文章编号：2095-2163(2023)01-0164-08 中图分类号：TP391;R319 文献标志码：A

基于 MPNet 与 BiLSTM 的 COVID-19 临床文本命名实体识别方法

蔡晓琼[1]，郑增亮[1]，苏前敏[1]，郭晶磊[2]
（1 上海工程技术大学 电子电气工程学院，上海 201620；2 上海中医药大学 基础医学院，上海 201203）

摘 要：随着生物医学研究与信息化技术的迅速发展，临床医学文献数量呈指数级增长，利用文本挖掘技术自动提取医学知识逐渐成为当前研究热点。针对目前新型冠状病毒肺炎（Corona Virus Disease 2019，COVID-19）临床文本研究匮乏、语料不足与标注质量不高等问题，本文结合 UMLS 医学语义网络和专家定义方式，制定医学实体标注规则，建立命名实体识别语料库，明确实体识别任务。其次，提出了一种基于 MPNet 与 BiLSTM 的 COVID-19 临床文本命名实体识别模型。通过预训练语言模型获得文本的向量化表示，解决了一词多义问题；采用双向长短期记忆网络，捕捉文本的长距离依赖；最后引入条件随机场，实现句子级序列注释，输出完整的最优标签序列。实验结果表明，MPNet-BiLSTM-CRF 模型在 COVID-19 临床命名实体识别数据集上取得了较好的表现。

关键词：COVID-19；命名实体识别；双向长短期记忆网络；条件随机场

Named entity recognition of COVID-19 clinical text based on MPNet and BiLSTM

CAI Xiaoqiong[1], ZHENG Zengliang[1], SU Qianmin[1], GUO Jinglei[2]
(1 College of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China;
2 School of Basic Medical Sciences, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China)

【Abstract】 With the rapid development of biomedical research and information technology, the amount of clinical medical literature is growing exponentially, and the automatic extraction of medical knowledge using text mining technology is gradually becoming a current research hotspot. T In view of the current lack of research on Corona Virus Disease 2019 (COVID-19) clinical texts, insufficient corpus, and low quality of labeling, this paper formulates medical entity labeling rules based on the UMLS medical semantic network and expert definition methods, establishes a named entity recognition corpus, and clarifies the entity recognition task. Secondly, a COVID-19 clinical text named entity recognition model based on MPNet and BiLSTM is proposed to obtain a vectorized representation of the text by pre-training the language model to solve the problem of multiple meanings of a word; a bidirectional long and short-term memory network is used in order to capture the long-range dependency of this paper; finally, a conditional random field is introduced to achieve sentence-level sequence annotation and output a complete sequence of optimal labels. The experimental results show that the MPNet-BiLSTM-CRF model achieves better performance on the COVID-19 clinical named entity identification dataset.

【Key words】 COVID-19; named entity recognition; bidirectional long short-term memory network; conditional random field

0 引 言

生物医学命名实体识别，是生物医学信息提取领域最基本的任务之一，其准确性对后续研究工作至关重要。随着生物医学研究与信息化技术的迅速发展，临床医学文献数量呈指数级增长，这些资源中存在大量非结构化数据，迫切需要开发一些自动化技术，来解决生物医学领域的单词分割、实体识别、关系抽取、主题分类等问题。由于生物医学实体的多样性和变异性，识别生物医学实体是一项非常具有挑战性的任务。目前，生物医学命名实体识别方法主要分为基于字典和规则的方法，以及基于深度学习的方法。

基于规则的命名实体识别方法主要依赖于人工

基金项目：“十三五”国家科技重大专项（2018ZX09711001-009-001）；上海市 2017 年度科技创新行动计划（17401970900）。

作者简介：蔡晓琼（1993-），女，硕士研究生，主要研究方向：自然语言处理、数据挖掘；苏前敏（1974-），男，博士，副教授，主要研究方向：智能信息处理、数据挖掘。

