DOI:10.3969/j.issn.1005-202X.2024.01.018

医学人工智能

基于BioBERT与BiLSTM的临床试验纳排标准命名实体识别

李盛青[1],苏前敏[1],黄继汉[2]
1.上海工程技术大学电子电气工程学院,上海 201620;2.上海中医药大学药物临床研究中心,上海 201203

【摘要】**目的:**提出一种基于BioBERT预训练模型的纳排标准命名实体识别方法(BioBERT-Att-BiLSTM-CRF),可自动提取临床试验相关信息,为高效制定纳排标准提供帮助。**方法:**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,并建立命名实体识别语料库以明确实体识别任务。BioBERT-Att-BiLSTM-CRF首先将文本转换为BioBERT向量并输入至双向长短期记忆网络以捕捉上下文语义特征;同时运用注意力机制来提取关键特征;最终采用条件随机场解码并输出最优标签序列。**结果:**BioBERT-Att-BiLSTM-CRF在纳排标准命名实体识别数据集上的效果优于其他基准模型。**结论:**使用BioBERT-Att-BiLSTM-CRF能更高效地提取临床试验的纳排标准相关信息,从而增强临床试验注册数据的科学性,并为临床试验纳排标准的制定提供帮助。

【关键词】纳排标准;命名实体识别;双向长短期记忆网络;条件随机场;临床试验
【中图分类号】R318 【文献标志码】A 【文章编号】1005-202X(2024)01-0125-08

Named entity recognition of eligibility criteria for clinical trials based on BioBERT and BiLSTM

LI Shengqing[1], SU Qianmin[1], HUANG Jihan[2]
1. School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China; 2. Center for Drug Clinical Research, Shanghai University of Traditional Chinese Medicine, Shanghai 201203, China

Abstract: Objective To present a named entity recognition method referred to as BioBERT-Att-BiLSTM-CRF for eligibility criteria based on the BioBERT pretrained model. The method can automatically extract relevant information from clinical trials and provide assistance in efficiently formulating eligibility criteria. **Methods** Based on the UMLS medical semantic network and expert-defined rules, the study established medical entity annotation rules and constructed a named entity recognition corpus to clarify the entity recognition task. BioBERT-Att-BiLSTM-CRF converted the text into BioBERT vectors and inputted them into a bidirectional long short-term memory network to capture contextual semantic features. Meanwhile, attention mechanisms were applied to extract key features, and a conditional random field was used for decoding and outputting the optimal label sequence. **Results** BioBERT-Att-BiLSTM-CRF outperformed other baseline models on the eligibility criteria named entity recognition dataset. **Conclusion** BioBERT-Att-BiLSTM-CRF can efficiently extract eligibility criteria-related information from clinical trials, thereby enhancing the scientific validity of clinical trial registration data and providing assistance in the formulation of eligibility criteria for clinical trials.

Keywords: eligibility criteria; named entity recognition; bidirectional long short-term memory network; conditional random field; clinical trial

前 言

根据世界卫生组织(World Health Organization, WHO)的定义,临床试验是一种以志愿者为主要研究对象的科学研究,旨在评估新的实验性药物、现有药品和医疗设备的疗效与安全性的系统性试验,对促进医学发展和提高人类健康具有积极作用[1]。美国临床试验注册中心是全球最大的临床试验注册平台之一,其数据覆盖范围广、数据质量高且更新及时,为临床试验的设计和优化提供有力的支持,其中的纳入排除标准(简称“纳排标准”)是决定受试者能否被纳入实验组的关键因素,也是试验成功的前提条件之一。

患者数量招募不足是过去的临床试验经常面临

【收稿日期】2023-08-20
【作者简介】李盛青,硕士研究生,研究方向:人工智能技术,E-mail: lsq1118@126.com
【通信作者】苏前敏,博士,副教授,研究方向:医学数据挖掘、医学数据分析,E-mail: suqm@sues.edu.com

的问题之一[2]，而这种情况通常与纳排标准存在直接或间接的关系。在早期的研究中，纳排标准的制定主要通过研究人员依靠手动检索与比较分析法进行，然而这两种方法费时且容易出现主观误差，严重影响筛选方案的制定效率和准确性。

