DOI:10.19652/j.cnki.femt.2305349

MRC-PBM:一种中文电子病历嵌套命名实体识别方法*

周佳伦 李琳宇 马洪彬 姜艳静
(三峡大学计算机与信息学院 宜昌 443000)

摘 要:中文电子病历实体包含大量的医学领域词汇并具有明显的嵌套特征。嵌套实体识别时往往存在目标实体定位不完整、不准确的问题。针对这一问题,提出了一种基于机器阅读理解的中文电子病历嵌套命名实体识别模型 MRC-PBM(machine reading comprehension-position information biaffine and MLP)。该模型将命名实体识别(named entity recognition,NER)转化为机器阅读理解任务,将中文电子病历文本和预定义的查询语句串联作为输入,使用基于医学的预训练模型 MC_BERT 获取词向量,然后通过双向长短期记忆网络模型(BiLSTM)和多粒度扩张卷积模型分别获取双向的特征信息以及单词之间的信息,得到相应的特征向量,最后使用 Hybrid-PBM 预测器进行实体预测。在嵌套和平面 NER 数据集上进行实验。实验表明,该模型在糖尿病语料和公开医学数据集上优于其他主流神经网络模型,$F1$ 值比基线模型提高了 1.21%~5.80%。

关键词:中文电子病历;命名实体识别;机器阅读理解;嵌套实体
中图分类号:TP391.1 **文献标识码**:A **国家标准学科分类代码**:510.4

MRC-PBM: A Chinese electronic medical Record nested named entity recognition method

Zhou Jialun Li Linyu Ma Hongbin Jiang Yanjing
(College of Computer and Information, China Three Gorges University, Yichang 443000, China)

Abstract: The Chinese electronic medical record entities contain a large number of medical domain vocabulary and have obvious nested features. When identifying nested entities, there is often a problem of incomplete or inaccurate location of the target entity. To address this problem, a Chinese electronic medical record nested named entity recognition model machine reading comprehension-position information biaffine and MLP(MRC-PBM), based on MRC is proposed. The model transforms named entity recognition (NER) into an MRC task, concatenating the Chinese EMR text and predefined query statements as input, using the medical-based pre-trained model MC_BERT to obtain word vectors, and then using a bidirectional long short-term memory network (BiLSTM) and a multi-granularity expansion convolution model to obtain bidirectional feature information and information between words, respectively, to obtain corresponding feature vectors. Finally, the Hybrid-PBM predictor is used to predict the entities. Experiments are conducted on nested and flat NER datasets. The experimental results show that the proposed model outperforms other mainstream neural network models on the diabetes corpus and public medical datasets, with $F1$ scores improved by 1.21% to 5.80% compared to baseline models.

Keywords: Chinese electronic medical record; machine reading comprehension; named entity recognition; nested entity

0 引 言

命名实体识别(named entity recognition,NER)技术旨在从非结构化文本中提取目标实体,是自然语言处理中的一项基本性任务[1]。根据实体中是否存在嵌套实体的情况,NER 可分为平面 NER 和嵌套 NER[2]。其中嵌套

收稿日期:2023-07-25
* **基金项目**:国家重点研究发展计划项目(2016YFC0802500)资助

NER 主要有 3 种类型：1）实体中包含一个或多个平面实体的实体；2）包含多个实体类型的实体；3）不连续实体。本文主要研究第 2 种嵌套实体类型，如在句子“其他症状心脏损害并不少见…”中，实体“心脏损害”（症状）包含了一个内嵌实体“心脏”（身体部位），两个实体共用了同一部分文本“心脏”。

中文电子病历（Chinese electronic medical records，CEMR）是医疗记录中的重要组成部分，作为患者临床信息的主要载体，其包含大量的医学信息，可为医生的后续诊断、治疗提供有价值的参考和依据，并且对于医学研究也具有重要的意义[3]。中文电子病历中含有较多的生物医药类词语，其中包含大量复杂的嵌套命名实体，通过识别这些嵌套命名实体可以获取更细粒度的信息[4]。