通讯作者：苏前敏 Email：suqm@ sues.edu.cn

收稿日期：2022-03-18

定义的词典和规则，可以根据特定领域的词典和句法-词汇模式来进行设计。一些著名的基于规则的命名实体识别模型包括 LaSIE-II、Sra、FASTUS 和 LTG 等[1-4]。与传统方法相比，引入深度学习技术的命名实体识别方法，通过在大规模语料库中进行训练与学习，自动提取语义特征，使模型具有较强的泛化能力，同时降低了人力与时间成本。目前常用的词嵌入方法是由 Mikolov 等[5]提出的 Word2Vec 模型。但该方法训练获得的词向量是静态的，无法解决一词多义问题。随着自然语言处理技术的发展，ELMo[6]、BERT[7]和 XLNet[8]等预训练语言模型的推出，为下游任务提供了极大的帮助，对模型进行微调后，在命名实体识别任务上取得了较好的效果。但 ELMo 输出的语义表征是单向的，无法获得更全面的双向信息；BERT 和 XLNet 采用的 MLM 和 PLM 方法各自存在局限性，在用于下游任务微调时，会造成预训练和微调不匹配。为了解决以上问题，南京大学和微软共同提出了基于 MLM 和 PLM 各自优点的预训练模型 MPNet[9]，弥补了 MLM 无法学习 tokens 之间依赖关系的不足，同时克服了 PLM 无法获得下游任务中可见完整信息的问题。

针对目前临床试验文本研究匮乏、语料不足与标注质量不高等问题，本文结合 UMLS 语义网络和专家定义方法建立了 COVID-19 临床实体识别语料库，并将预训练语言模型 MPNet 引入 COVID-19 临床实体识别任务中，提出了一种基于 MPNet 与 BiLSTM 的医学实体识别模型。

1　MPNet-BiLSTM-CRF 命名实体识别模型

本文提出的 MPNet-BiLSTM-CRF 命名实体识别模型主要结构如图 1 所示。采用 MPNet 预训练模型作为嵌入层对输入进行语义提取生成动态词向量；采用 BiLSTM 捕捉长距离依赖关系；最后由 CRF 推理层进行最佳标注序列解码，预测出全局最优标签。该方法在经典 BiLSTM-CRF 的基础上进行了改进，引入了 MPNet 语言模型，在预测掩码标记的同时以更多的信息为条件，从而获得更好的学习表征，并减少了微调阶段的差异。