与复杂繁琐的手动流程相比，经过严格测试的人工智能算法能快速、高效地制定纳排标准，提高临床试验的效率和质量，同时缩短试验的周期并降低成本。因此，采用人工智能算法辅助筛选方案的制定和优化已成为临床试验设计和实施的研究热点。作为信息抽取的子任务，命名实体识别可以识别处理医学中的专有名词（如药物名称等）[3]。但是，由于生物医学实体的多样性与变异性，识别生物医学实体是一项具有挑战性的任务。生物医学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。基于规则和字典的命名实体识别方法主要利用已有的标准术语词典及匹配算法识别文本中出现的术语，并结合领域专家的观点建立词典或规则模板[4-5]。医学领域专业术语众多，随着新的实体名称不断出现，词典的及时更新将面临巨大挑战；此外，单纯依赖传统的词典匹配方法也难以达到较高的性能，通常需要结合其他方法使用[6]。虽然基于复杂规则的系统精确率高，但随着规则变得越来越特殊，召回率会越来越低。因此，该方法通常与机器学习方法相结合以提高模型性能[7-8]。机器学习已被广泛应用于序列标注问题的研究中，对序列中的每个单词赋予特定标签，通过输入单词序列，输出相应的实体和预测结果。机器学习主要解决两个问题，即确定实体边界和预测实体类型，并且能为每个实体分配特定标签，以表明其开始、中间和结束等词位信息。

自然语言处理（Natural Language Processing, NLP）的深度学习技术不断进步，为生物医学文本挖掘模型带来新的可能性。2015年，百度研究院提出深度学习应用于命名实体识别的模型，即双向长短期记忆网络-条件随机场（Bidirectional Long Short-Term Memory Network - Conditional Random Field, BiLSTM-CRF），该模型通过深度建模上下文信息，再利用条件随机场解码整个句子的标签。如今，NLP已不再是单个模型处理单个任务，而是在大量语料上预训练通用模型，并对特定下游任务进行微调，ELMO、GTP、BERT等微调后的模型在许多NLP任务上都表现优异。自动纳排标准采用了多种方法，包括基于模式匹配和规则的EliXR[9]、EliXR-TIME[10]和ERGO系统等。此外，还有大量的研究集中在信息抽取方面，如EliIE[11]和Criteria2Query[12]以及Tseo等[13]的工作。

针对目前临床试验纳排标准标注语料匮乏和术语专业性强等问题，本研究参考医学术语系统UMLS，并结合医学专家定义和纳排标准数据的特点，预先制定纳排标准实体标注规则，采用BIO标注方式创建基于纳排标准的训练语料库，并将预训练语言模型BioBERT引入纳排标准实体识别任务中，提出一种基于BioBERT与BiLSTM的医学实体识别模型。

1 BioBERT-Att-BiLSTM-CRF命名实体识别模型

本研究提出的医学实体识别模型BioBERT-Att-BiLSTM-CRF的整体架构如图1所示。首先将英语文本转换成BioBERT的输入格式。然后将其输入BioBERT网络中以识别句子特征。BioBERT通过编码层的BiLSTM捕捉词向量中的长距离依赖关系，得到句子中每个单词的正确标签；使用注意力机制提取重要特征并减少噪声干扰。最后采用CRF层为最终预测的标签引入约束条件，以提高标签预测的准确率。该方法在经典BiLSTM-CRF的基础上进行改进，引入BioBERT语言模型以及注意力机制，进一步提高命名实体识别的准确性和效率。

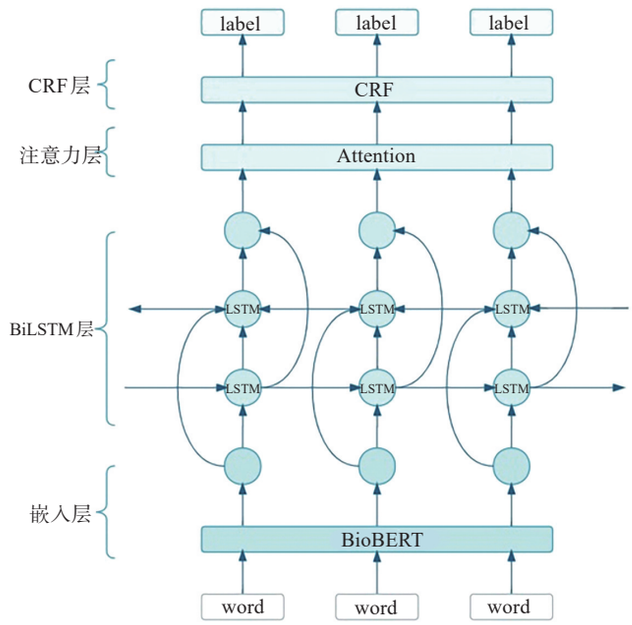


图1 BioBERT-Att-BiLSTM-CRF模型架构图

Figure 1 BioBERT-Att-BiLSTM-CRF model architecture

1.1 BioBERT模型

BioBERT是针对生物医学领域的语言表示模型[14]。该模型利用大规模的生物医学语料库进行预训练。BioBERT与BERT具有相同的架构，编码器均采用双向Transformer，并且基于注意力机制表示文

本序列的上下文关系，能很好地并行计算和捕获长距离文本特征。BioBERT的Embedding包括3种不同的嵌入特征，分别是Token Embeddings、Segment Embeddings和Position Embeddings。具体嵌入特征如图2所示。