中文电子病历领域中，命名实体识别的研究主要集中与平面命名实体，如 Chen 等[5]提出了一种基于医学 MC_BERT 的混合神经网络模型，通过 MC_BERT 预训练模型来表达医学领域的相关词汇信息，使用深度网络模型充分提取复杂文本中丰富的实体特征。李博等[6]提出一种完全基于注意力机制的神经网络模型，该模型对 Transformer 模型进行训练优化，使用条件随机场对医学文本特征进行分类识别。Chiu 等[7]提出卷积神经网络＋双向长短期记忆（CNN＋BiLSTM）的神经网络结构，该模型可以自动提取单词级和字符级的特征。顾佼佼等[8]提出了一种基于 BERT 和知识蒸馏的命名实体识别模型，对字符融合远程标签词边界特征得到特征融合向量，送入 BERT 生成动态字向量表示，大大提高了模型效率。上述研究在平面 NER 任务中取得了不错的效果，然而，在实际应用中，文本中的实体往往具有嵌套结构。为了解决嵌套 NER 问题，研究人员提出了多种方法，如 Ju 等[9]提出了第 1 个神经分层模型，通过内外动态叠加平面 NER 层来识别嵌套实体。Jiang 等[10]提出了一种新的基于机器阅读理解（machine reading comprehension，MRC）的高性能实体提取模型，通过 MRC 模式来提取嵌套实体，并设计 2D pro-encoding 模块加强实体定位，从而得到较好效果。Yu 等[11]提出了一种基于跨度的方法，通过 Biaffine 解码器来获取每个跨度对应的实体得分矩阵。Li 等[12]提出了一种统一的 NER 框架，将 NER 任务转化成为一种词对的关系分类任务，充分考虑词与词之间的关系，从而有效地处理平面和嵌套实体。但是这些方法不适用于中文医学领域且在识别嵌套命名实体时存在实体识别不准确、遗漏等问题。为了解决上述问题，本文提出了一种新的中文电子病历命名实体识别模型 MRC-PBM（position information biaffine and MLP）。该模型不同于基于序列标注任务的解决方法，是将命名实体识别任务转换为机器阅读理解的片段抽取任务，通过多轮问答，从而将平面和嵌套实体全部提取出来。在编码阶段，本文使用医学领域的预训练模型 MC_BERT 获取词向量，并使用 BiLSTM 和多粒度扩张卷积网络从而更好地提取特征向量。在结果预测阶段，本文设计了一种包含实体相对位置信息的预测器 Hybrid-PBM，进一步提高模型识别嵌套实体的能力。

1 基于机器阅读理解的命名实体识别方法

为了解决实体预测阶段目标实体定位不完整或不准确的问题，本文提出了 MRC-PBM 模型，该模型主要由编码层、Hybrid-PBM 预测层以及输出层 3 个部分组成，如图 1 所示。通过编码层获取词嵌入，然后使用 Hybrid-PBM 预测器来获取打分矩阵，最后通过输出层得到最终的 NER 结果。

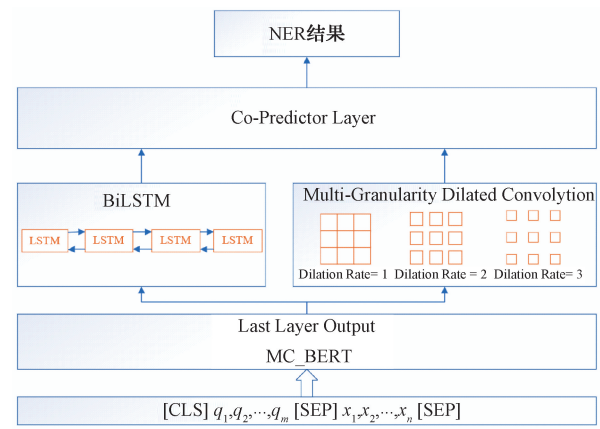

图 1 模型架构

1.1 编码层

编码层使用机器阅读理解框架，因此需要提前为所有目标实体类型设计查询问题。例如，当从中文电子病历中提取问诊科室类型的实体时，查询问题可以为“从文本中查询出科室，包括眼科、儿科、外科、内科、急诊室等”。将预定义查询问题 $Question=\{q_1,q_2,\cdots,q_m\}$ 和文本句子 $Context=\{x_1,x_2,\cdots,x_n\}$ 的串联作为预训练模块的输入。其中，$Question$ 表示目标实体类型的查询问题，m 表示生成查询问题的长度，$Context$ 表示输入的文本句子，n 表示序列的长度。查询问题的作用是为模型提供额外的实体类型信息，从而在后续的编码过程中，能够充分地提取实体信息。因此本文使用关键字来代替完整语义的方法，如“科室、眼科、儿科、外科、内科、急诊室”这样的格式，以达到减小输入长度和降低模型的计算量。