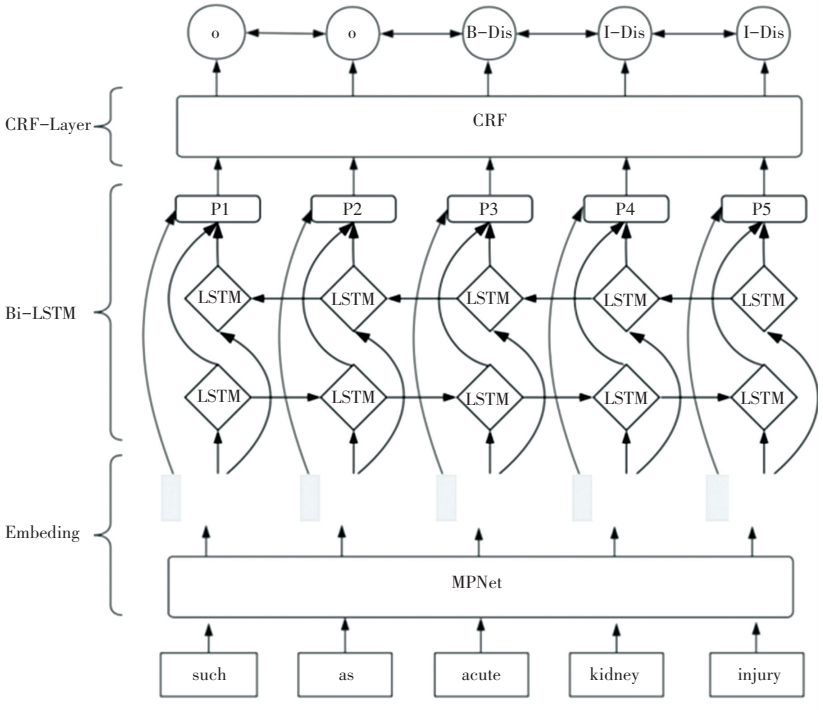

图 1　MPNet-BiLSTM-CRF 模型结构图

Fig. 1　The structure of MPNet-BiLSTM-CRF model

1.1　MPNet 预训练模型

MPNet 模型的注意力掩码机制：首先设长度 $n=6$ 的输入序列为 $(x_1,x_2,x_3,x_4,x_5,x_6)$，若随机生成的序列为 $(x_5,x_4,x_2,x_6,x_3,x_1)$，预测值分别是 x_6、x_3 和 x_1，则非预测序列表示为 $(x_5,x_4,x_2,$[MASK]，[MASK]，[MASK]$)$，对应位置序列为 $(P_5,P_4,P_2,P_6,P_3,P_1)$。其次，为了能让预测部分的[MASK]看到之前预测的 tokens，MPNet 采用了 PLM 双流自注意力机制完成自回归生成，并为内容流和查询流设置了不同的掩蔽机制。例如，MPNet 在预测上述序列中的 x_3 时，能在非预测部分看到 $(x_5+P_5,x_4+P_4,x_2+P_2)$，同时在预测部分看到之

前预测的 $(x_6 + P_6)$，从而避免 MLM 中依赖关系遗漏的问题。此外，为了确保预训练中输入信息与下游任务中输入信息的一致性，MPNet 在非预测部分增加了掩码符号和位置信息（[MASK] + P_6，[MASK] + P_3，[MASK] + P_1），使模型能看到完整的句子。当预测 x_3 时，能在非预测部分看到原始的 $(x_5 + P_5, x_4 + P_4, x_2 + P_2)$ 以及引入了额外 tokens 和位置信息的（[MASK] + P_3，[MASK] + P_1），同时在预测部分看到之前预测的 $(x_6 + P_6)$。通过上述办法对查询流和内容流进行位置补偿后的模型，能够大幅减少预训练与微调之间输入不一致的问题。

假定句子为“to take up seasonal flu vaccination”，模型输入序列为［to，take up，seasonal，flu，vaccination］，需要预测的 token 是［seasonal，flu，vaccination］，MPNet 的因子化如下：

$$\log P(\text{seasonal} \mid \text{to take up [MASK] [MASK] [MASK]}) + \log P(\text{flu} \mid \text{to take up [seasonal] [MASK] [MASK]}) + \log P(\text{vaccination} \mid \text{to take up [seasonal] [flu] [MASK]}) \quad (1)$$

1.2 Transformer

自 Vaswani 等[10]提出 Transformer 这一基于自注意力机制的深度学习模型以来，已被广泛应用到 NLP 领域中解决各种复杂问题，几款主流的预训练语言模型（如 BERT、XLNet 和 ALBERT 等）都以 Transformer 作为其骨干网络。传统语言模型通常基于 CNN 或 RNN 编码器进行训练，在长周期语境建模中能力较为欠缺，并且在学习单词表征时存在位置偏差，尤其 RNN 是按顺序处理输入的，即一个字一个字地处理，对计算机硬件的并行能力要求较高。为了克服现有深度学习模型的缺点，Transformer 在每个关键模块中引入了 Attention 机制，大幅提升了模型对文本的特征提取能力。此外，Transformer 还引入了多头注意力机制（Multi-Head Attention），使模型能够使用各个序列位置的各个子空间的表征信息来进行序列数据处理，其相当于多个不同的自注意力模块的集成，从而构成蕴含完整语料信息的多粒度特征。标准的 Transformer 结构如图 2 所示。