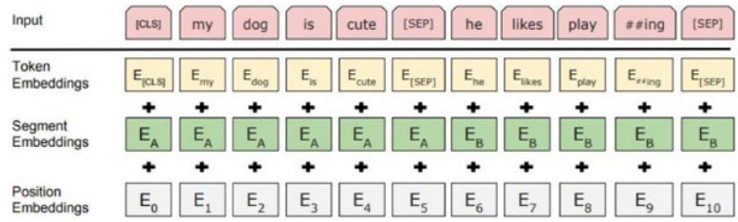


图2 Embedding示意图

Figure 2 Embedding diagram

为提高模型效果，该模型结合了“掩码语言模型（Masked Language Model）”和“下一句预测（Next Sentence Prediction）”两个任务。在掩码语言模型中，每次会随机选取文本序列中15%的词，其中，80%的词被掩盖，即用特殊的标记符号替换原始词汇；10%的词被替换为随机生成的其他词；10%则直接保留原始词，不进行任何处理。模型需要根据上下文信息预测被掩盖的词。下一句预测任务则需要输入两个句子A和B，从中随机选取两个句子进行训练，其中，50%的概率是连续的上下文，另外50%的概率则是不连续的。模型需要判断句子B是否为句子A的下一句，以判断这两个句子之间的关系。

在生物医学语料库上进行预训练后，BioBERT在多种生物医学文本挖掘任务中的表现远远超过BERT和其他先进的模型。在生物医学命名实体识别和生物医学关系提取任务中，BioBERT的F1值分别提高0.62%和2.80%；而在生物医学问答任务中，平均倒数排名提高12.24%，表现显著优于其他模型[14]。

BioBERT的预训练和微调概述如图3所示。首先，采用通用领域预训练的BERT权重对BioBERT进行初始化；随后，通过使用生物医学领域语料库（涵盖PubMed和PMC中的文本）对BioBERT进行进一步预训练；最后，对BioBERT进行微调，并通过命名实体识别、关系抽取和问答等任务进行评估。

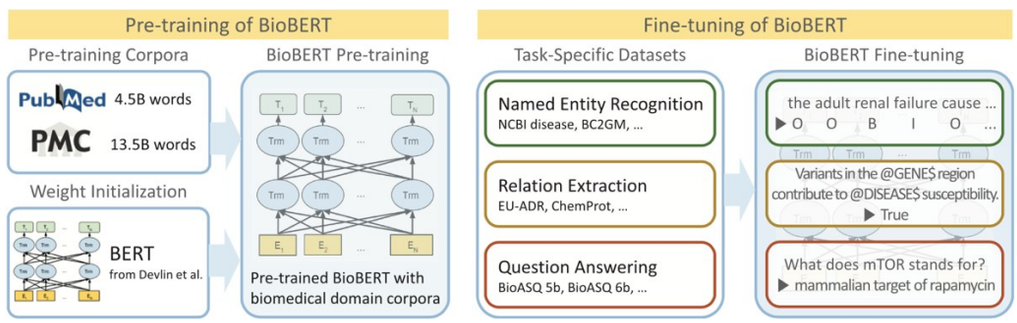


图3 BioBERT的预训练和微调

Figure 3 Pre-training and fine-tuning of BioBERT

1.2 BiLSTM模型

传统的循环神经网络在处理长文本时可能会出现梯度消失或梯度爆炸的问题。为解决这个问题，Hochreiter等[15]提出长短期记忆网络（LSTM）。相比传统的循环神经网络模型，LSTM通过增加门控机制和记忆单元的方式来捕捉长距离依赖关系。门控机制主要用于存储文本特征，而记忆单元则用于筛选已经存储的信息。LSTM模型通过累加更新的方式来传递信息，避免在处理长文本时可能出现的问题。LSTM的单元结构如图4所示。其中，X_t表示t时刻的输入向量，C_t表示记忆细胞，$\widetilde{C}_t$表示中间状态，h_t表示隐藏状态，f_t表示遗忘门，i_t表示输入门，o_t表示输