1）MC_BERT 模型

MC_BERT 以 Bert 模型[13]提为基础，并使用医学数据进行了训练，提升了 MC_BERT 模型在医学领域的表现。在本文中，本文使用了 MC_BERT[14]作为预训练模型。由于生物医学领域的文本在结构和词汇分布上与通用领域有很大差异，因此 MC_BERT 采用了全实体和全跨度的掩码预测方式，遮盖类似于“胃疼”的医疗实体以及“胃部一阵一阵疼”，“胃不舒服有点痛”等与“胃疼”意思相近的短语，以引入生物医学相关知识。本文选择 MC_

BERT 模型的最后一个隐藏层表示作为该模型的输出 $\boldsymbol{H}_0$：

$$\boldsymbol{H}_0 = MC\_BERT(x_i)[-1] \in \boldsymbol{R}^{n\times d} \tag{1}$$

式中：d 是隐藏大小。

2)BiLSTM 模型

长短期记忆（LSTM）是一种特殊的循环神经网络（RNN）模型，相较于一般的 RNN 结构，LSTM 引入了输入门、遗忘门和输出门 3 个门结构，从而实现了一种特定的学习机制，学习到的哪些信息需要被记忆、更新以及注意[15]。通过这种机制，LSTM 模型可以获取更多的有用信息，并有效地解决了长文本序列训练中的梯度消失和梯度爆炸问题。LSTM 中神经元的计算过程如下：

$$f_t = \sigma(\boldsymbol{w}_f h_{t-1} + \boldsymbol{u}_f x_t) \tag{2}$$

$$i_t = \sigma(\boldsymbol{w}_i h_{t-1} + \boldsymbol{u}_i x_t) \tag{3}$$

$$o_t = \sigma(\boldsymbol{w}_o h_{t-1} + \boldsymbol{u}_o x_t) \tag{4}$$

$$c_t = f_t \odot c_{t-1} + i_t \odot \tanh(\boldsymbol{w}_c h_{t-1} + \boldsymbol{u}_c x_t) \tag{5}$$

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

式中：f_t 表示遗忘门；i_t 表示输入门；o_t 表示输出门；c_t 表示存储单元；h_t 表示隐藏层；σ 是 sigmoid 激活函数；$\boldsymbol{w}_f$、$\boldsymbol{w}_i$、$\boldsymbol{w}_o$、$\boldsymbol{w}_c$、$\boldsymbol{u}_f$、$\boldsymbol{u}_i$、$\boldsymbol{u}_o$ 和 $\boldsymbol{u}_c$ 表示在训练过程中学习的权重矩阵；$\odot$ 表示点积运算。

BiLSTM 是前向 LSTM 和后向 LSTM 组成的模型，通过双向学习文本信息，在句子层面上更好地捕捉到较远距离的语义依赖关系[16]。为了进一步增强上下文建模，本文采用 BiLSTM 生成特征表示，即 $\boldsymbol{H} = \{h_1, h_2, \cdots, h_n\} \in \boldsymbol{R}^{n\times d_h}$，其中 d_h 是特征表示的维度。

3)多粒度扩张卷积模型

CNN 主要由输入层、卷积层、池化层和全连接层组成。其中卷积层是 CNN 的核心层，通过使用多个大小不一的卷积核并行处理文本特征，从而提高特征向量的计算效率。池化层使用最大池化操作提取卷积层中最重要的特征表示，进而得到基于 CNN 的文本特征向量[17]。

本文使用多个二维扩张卷积（dilated convolution，DConv），每个 DConv 具有不同的扩张率 L，以捕捉不同距离单词之间的相互作用，从而获取更多单词之间的信息，提高预测的准确性。每个扩张卷积计算公式如下：

$$\boldsymbol{Q}^L = \sigma_1(DConv_L(\boldsymbol{H}_0)) \tag{7}$$

其中，$\boldsymbol{Q}^L \in \boldsymbol{R}^{n\times n\times d_h}$，表示扩张卷积与扩张率 L 的输出，σ_1 是 GELU 激活函数[18]。如果 $L \in [1,2,3]$，本文就可以得到最终的特征表示 $\boldsymbol{Q} = [Q^1, Q^2, Q^3] \in \boldsymbol{R}^{n\times n\times 3d_h}$。

1.2 Hybrid-PBM 预测器

Hybrid-PBM 预测模块致力于进一步提高模型识别嵌套实体的能力，Hybrid-PBM 预测模块结构如图 2 所示。

首先，该模块分别从 BiLSTM 和多粒度扩张卷积模型获取特征表示 B 和 Q。将特征表示 B 输入到两个单独的前馈神经网络（FFNN）为跨度的开始/结束创建不同的