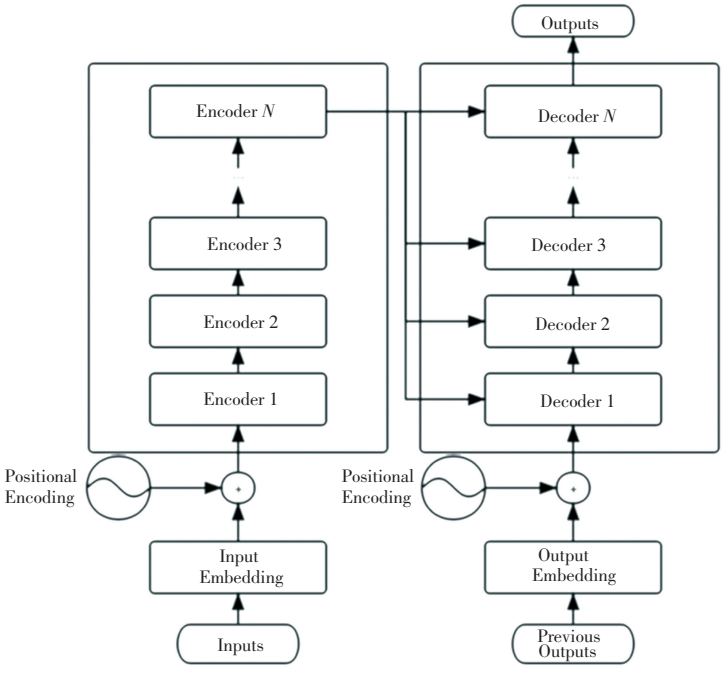

图 2 标准 Transformer 结构图

Fig. 2 Standard Transformer structure diagram

通过图 2 可知，Transformer 由 encoder 和 decoder 两部分组合而成。其中，encoder 通过 6 个编码器叠加构成，每个 encoder 中都包含一个自注意力层和前馈神经网络层，两个子层之间通过残差网络结构进行连接，后接一个正则化层。为了方便各层之间的残差连接，模型中所有子层的输出维度均为 512。Decoder 也由 6 个完全相同的解码器叠加组成，除了与编码器完全一致的两个子层外，Decoder 还设置了一个多头注意力层。多头注意力层是由多个自注意力层拼接而成的，可以捕捉到单词之间各维度上的相关系数。

1.3　双向长短期记忆网络(BiLSTM)模型

为了解决传统循环神经网络(Recurrent Neural Network，RNN)在训练较长语句时可能导致梯度爆炸和梯度消失问题，Hochreiter 等人[11]提出了长短期记忆网络(Long Short-Term Memory，LSTM)，该模型能够在长文本训练中捕捉长距离依赖特征。与标准 RNN 模型相比，LSTM 在其结构基础上增加了门控机制和记忆单元两个模块。其中，记忆单元用于存储文本特征，门控机制则对记忆单元中的存储信息进行筛选。LSTM 模型分别设置了输入门、遗忘门和输出门保护和控制细胞状态，通过累加更新传递信息的方式，消除了 RNN 模型在处理长文本任务时可能出现的问题，其单元结构如图 3 所示。

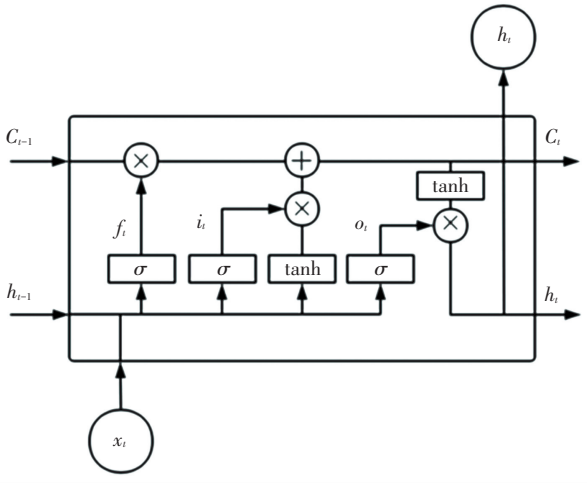

图 3　LSTM 单元结构图

Fig. 3　LSTM unit structure

LSTM 模型是由 t 时刻的输入词 X_t、细胞状态 C_t、临时细胞状态 $\tilde{C}_t$，隐含状态 h_t、遗忘门 f_t、记忆门 i_t 及输出门 o_t 组成，通过对细胞状态中信息遗忘和记忆新的信息，使得对后续时刻计算有用的信息得以传递，而无用的信息被摒弃，并在每个时间步都会输出隐含状态 h_t。其中遗忘、记忆与输出通过上一时刻的隐含状态 h_{t-1} 和当前输入 X_t 计算出来的遗忘门 f_t、记忆门 i_t 和输出门 o_t 来控制，遗忘门计算如公式(2)：

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

其中，输入为前一时刻的隐含状态 h_{t-1} 和当前时刻的输入词 X_t。

记忆门的值 i_t 和临时细胞状态 $\tilde{C}_t$ 分别由公式(3)和公式(4)计算得出：

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{3}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{4}$$

