. Bioinformatics, 2019, 35(10): 1745-1752.

[8] DENG N, FU H, CHEN X. Named entity recognition of traditional Chinese medicine patents based on bilstm-crf[J]. Wireless Communications and Mobile Computing, 2021(02): 1-12.

[9] ZHANG Z Z, ZHANG Z W, CHEN H Y, et al. A Joint Learning fr-amework With BERT for Spoken Language Understanding[J]. IEEE Access, 2019, 7: 168849-168858.

[10] 沈同平，俞磊，金力，等. 基于 BERT-bilstm-crf 模型的中文实体识别研究[J]. 齐齐哈尔大学学报(自然科学版)，2022，38(01)：26-32.

[11] 王子牛，姜猛，高建瓴，等. 基于 BERT 的中文命名实体识别方法[J]. 计算机科学，2019，46(S2)：138-142.

[12] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Preceedings of NIPS 2017. Long Beach: Neural information processing systems foundation, 2017:5999-6009.

[13] LI Z, CHENG N, SONG W. Research on Chinese event extraction method based on RoBERTa-WWM-CRF[C]// Preceedings of 12th International Conference on Software Engineering and Service Science. NY:IEEE, 2021:100-104.

(下转第 75 页)