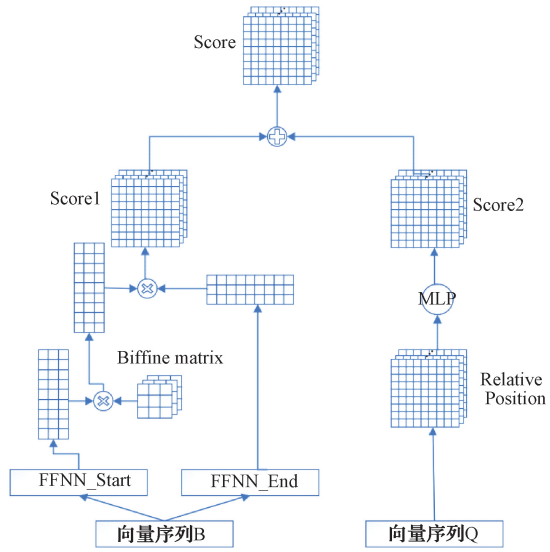

图 2 联合编码模型结构

表示 H_s/H_e，使模型能够分别识别跨度的开始和结束，从而提高准确性。在句子上使用 Biaffine 模型来创建一个评分矩阵 Score_1。跨度 i 的分数为：

$$\boldsymbol{H}_s(i) = FFNN_s(x_{si}) \tag{8}$$

$$\boldsymbol{H}_e(i) = FFNN_e(x_{ei}) \tag{9}$$

$$\boldsymbol{Score}_1(i) = \boldsymbol{H}_s(i)^{\mathrm{T}}\boldsymbol{U}_m\boldsymbol{H}_e(i) + W_m(\boldsymbol{H}_s(i) \oplus \boldsymbol{H}_e(i)) + \boldsymbol{B}_m \tag{10}$$

式中：si 和 ei 是跨度 i 的开始和结束索引；$\boldsymbol{U}_m \in \boldsymbol{R}^{k\times c\times k}$；$\boldsymbol{W}_m \in \boldsymbol{R}^{2k\times c}$；$k$ 为 FFNN 输出层的大小；c 为实体类型的数量；$\boldsymbol{Score}_1(i) \in \boldsymbol{R}^{n\times n\times k}$；$\boldsymbol{B}_m$ 是偏差。

其次，本文引入实体相对位置信息组件到预测器中，以提高实体预测的准确度。例如，当输入文本为“二甲双胍：0.25 g、格列苯脲：2.5 mg、阿卡波糖：50 mg”时，需要识别出多个实体，如果没有输入实体相对位置信息，预测器对实体的长度和跨度都不够敏感，容易将任意两个实体的首尾组合预测为一个目标实体，例如预测出“二甲双胍：0.25 g、格列苯脲”这样的实体。相反，如果有实体相对位置信息，预测器就能对实体的长度和跨度更敏感，从而更准确地识别出真正的实体。因此，本文将特征表示 $\boldsymbol{Q} = [q_1, q_2, \cdots, q_n]$ 通过变换公式 $\boldsymbol{p}_{si,\partial} = W_{p,\partial} \cdot q_{si} + b_{p,\partial}$ 和 $k_{si,\partial} = W_{k,\partial} \cdot q_{si} + b_{k,\partial}$，得到向量序列 $\boldsymbol{p} = [p_{1,\partial}, p_{2,\partial}, \cdots, p_{n,\partial}]$ 和 $\boldsymbol{k} = [k_{1,\partial}, k_{2,\partial}, \cdots, k_{n,\partial}]$，它们是识别第 ∂ 种类型实体所用的向量序列。本文通过式(11)计算跨度 I 且类型为 ∂ 的实体打分。

$$\boldsymbol{s}_\partial(si, ei) = \boldsymbol{p}^{\mathrm{T}}_{si,\partial} \cdot \boldsymbol{k}_{ei,\partial} \tag{11}$$