根据输入为记忆门的值 i_t、遗忘门的值 f_t、临时细胞状态 $\tilde{C}_t$ 和上一时刻细胞状态 C_{t-1} 计算 t 时刻细胞状态 C_t 为

$$C_t = \sigma(f_t \cdot C_{t-1} + i_t \cdot \tilde{C}_t) \tag{5}$$

前一时刻的隐含状态 h_{t-1}、当前时刻的输入词 X_t 和当前时刻细胞状态 h_t，计算 t 时刻输出门的值 o_t 和隐含状态 h_t 为

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{6}$$

$$h_t = o_t * \tanh(C_t) \tag{7}$$

其中，σ 为 *sigmoid* 函数；tanh 为双曲正切激活函数；$\boldsymbol{W}$ 和 $\boldsymbol{b}$ 分别表示链接两层的权重矩阵和偏置向量。经过 LSTM 模型计算后，最终即可得到与句子长度相同的隐含状态序列 $\{h_0, h_1, \cdots, h_{n-1}\}$。

对于处理 NLP 任务(尤其对序列标注任务)，上下文内容无论对单词、词组还是字符，在整个研究过程中都尤为重要。通常情况下，LSTM 常用单元为前向传播，然而在研究序列问题时，前向 LSTM 无法处理下文的内容信息，从而导致模型无法学习到下文的知识，影响最终模型效果。而双向长短期记忆网络(Bi-directional Long Short-Term Memory，BiLSTM)既能获取上文信息，又能捕获下文内容，对双向信息都能进行记忆，通过同时得到前后两个方向的输出，来提高整个 NLP 模型的性能。BiLSTM 模型结构如图 4 所示。

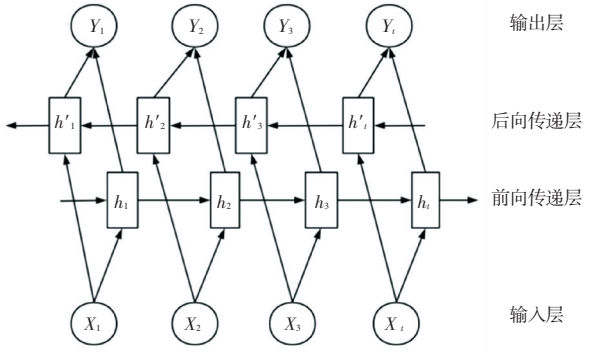

图 4　BiLSTM 模型结构图

Fig. 4　BiLSTM model structure

1.4　条件随机场(CRF)

条件随机场(Conditional Random Field，CRF)是 Lafferty 等[12]基于最大熵模型和隐马尔可夫模型所提出的一种判别式概率无向图学习模型，常用于标注和切分有序数据的条件概率模型。在序列任务标注任务中，令 $X = (X_1, X_2, \cdots, X_n)$ 表示模型的观察序列，$Y = (Y_1, Y_2, \cdots, Y_n)$ 表示状态序列，$P(Y \mid X)$ 为线性链条件随机场，则在随机变量 X 取值为 x 的条件下，计算状态序列 Y 取值为 y 的条件概率分

布 $P(Y \mid X)$,计算公式为

$$P(y \mid x) = \frac{1}{Z(x)} \exp\Big(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} u_l s_l(y_i, x, i) \Big) \tag{8}$$

其中,t_k、s_l 为特征函数,其函数取值为 0 或 1(当满足特征条件时,函数取值为 1,反之为 0);λ_k、u_l 为对应权重;$Z(x)$ 为归一化因子,其计算公式为

$$Z(x) = \sum_{y} \exp\Big(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} u_l s_l(y_i, x, i) \Big) \tag{9}$$

为了对上述公式进行简化,可将转移特征、状态特征及其权重用统一的符号进行表示,简化后的公式为

$$P(y \mid x) = \frac{1}{Z(x)} \exp \sum_{k=1}^{K} w_k f_k(y, x) \tag{10}$$

其中,

$$Z(x) = \sum_{y} \exp\Big(\sum_{k=1}^{K} w_k f_k(y, x) \Big) \tag{11}$$

CRF 中最重要的就是训练模型的权重。通常情况下,使用最大化对数似然函数进行 CRF 模型的训练,通过式(12)计算在给定条件下,标签序列 Y 的条件概率:

$$L = \log\left(P(y \mid x)\right) \tag{12}$$

最后,在 CRF 模型预测过程中采用维特比(Viterbi)算法来求解全局最优序列,通过该算法可以计算出与预测对象对应的最大概率标签,计算公式为

$$\hat{y} = \operatorname{argmax} P(y \mid x) \tag{13}$$

2 实验结果与分析

2.1 实验数据

本文使用的是美国临床试验注册中心(ClinicalTrials,CT)中的 COVID-19 相关临床试验注册数据,CT 官方网址为 https://clinicaltrials.gov。CT 中收录了临床研究者在世界各地进行的私人或公共资助的临床研究项目,其中包含有关人类志愿者医学研究的信息。随着 COVID-19 的爆发,越来越多相关的临床试验在此平台进行了注册,临床记录中富含 COVID-19 相关的临床医学知识。考虑到试验结束前可能存在信息更新不完整的问题,本文基于已完成的 697 项干预性试验数据进行 COVID-19 临床试验的命名实体识别实验。

2.2 实验配置

本研究采用的 NER 模型基于 PyTorch 深度学习框架,实验环境配置见表 1。

表 1 实验环境配置

Tab. 1 Experimental environment configuration

项目	实验环境
操作系统	Ubuntu16.04
CPU	i7-11370H@3.3GHz
GPU	RTX3080(16G)
Python 版本	3.7.0
PyTorch 框架	1.7.1

本研究采用微软发布的 MPNet 模型,由 12 个 Transformer 层组合而成,隐藏层维度设为 768,12 个注意头模式;使用 GELU 作为其激活函数,BiLSTM 隐藏单元为 128。在训练阶段,MPNet-BiLSTM-CRF 的最大序列长度为 256,batch_size 为 128,MPNet 学习率设为 3e-5,Dropout 为 0.1,其他模块 Dropout 为 0.3,并通过 Adam 优化算法对模型进行训练。

2.3 概念定义与标注策略

临床试验注册中的非结构化文本主要涉及了具有特定意义的相关医学实体。例如:药物名称“Remdesivir”、医疗程序“bronchoalveolar lavage”、疾病名称“COVID-19”等等。不同的研究对医疗实体的标注规则和定义都有一定差异,统一医学语言系统(Unified Medical Language System,UMLS)收录了超过 500 万条生物医学术语,至少 200 万种医学概念,目前已广泛应用于文献分类、临床研究和中英文电子病历等领域中。本研究参考了 UMLS 定义的实体类别以及文献[13]中提出的医学实体标注规范,对临床试验注册内容规定了实体标准和含义,结合文本内容定义了 8 种类别的临床实体,COVID-19 临床文本命名实体识别示例见表 2。

表 2 实体类型定义与示例

Tab. 2 Entity type definitions and examples

序号	实体类别	实体英文	示例
1	疾病	Disease	COVID-19
2	药物	Drug	Remdesivir
3	临床表现	Symptom	cough
4	医疗程序	Procedure	bronchoalveolar lavage
5	程度	Severity	critical ill
6	医疗设备	Equipment	bronchoscopy
7	医学检验项目	Item	vital signs
8	身体	Body	Lung

NER 任务旨在提取文本中的命名实体,如名称或带有适当 NER 类的标签,本文使用 BIO 标注方式对序列进行标注。在这种格式中,不属于实体的标

记被标记为“O”，“B”标记对应实体的第一个单词，“I”标记对应同一实体的其余单词。“B”和“I”标签后跟连字符（或下划线），后跟实体类别缩写（如 Dru、Dis、Syp 等），表 3 是对 COVID-19 临床实体预测标签的示例。

表 3　实体预测标签定义

Tab. 3　Entity prediction label definitions

序号	实体类别	开始标签	中间标签
1	疾病 Disease	B-Dis	I-Dis
2	药品名称 Drug	B-Dru	I-Dru
3	临床表现 Symptom	B-Sym	I-Sym
4	医疗程序 Procedure	B-Pro	I-Pro
5	程度 Severity	B-Ser	I-Ser
6	医疗设备 Equipment	B-Equ	I-Equ
7	医学检验项目 Item	B-Ite	I-Ite
8	身体 Body	B-Bod	I-Bod