出门。遗忘门确定前一个步长中保留或摒弃哪些信息，输入门用于处理当前序列位置的输入，控制记忆单元决定存储哪些重要信息，输出门确定下一个隐藏状态。

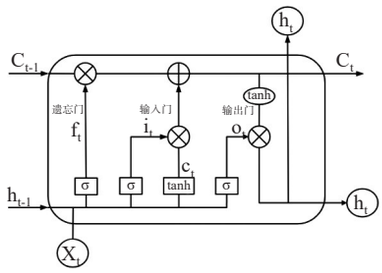


图4 LSTM单元结构图

Figure 4 LSTM unit structure

遗忘门基于前一时刻的隐藏状态h_{t-1}和当前时刻的输入词X_t计算得出，具体公式如下：

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

输入门的值i_t和中间状态$\widetilde{C}_t$的计算公式如下：

$$i_t = \sigma(W_i \cdot [h_{t-1}, X_t] + b_i) \tag{2}$$

$$\widetilde{C}_t = \tan(W_C \cdot [h_{t-1}, X_t] + b_c) \tag{3}$$

t时刻的细胞状态C_t基于输入门的值i_t、遗忘门的值f_t、中间状态$\widetilde{C}_t$和前一时刻细胞状态C_{t-1}计算得出，具体公式如下：

$$C_t = \sigma(f_t \cdot C_{t-1} + i_t \cdot \widetilde{C}_t) \tag{4}$$

t时刻输出门的值o_t和隐藏状态h_t由前一时刻的隐藏状态h_{t-1}、当前时刻的输入词X_t和当前时刻隐藏状态h_t计算得出，计算公式如下：

$$o_t = \sigma(\boldsymbol{W}_o \cdot [h_{t-1}, X_t] + \boldsymbol{b}_o) \tag{5}$$

$$h_t = o_t * \tan(C_t) \tag{6}$$

其中，σ为sigmoid函数，其输出值范围在0到1之间，用于表示需要保留或遗忘的信息的比例；$\boldsymbol{W}$和$\boldsymbol{b}$分别表示链接两层的权重矩阵和偏置向量，通过反向传播算法更新，使得模型可以适应输入数据的特征。

然而，LSTM模型还存在一些缺陷。通常情况下，前向LSTM无法处理下文的内容信息，从而限制模型在学习下文信息时的表现，对模型的最终性能产生不良影响，特别是在处理序列标注任务等NLP任务时，上下文信息对于单词、词组甚至字符都非常重要。为解决这个问题，有学者提出BiLSTM。BiLSTM本质上仍是一个循环神经网络，它将前向和后向LSTM网络连接在一起，同时考虑前后两个方向的内容信息来提高整个NLP模型的性能。BiLSTM的结构如图5所示。

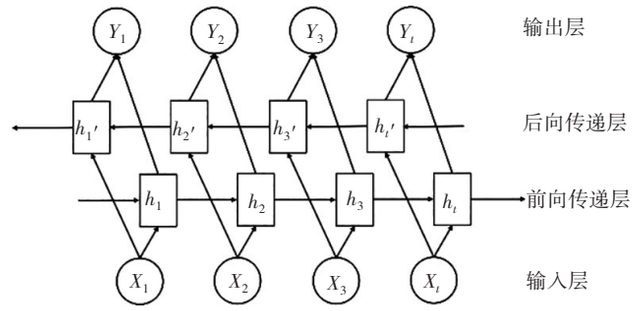


图5 BiLSTM结构图

Figure 5 BiLSTM model structure

1.3 注意力机制

2014年，Minh等[16]首次将注意力机制引入循环神经网络模型，并用于图像处理。随后，注意力机制逐渐被应用到NLP任务中。2017年，谷歌团队首次在文本表达中使用注意力机制[17]。该机制的工作原理是通过函数计算当前输入模块与整个输入信息之间的相似性，从而计算出每个输入模块对于当前输出的重要性，并将其作为权重赋予输入语句，最终得到注意力分布α_i，用于调整不同输入模块在生成输出时的贡献度，从而提高整个模型的性能。

注意力打分机制$f(Q,K_i)$的公式如下：

$$f(Q,K_i)\begin{cases} Q^{\mathrm{T}}K_i & \text{点积模型} \\ Q^{\mathrm{T}}\boldsymbol{W}K_i & \text{双线性模型} \\ \boldsymbol{W}[Q \cdot K_i] & \text{缩放点积模型} \\ \boldsymbol{V}^{\mathrm{T}}\tanh(\boldsymbol{W}Q + \boldsymbol{U}K_i) & \text{加性模型} \end{cases} \tag{7}$$

其中，K和Q表示键和查询，$\boldsymbol{V}$、$\boldsymbol{W}$和$\boldsymbol{U}$是需要通过网络训练学习得到的参数矩阵。

然后，使用softmax函数将其归一化得到概率分布，从而得到每个键的权重。具体公式如下：

$$\alpha_i = \mathrm{softmax}(f(Q,K_i)) = \frac{\exp(f(Q,K_i))}{\sum_j \exp(f(Q,K_i))} \tag{8}$$

最后，将权重和对应的值$\boldsymbol{V}$进行加权求和，得到最终输出，公式如下：

$$\mathrm{Attention}(Q,K,\boldsymbol{V}) = \sum_i \alpha_i \boldsymbol{V}_i \tag{9}$$