然后使用一个变换矩阵 $\boldsymbol{R}_{si}$，满足 $\boldsymbol{R}^{\mathrm{T}}_{si} \cdot \boldsymbol{R}_{ei} = \boldsymbol{R}_{ei-si}$[19] 分别应用到 $\boldsymbol{p}$、$\boldsymbol{k}$ 中，如式(12)所示，从而显式地往打分 $\boldsymbol{s}_\partial(si, ei)$ 中注入实体相对位置信息。

$$\boldsymbol{s}_\partial(si, ei) = (\boldsymbol{R}_{si} \cdot \boldsymbol{p}_{si,\partial})^{\mathrm{T}} \cdot (\boldsymbol{R}_{ei} \cdot \boldsymbol{k}_{ei,\partial}) = \boldsymbol{p}^{\mathrm{T}}_{si,\partial} \cdot \boldsymbol{R}^{\mathrm{T}}_{si} \cdot \boldsymbol{R}_{ei} \cdot \boldsymbol{k}_{ei,\partial} = \boldsymbol{p}^{\mathrm{T}}_{si,\partial} \cdot \boldsymbol{R}_{ei-si} \cdot \boldsymbol{k}_{ei,\partial} \tag{12}$$

本文将最终的打分矩阵 s_3 输入到多层感知器(MLP)中计算特征表示 Q 所对应的打分矩阵 $\boldsymbol{Score}_2$:

$$\boldsymbol{Score}_2 = MLP(S_3) \tag{13}$$

其中,$\boldsymbol{Score}_2 \in \boldsymbol{R}^{n\times n\times k}$。

最后,本文将打分矩阵 $\boldsymbol{Score}_1$ 和 $\boldsymbol{Score}_2$ 相加得到最终的打分矩阵 $\boldsymbol{Score}$:

$$\boldsymbol{Score} = \boldsymbol{Score}_1 + \boldsymbol{Score}_2 \tag{14}$$

其中,$\boldsymbol{Score} \in \boldsymbol{R}^{n\times n\times k}$。

1.3 输出层

本文使用感知器对不同跨度进行预测:

$$\boldsymbol{P} = Sigmoid(\boldsymbol{w}_0 \cdot \boldsymbol{Score} + \boldsymbol{b}) \tag{15}$$

其中,$\boldsymbol{w}_0 \in \boldsymbol{R}^{c\times k}$,$\boldsymbol{b} \in \boldsymbol{R}^{c}$,$\boldsymbol{P} \in \boldsymbol{R}^{n\times n\times c}$。

在训练时,本文对开始和结束指数的预测有两个损失:

$$L_{\text{start}} = CE(P_{\text{start}}, Y_{\text{start}}) \tag{16}$$

$$L_{\text{end}} = CE(P_{\text{end}}, Y_{\text{end}}) \tag{17}$$

开始-结束索引匹配损失为:

$$L_{\text{span}} = CE(P_{\text{start,end}}, Y_{\text{start,end}}) \tag{18}$$

总体训练目标最小化如下:

$$L = \alpha L_{\text{start}} + \beta L_{\text{end}} + \gamma L_{\text{span}} \tag{19}$$

其中,L 为最小化的总体训练目标,α、β、$\gamma \in [0,1]$ 是超参数控制对整体训练目标的贡献。P_{start} 和 P_{end} 分别表示每个字符作为与查询问题相关的开始索引和结束索引的概率分布。Y 是所有可能的标签类型的预定义列表(例如,sym、bod 等),Y_{start} 和 Y_{end} 分别作为训练集中开始索引和结束索引的字符的标签。CE 为交叉熵损失函数。

2 实 验

2.1 实验数据集

为了评估本文的模型,本文在 3 个数据集上进行了实验,验证模型的实体提取能力和泛化能力。

糖尿病(Diabetes)数据集是来自 1 120 份某医院糖尿病患者的电子病历。数据集信息以及实体信息,如表 1、2 所示。此外,Diabetes 数据集不包含嵌套实体。

表 1 Diabetes 数据集信息

数据集	数据量	实体类型
训练集	900	并发症(Complication)、疾病(Disease)、检查项目(Test)、检查数值(Test_value)、药物(Drug)、自我检查症状(self-check symptoms)、诊断症状(Examination symptoms)、手术(Operation)
验证集	220	

CMeEE 数据集是由阿里云天池承办的中文医疗信息处理挑战榜 CBLUE[20](https://tianchi.aliyun.com/dataset/95414)中命名实体识别任务所使用的数据集,是中文医疗文本作为自然语言处理任务的标准数据集,如表 3 所示。

表 2 Diabetes 实体信息

实体名称	例子	数量	平均长度
并发症	糖尿病视网膜病变检查及化验	1 082	6.98
疾病	糖尿病控制不佳,住院调理	478	4.18
检查数值	眼底检查:视力 右 0.4 左 0.2	206	7.07
药物	体检:血压 200/100 mmHg	235	7.02
自我症状	用胰岛素有一年时间	396	14.89
诊断症状	视力下降比较严重	805	11.28
诊断症状	做检查眼底出血,玻璃体积血	647	10.61
手术	特别应该打激光手术	237	6.12