2.4　评价指标

本研究采用精确率（P）、召回率（R）和 $F1$ 值（$F1$）作为模型的评价指标。P 是指正确识别的实体占全部预测实体的比重，R 是正确识别的实体占语料库中所有实体的比重。各指标对应的计算公式如下：

$$P = \frac{TP}{TP + FP} \tag{14}$$

$$R = \frac{TP}{TP + FN} \tag{15}$$

$$F1 = \frac{2PR}{P + R} = \frac{2TP}{2TP + FP + FN} \tag{16}$$

其中，TP 指正确地将正例预测为正的数量；FP 指错误地将负例预测为正的数量；FN 指错误地将正例预测为负的数量。

2.5　对比实验

为了验证本文提出的 MPNet-BiLSTM-CRF 融合模型对 COVID-19 临床试验注册实体具有较好的识别效果，与以下几种方法进行对比实验：

（1）经典的 BiLSTM-CRF 模型：该模型采用 word2vec 训练得到的词，嵌入向量作为输入，然后通过双向 LSTM 层和 CRF 完成编码与识别任务。目前该模型已广泛应用于中英文生物医学 NER 任务中，并取得了良好的效果。

（2）Att-BiLSTM-CRF 融合模型[14]：该模型通过引入 Attention 机制，确保模型能够专注于标记本文中同一 token 的多个实例之间的一致性。

（3）XLNet-BiLSTM-CRF 模型[15]：该模型使用 XLNet 预训练语言模型提取句子特征，然后将经典神经网络模型与获取的特征相结合，在公共医疗数据集上识别效果较好。

（4）BERT-BiLSTM-CRF 模型：通过使用 BERT 替换上个模型中的 XLNet 方法。与现有方法相比，该模型在英文生物医学 NER 任务中具有更好的表现[16]。

2.6　结果分析

对于不同的实体类型，各模型的实验结果见表 4。由表 4 可知，本文提出的 MPNet-BiLSTM-CRF 模型在大部分实体上表现较好，少数几类实体上表现略逊于基于 XLNet 的融合模型，但 $F1$ 值的差别不明显。在所有模型的实验结果中，“Disease”、“Symptom”和“Severity”的 $F1$ 值较高，这是由于在摘要中这 3 类实体结构简单且包含的信息种类较少，模型能够充分学习这些文本的特征。通过分析另外几种实体发现，存在训练数据集较少导致过拟合现象，另外部分实体结构复杂且出现次数少（例如为特定情况下 COVID-19 防治选用的药物、检查或治疗措施），导致模型难以充分提取其特征。在后续 COVID-19 临床试验摘要的 NER 任务中，可以通过适量加入专业的英语语料库来增加语义特征，从而优化模型的识别能力。

由表 5 中展示的实验结果可知，在 COVID-19 临床试验注册数据集中，MPNet-BiLSTM-CRF 模型与其它 4 种模型相比，整体的精确率、召回率和 $F1$ 值都有所提高。经典 BiLSTM-CRF 模型实体识别的 $F1$ 值为 69.42%，引入 Attention 机制后的模型 $F1$ 值提升了 2.49%。注意力矩阵能够计算当前的目标单词与序列中所有单词的相似性，通过权重矩阵为不同重要程度的单词分配相应的权重值，计算出文本的全局向量作为 BiLSTM 输出的加权和。对比 ERT-BiLSTM-CRF 和 Att-BiLSTM-CRF 模型，前者实验结果的 3 项指标均有提升，$F1$ 值比后者提升了 1.26%，说明 BERT 预训练语言模型能更好地捕捉语义关系和单词特征。BERT-BiLSTM-CRF 与 XLNet-BiLSTM-CRF 模型相比，精确率分别是 72.60%和 74.57%，$F1$ 值分别为 73.17%和 74.01%。相比之下，基于 XLNet 的 NER 模型效果略优于 BERT。而本文提出的 MPNet-BiLSTM-CRF 融合模型与 XLNet-BiLSTM-CRF 模型相比，$F1$ 值提高了 1.06%，通过使用融合了 BERT 和 XLNet 优点的 MPNet 预训练语言模型，增强了序列的特征表示，同