1.4 CRF

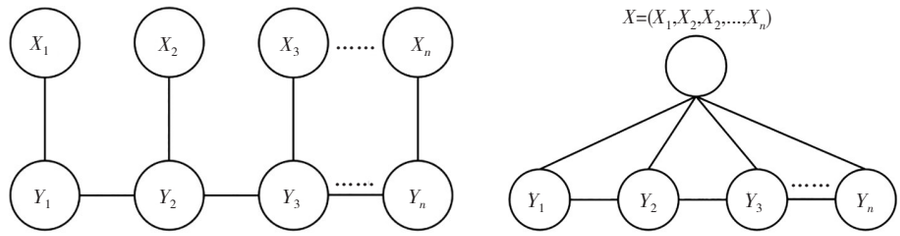
CRF是一种判别式概率无向图学习模型，它是在隐马尔可夫模型和最大熵模型的基础上发展而来的。CRF是一种非参数化的统计学习方法，它可以在标注和切分有序数据的条件概率模型中发挥重要作用，其一般定义如下[18]：设输入序列X和输出序列Y为随机变量，给定输入序列X的情况下，CRF可以计算输出序列Y的条件概率分布$P(Y|X)$。假设由随机变量Y构成随机无向图$G = (V,E)$，其中

$Y=\{Y_v|v\in V\}$是以图中节点v为索引的随机变量集合。在给定X的条件下，若每个随机变量Y_v都满足马尔可夫属性，即对于任意节点v均满足式(10)，则条件概率分布$P(Y|X)$被称为条件随机场。

$$P\left(Y_v|X,Y_w,w\neq v\right)=P(Y_v|X,Y_w,w\sim v) \tag{10}$$

其中，$w\sim v$表示与顶点v直接相邻的所有顶点w；$w\neq v$表示除顶点v以外的所有顶点；Y_v与Y_w为顶点v与w对应的随机变量。

在实际应用中，由于线性链CRF模型训练时间较短且操作便捷，因此其应用最为广泛。图6展示了两种主要的线性链条件随机场的图结构。



a：最大团是相邻两个顶点的集合　　b：最大团是相邻三个顶点的集合

图6 线性链的条件随机场图

Figure 6 Graphs of conditional random field of linear chain

在序列标注任务中，模型的观察序列通常使用$X=(X_1,X_2,\cdots,X_n)$表示，状态序列使用$Y=(Y_1,Y_2,\cdots,Y_n)$表示。在给定随机变量X取值为x的情况下，可以计算随机变量Y取值为y的条件概率分布$P(y|x)$，计算公式可参考式(11)：

$$P(y|x)=\frac{1}{Z(x)}\exp\left(\sum_{i,k}\lambda_k t_k\left(y_{i-1},y_i,x,i\right)+\sum_{i,l}u_l s_l(y_i,x,i)\right) \tag{11}$$

其中，t_k、s_l为特征函数，当特征条件被满足时才取值为1，否则为0；λ_k、u_l为对应权重。

$Z(x)$为归一化因子：

$$Z(x)=\sum_y\exp\left(\sum_{i,k}\lambda_k t_k\left(y_{i-1},y_i,x,i\right)+\sum_{i,l}u_l s_l(y_i,x,i)\right) \tag{12}$$

为简化上述公式，用一个统一的符号来表示转移特征、状态特征及其权重。式(13)表示为简化后的模型：

$$P\left(y|x\right)=\frac{1}{Z(x)}\exp\sum_{k=1}^{K}w_k f_k(y,x) \tag{13}$$

其中，

$$Z(x)=\sum_y\exp\left(\sum_{k=1}^{K}w_k f_k(y,x)\right) \tag{14}$$

通过式(15)计算条件概率：

$$L=\log\left(P(y|x)\right) \tag{15}$$

使用Viterbi算法来求解最大概率标签：

$$\hat{y}=\arg\max P(y|x) \tag{16}$$

2 实验结果与分析

2.1 实验数据

本研究基于美国临床试验注册中心的相关临床试验注册数据进行研究，这些研究项目由临床研究者们在世界各地进行，其中包括有关人类志愿者医学研究的多方面信息，如疾病、干预措施、研究的标题、试验设计、纳排标准以及进行研究的地点等。数据集共有4 000条临床试验标准，其中纳排标准数据达25 294条，训练集与验证集按8:2比例划分。

2.2 实验配置

本研究的实验环境采用Python语言以及开源的深度学习框架Pytorch。具体环境设置见表1。本研究使用BioBERT模型，该模型的隐藏层共有12层，每层有768个节点，并且使用12个注意力头。BioBERT模型使用GELU作为激活函数，BiLSTM的隐藏单元数为128。在训练过程中，BioBERT-Att-BiLSTM-CRF的最大序列长度为512，batch_size为192。此外，BioBERT学习率设置为5e-5，Dropout为0.1。

表1 实验环境配置

Table 1 Experimental environment configuration

项目	实验环境
操作系统	Windows10
CPU	i7-11370H@3.3 GHz
GPU	RTX3080(16 G)
Python版本	3.7.0
PyTorch框架	1.7.1