时弥补了两者的缺陷,从而提高了模型整体的识别能力。此外,MPNet 模型的识别能力提高后,速度并未有明显下降,因此 MPNet 作为词嵌入层将具有更优秀的表现。

表 4 不同实体类型的识别结果

Tab. 4 Identification results of different entity types

模型	指标	实体类型							
		Disease	Drug	Symptom	Procedure	Severity	Equipment	Item	Body
BiLSTM-CRF	P	80.89	73.21	77.88	65.05	83.23	60.87	62.12	69.05
	R	78.35	71.75	73.64	64.42	81.13	61.54	64.06	64.93
	*F*1	79.60	72.47	75.70	64.73	82.17	61.20	63.08	66.93
Att-BiLSTM-CRF	P	81.53	74.24	78.50	69.11	83.95	68.04	68.30	70.16
	R	79.61	72.95	76.36	67.18	85.53	63.01	66.54	63.97
	*F*1	80.56	73.59	77.42	68.13	84.73	65.43	67.41	66.92
BERT-BiLSTM-CRF	P	82.33	74.25	82.14	71.61	86.06	63.54	66.55	66.67
	R	79.77	73.88	81.42	68.79	87.65	68.96	70.34	70.31
	*F*1	81.03	74.06	81.78	70.17	86.85	66.14	68.39	68.44
XLNet-BiLSTM-CRF	P	83.13	75.87	83.04	71.07	87.67	68.68	67.75	71.76
	R	79.92	74.75	81.58	68.90	87.73	67.10	70.04	69.12
	*F*1	81.49	**75.31**	**82.30**	69.97	**87.70**	67.88	68.88	70.42
MPNet-BiLSTM-CRF	P	84.80	75.43	83.49	73.48	89.44	68.45	69.45	73.88
	R	80.61	74.88	79.13	72.37	85.71	70.07	70.22	74.60
	*F*1	**82.65**	75.15	81.25	**72.92**	87.54	**69.25**	**69.83**	**74.24**

表 5 各模型整体对比结果

Tab. 5 Overall comparison results for each model %

模型	精确率	召回率	*F*1 值
BiLSTM-CRF	69.86	68.98	69.42
BiLSTM-Attention-CRF	73.18	70.68	71.91
BERT-BiLSTM-CRF	72.60	73.75	73.17
XLNet-BiLSTM-CRF	74.57	73.45	74.01
MPNet-BiLSTM-CRF	75.55	74.60	75.07

3 结束语

本文提出了一种基于 MPNet 和 BiLSTM 的医学实体识别模型,联合 BiLSTM 网络适应长文本特征提取和 CRF 序列标注方法,能适应于 COVID-19 临床试验注册文本中新兴医学实体的识别任务。实验设置了多组对比模型以验证本文方法的有效性,结果表明其识别性能优于基准模型以及近年来被广泛研究的基于主流预训练模型的实体识别方法,并且能够较好地实现 COVID-19 相关临床医学实体的识别任务,对医学领域相关研究具有一定参考价值。本实验数据仅包含 697 份临床试验注册记录中的摘要文本,存在实体种类多但数量不均衡的问题,因此在接下来的工作中,将纳入更多临床文本来丰富语料库,为挖掘 COVID-19 临床文本中隐含的医学知识与临床价值做准备。

参考文献

[1] GAIZAUSKAS R, HUMPHREYS K, CUNNINGHAM H, et al. University of Sheffield: description of the LaSIE system as used for MUC-6 [C]//Proceedings of the 6th conference on Message understanding. Columbia, Maryland: Association for Computational Linguistics. 1995: 207.

[2] AONE C, HALVERSON L, HAMPTON T, et al. SRA: Description of the IE2 system used for MUC-7 [C]//Seventh Message Understanding Conference (MUC-7): Proceedings of a Conference Held in Fairfax, Virginia, April 29-May 1, 1998: 1998.

[3] APPELT D, HOBBS J R, BEAR J, et al. SRI International FASTUS systemMUC-6 test results and analysis [C]//Sixth Message Understanding Conference (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995: 1995.

[4] MIKHEEV A, MOENS M, GROVER C. Named entity recognition without gazetteers[C]//Ninth Conference of the European Chapter of the Association for Computational Linguistics. 1999: 1-8.

[5] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [J]. arXiv preprint arXiv:1301.3781, 2013.

(下转第 177 页)