2.3 概念定义与标注文本

在临床试验纳排标准中，需要对特定意义的医学实体进行标注，如疾病名称“Gout”、治疗方式“Dialysis”以及过敏症“Quercetin”等。UMLS已被广泛应用于电子病历、临床研究以及文献分类等领域，旨在解决不同研究对医疗实体的定义和标注规则存在的差异问题。UMLS包含多种来自不同领域的医学词汇和术语，其中所包含的生物医学术语数量超过五百万，涵盖了至少两百万种医学概念。本研究考虑Zhang等[19]提出的医学实体标注规范，并结合UMLS定义的实体类别，通过对临床试验的专业知识和纳排标准的综合分析，最终定义年龄（Age）、疾病（Disease）、治疗方式（Treatment）、过敏症（Allergy）、性别（Gender）以及妊娠（Pregnancy）共6种类别的临床实体，表2列举了具体实体类型及其含义。

表2 实体类型及相关含义对照表

Table 2 Entity type and the corresponding meaning

序号	实体类别	实体含义
1	年龄	受试者的年龄要求
2	疾病	表示能被治疗的病因或医生对病人做出的诊断，包括常见疾病、综合征、中毒/受伤、器官/细胞受损等
3	治疗方式	表示用来预防、治疗及诊断疾病的化学物质，在UMLS中对应临床药物、糖皮质激素、疫苗、抗生素等
4	过敏症	受试者存在的过敏症
5	性别	受试者的性别要求
6	妊娠	受试者是否怀孕

命名实体识别的目的是识别出文本中的实体，需要对数据进行标注，本研究使用BIO标注方式进行标注。若标记非实体则标注为“O（Other）”，若为实体的第一个单词则标注为“B（Begin）”，若为同一实体的其余单词则标注为“I（Internal）”。实体类别缩写（如Dis、Gen、Pre等）接在“B”和“I”标签后，用连字符（或下划线）分割，表3是对临床纳排标准文本中预测标签的示例。

2.4 评价指标

为评估模型的性能，本研究采用精确率（Precision, P）、召回率（Recall, R）和F1值（F1-Score, F1）来评估模型的性能。其中，精确率表示正确识别的实体占识别出的实体总量的比例，召回率表示正确识别的实体占标准结果中实体总量的比例，F1值是精确率和召回率的调和平均值。各指标对应的计算公式如下：

表3 临床纳排标准实体预测标签定义

Table 3 Clinical eligibility criteria entity prediction label definitions

序号	实体类别	开始标签	中间标签
1	年龄	B-AGE	I-AGE
2	疾病	B-DIS	I-DIS
3	治疗方式	B-TRE	I-TRE
4	过敏症	B-ALL	I-ALL
5	性别	B-GEN	I-GEN
6	妊娠	B-PRE	I-PRE

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

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

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

其中，TP表示正确地预测为正例的实际正例样本数量数量，FP表示错误地预测为正例的实际负例样本数量，FN则表示错误地预测为负例的实际正例样本数量。

2.5 对比实验

为验证本研究提出的BioBERT-Att-BiLSTM-CRF模型在临床试验纳排标准实体识别任务中的识别效果，设计了以下几种方法进行对比实验：(1)BiLSTM-CRF模型，输入为word2vec训练得到的词向量，经过BiLSTM层后输出每个标记的概率，最后通过CRF层进行实体识别。该模型在中英文生物医学实体识别任务中被广泛应用，取得了良好的效果。(2)Att-BiLSTM-CRF模型，引入注意力机制以确保模型能关注标记本研究中同一token的多个实例之间的一致性。(3)BERT-BiLSTM-CRF模型，使用BERT预训练模型提取句子特征，再将获取的特征与经典BiLSTM-CRF模型相结合。(4)BERT-Att-BiLSTM-CRF模型，在上个模型的基础上引入注意力机制，从而更好地利用上下文信息和全局信息。(5)BioBERT-Att-BiLSTM-CRF模型，采用BioBERT替换上个模型中的BERT，与现有方法相比，该模型被证实对公共医疗数据集的实体识别具有较好的表现。

2.6 结果分析

各模型对比结果见表4。整体实验结果表明BioBERT-Att-BiLSTM-CRF模型在临床试验纳排标准实体识别中表现最佳。相较BiLSTM-CRF，Att-BiLSTM-CRF模型的精确率提高2.59%，召回率提高2.84%，F1值提高2.72%，表明该模型可以通过引入注

意力机制来提高整个识别任务的准确率，从而更好地捕捉上下文关系和实体特征。BERT预训练语言模型的优越性在于其能更好地学习上下文信息，从而提高模型的泛化能力。因此，在实体识别任务中引入注意力机制和BERT预训练语言模型可以相互协作，提高模型的性能和准确率，从而更好地处理NLP中的实际问题。比较BERT-Att-BiLSTM-CRF和BioBERT-Att-BiLSTM-CRF模型的表现，前者的精确率、召回率和F1值分别是76.43%、76.95%和76.69%，后者分别为77.51%、77.30%和77.40%。相较BERT，基于BioBERT的实体识别模型效果更出色。因此，在进行多轮训练时，应选择BioBERT模型作为词嵌入层，并引入注意力机制，以提高模型效果。

表4 各模型整体对比结果(%)

Table 4 Comparison among different models (%)

模型	精确率	召回率	F1值
BiLSTM-CRF	70.65	69.52	70.08
Att-BiLSTM-CRF	73.24	72.36	72.80
BERT-BiLSTM-CRF	75.32	74.14	74.73
BERT-Att-BiLSTM-CRF	76.43	76.95	76.69
BioBERT-Att-BiLSTM-CRF	77.51	77.30	77.40

训练完成后，该模型被用于临床试验筛选方案实体识别任务中，以自动并高效地识别其中的实体。该模型对不同实体类型的识别结果如表5所示，整体而言，该模型在临床病历命名实体识别任务中展现了相对均衡的综合性能，总体F1值为77.40%。其中年龄和性别实体类型的表现相对较为显著，F1值分别达到82.82%和86.29%，疾病和治疗方式实体类型的性能也较为良好，F1值分别达到77.18%和77.23%。

表5 BioBERT-Att-BiLSTM-CRF对不同实体类型的识别结果(%)

Table 5 Results of different entity types identified using BioBERT-Att-BiLSTM-CRF (%)

实体类型	精确率	召回率	F1值
年龄	83.32	82.33	82.82
疾病	80.36	74.25	77.18
治疗方式	72.88	82.14	77.23
过敏症	71.56	71.61	71.58
性别	86.52	86.06	86.29
妊娠	70.42	67.41	68.88
总体	77.51	77.30	77.40

3 讨 论

过去的研究主要集中在临床病例数据集，例如孙安等[20]、张柏嘉[21]、唐国强等[22]、曹春萍等[23]以及万泽宇等[24]的工作，均专注于命名实体识别任务，提出了多样化的方法和技术以改善模型性能。他们解决了实体知识边界划分不明确、复合实体知识识别困难以及学习标签的依赖关系等问题，从而改善医学领域复合实体知识识别的效果，为临床病历文本中的复合实体识别提供可借鉴的方法。另一方面，蔡晓琼等[25]尝试对COVID-19临床文本进行命名实体识别，但其数据仅限于COVID-19的临床试验注册记录中的摘要文本，且存在实体种类多但数量不均衡的问题。

与之前的研究相比，本研究关注临床试验纳排标准。这种聚焦相较于临床病例数据集更精准，相比摘要文本能提供更多的信息。进一步而言，基于纳排标准的命名实体识别，能构建纳排标准知识图谱，为医疗工作者制定纳排标准提供有效可靠的方案。本研究将提升医学实体识别的水平，并通过提高纳排标准的精度，进一步推动临床试验的科学性和准确性。

4 结 论

本研究提出一种医学实体识别模型，采用基于BioBERT预训练语言模型的词嵌入技术，并融合了BiLSTM和CRF序列标注方法，引入注意力机制，可用于识别临床试验纳排标准中的新兴医学实体。多组实验对比验证了本研究方法的有效性。结果表明该模型的识别性能优于基准模型和主流预训练模型的实体识别方法，同时也能有效地完成相关实体的识别任务。

在接下来的工作中，将在纳排标准命名实体识别的基础上进行关系抽取，并设计纳排标准领域的知识图谱，从而为后续的纳排标准辅助决策提供支持。此外，本研究的文本分析主要基于英文，对中文文本的处理仍有待提升，可以借鉴现有的中文文本处理技术。本研究为临床试验纳排标准的自动化处理提供了有力支持，并为未来的相关研究提供了新的发展方向。

【参考文献】

[1] Laine C, Horton R, Angelis C, et al. Clinical trial registration: looking back and moving ahead[J]. N Engl J Med, 2007, 356(26): 2734-2736.
[2] Shah P, Kendall F, Khozin S, et al. Artificial intelligence and machine learning in clinical development: a translational perspective[J]. NPJ Digit Med, 2019, 2(1): 69.
[3] 王怡, 白雪, 崔胜男, 等. 临床医学命名实体识别的病历质量筛选标

准研究[J]. 中国卫生质量管理, 2018, 25(6): 34-36.
Wang Y, Bai X, Cui SN, et al. A study on medical record quality screening criteria for clinical medicine named entity identification[J]. China Health Quality Management, 2018, 25(6): 34-36.
[4] Coletti MH, Bleich HL. Medical subject headings used to search the biomedical literature[J]. J Am Med Inform Assoc, 2001, 8(4): 317-323.
[5] Schuemie MJ, Mons B, Weeber M, et al. Evaluation of techniques for increasing recall in a dictionary approach to gene and protein name identification[J]. J Biomed Inform, 2007, 40(3): 316-324.
[6] Blaschke C, Valencia A. The frame-based module of the SUISEKI information extraction system[J]. IEEE Intell Syst, 2002, 17(2): 14-20.
[7] Corney DP, Buxton BF, Langdon W, et al. BioRAT: extracting biological information from full-length papers[J]. Bioinformatics, 2004, 20(17): 3206-3213.
[8] Fundel K, Küffner R, Zimmer R. RelEx-relation extraction using dependency parse trees[J]. Bioinformatics, 2007, 23(3): 365-371.
[9] Weng C, Wu X, Luo Z, et al. EliXR: an approach to eligibility criteria extraction and representation[J]. J Am Med Inform Assoc, 2011, 18 (Suppl 1): i116-i124.
[10] Boland MR, Tu SW, Carini S, et al. EliXR-TIME: a temporal knowledge representation for clinical research eligibility criteria[J]. AMIA Jt Summits Transl Sci Proc, 2012, 2012: 71-80.
[11] Kang T, Zhang S, Tang Y, et al. EliIE: an open-source information extraction system for clinical trial eligibility criteria[J]. J Am Med Inform Assoc, 2017, 24(6): 1062-1071.
[12] Yuan C, Ryan PB, Ta C, et al. Criteria2Query: a natural language interface to clinical databases for cohort definition[J]. J Am Med Inform Assoc, 2019, 26(4): 294-305.
[13] Tseo Y, Salkola M, Mohamed A, et al. Information extraction of clinical trial eligibility criteria[J]. arXiv reprient arXiv:2006. 07296v6, 2020.
[14] Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining[J]. Bioinformatics, 2020, 36(4): 1234-1240.
[15] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural Comput, 1997, 9(8): 1735-1780.
[16] Mnih V, Heess N, Graves A. Recurrent models of visual attention[J]. arXiv reprient arXiv:1406.6247, 2014.
[17] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. arXiv reprient arXiv:1706.03762, 2017.
[18] Lafferty JD, Mccallum A, Pereira FC. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]// Proceedings of the Eighteenth International Conference on Machine Learning. Morgan Kaufmann Publishers Inc. 2001: 282-289.
[19] Zhang H, Zong Y, Chang B, et al. Medical entity annotation standard for medical text processing[C]//Proceedings of the 19th Chinese National Conference on Computational Linguistics. 2020: 561-571.
[20] 孙安, 于英香, 罗永刚, 等. 序列标注模型中的字粒度特征提取方案研究-以CCKS2017: Task2临床病历命名实体识别任务为例[J]. 图书情报工作, 2018, 62(11): 103-111.
Sun A, Yu YX, Luo YG, et al. Research on word granularity feature extraction scheme in sequence annotation model-taking CCKS2017: Task2 clinical medical record named entity recognition task as an example[J]. Library Intelligence Work, 2018, 62(11): 103-111.
[21] 张柏嘉. 临床心脏病医疗文本命名实体识别方法研究[D]. 哈尔滨: 哈尔滨工程大学, 2020.
Zhang BJ. Research on named entity recognition method for clinical cardiology medical texts[D]. Harbin: Harbin Engineering University, 2020.
[22] 唐国强, 高大启, 阮彤, 等. 融入语言模型和注意力机制的临床电子病历命名实体识别[J]. 计算机科学, 2020, 47(3): 211-216.
Tang GQ, Gao DQ, Ruan T, et al. Named entity recognition for clinical electronic medical records incorporating language models and attention mechanisms[J]. Computer Science, 2020, 47(3): 211-216.
[23] 曹春萍, 关鹏举. 基于E-CNN和BLSTM-CRF的临床文本命名实体识别[J]. 计算机应用研究, 2019, 36(12): 3748-3751.
Cao CP, Guan PJ. Named entity recognition of clinical text based on E-CNN and BLSTM-CRF[J]. Computer Application Research, 2019, 36(12): 3748-3751.
[24] 万泽宇, 龚庆悦, 李铁军, 等. 基于自适应词嵌入RoBERTa-wwm的名中医临床病历命名实体识别研究[J]. 软件导刊, 2022, 21(12): 58-62.
Wan ZY, Gong QY, Li TJ, et al. A study on named entity recognition of famous Chinese medicine clinical records based on adaptive word embedding RoBERTa-wwm[J]. Software Guide, 2022, 21(12): 58-62.
[25] 蔡晓琼, 郑增亮, 苏前敏, 等. 基于MPNet与BiLSTM的COVID-19临床文本命名实体识别方法[J]. 智能计算机与应用, 2023, 13(1): 164-170.
Cai XQ, Zheng ZL, Su QM, et al. An MPNet and BiLSTM based method for COVID-19 clinical text named entity recognition[J]. Intelligent Computers and Applications, 2023, 13(1): 164-170.

（编辑：谭斯允）