表 3 CMeEE 数据集信息

数据集	数据量	实体类型
训练集	16 000	身体(Bod)、科室(Dep)、疾病(Dis)、药物(Dru)、医疗设备(Equ)、医疗程序(Ite)、微生物类(Mic)、医学检验项目(Pro)、临床表现(Sym)
验证集	4 000	

CMeEE_Nested 数据集是在 CMeEE 数据集基础上通过脚本提取出的 2 000 条包含嵌套实体的文本,目的是检验模型处理嵌套实体的能力,如表 4、5 所示。

表 4 CMeEE_Nested 数据集信息

数据集	数据量	实体类型
训练集	1 600	身体(Bod)、科室(Dep)、疾病(Dis)、药物(Dru)、医疗设备(Equ)、医疗程序(Ite)、微生物类(Mic)、医学检验项目(Pro)、临床表现(Sym)
验证集	400	

表 5 CMeEE_Nested 实体信息

实体名称	例子	数量	平均长度
疾病	四环素族:加重氮质血症	1 024	9.58
临床表现	咽肌痉挛、进行性瘫痪为特征	1 982	21.60
医疗程序	胰高血糖素试验血糖反应正常	441	6.20
医疗设备	CT 见肿瘤为高密度	59	3.24
药物	饮食受限制导致维生素 A 缺乏	144	8.06
医学检验	颅内压增高中枢神经系统肿瘤	363	5.03
身体	交感神经受累可致唾液分泌	1 815	5.92
科室	ICU 集中了各种诊疗仪器	14	2.14
微生物类	母亲孕早期感染巨细胞病毒	63	4

2.2 评估指标

本文采用精确度(precision)、召回率(recall)和 $F1$ 分数作为评估指标用于衡量平面和嵌套 NER 的性能，计算公式如下：

$$Precision = \frac{TP}{TP + FP} \tag{20}$$

$$Recall = \frac{TP}{TP + FN} \tag{21}$$

$$F1 = 2 \cdot Precision \cdot \frac{Recall}{Precision + Recall} \tag{22}$$

式中：TP 为真正类；FP 为假正类；FN 为假负类。

2.3 参数设置

所有实验均在 Tesla T4 16 G 的服务器上进行，实验超参数设置如表 6 所示。

表 6 实验超参数

参数	Diabetes	CMeEE
Epoch	150	30
Learning rate	8×10^{-6}	8×10^{-6}
Dropout	0.2	0.2
Batch_size	8	4
BiLSTM_size	768	768
CNN_size	512	512
Biaffine_size	512	512
Inner_size	64	64

2.4 对比实验

在 Diabetes 数据集、CMeEE 数据集和 CMeEE_Nested 数据集上进行对比实验，以验证 MRC_PBM 模型的泛化能力和整体性能，实验结果如表 7 所示。其中比较基线模型都是具有代表性的模型，如 Bert_Seq2Seq 模型、MRC_Baffine 模型等。

表 7 模型结果 (%)

数据集	模型	精确度	召回率	$F1$ 值
Diabetes	Bert_Seq2Seq[21]	86.09	83.38	84.71
	MRC_Baffine[22]	89.23	89.44	89.34
	Bert_MRC[23]	89.02	88.49	88.75
	MRC-PBM	90.51	90.51	90.51
CMeEE	Bert_Seq2Seq	57.78	59.80	57.78
	MRC_Baffine	61.49	61.78	61.63
	Bert_MRC	61.72	61.39	61.55
	MRC-PBM	64.59	61.01	62.76
CMeEE_Nested	Bert_Seq2Seq	69.57	45.26	54.84
	MRC_Baffine	61.51	63.83	62.65
	Bert_MRC	61.42	62.67	62.04
	MRC-PBM	62.46	63.21	62.83

从表 7 和图 3 可看出，本文提出的 MRC-PBM 模型的实体提取性能超过了其他 3 种基线模型，其 $F1$ 值与基线模型相比有 0.79%～7.99%的提高。在 3 种数据集中，基于序列标注方法的 Bert_Seq2Seq 模型表现都不如基于机器阅读理解框架的模型。这主要是因为本文将预定义的查询问题与文本句子串联在一起，为模型提供了额外的实体类型信息。基于机器阅读理解框架不仅提高了模型识别实体的能力，而且在 Diabetes 数据集数据量有限的情况下仍有出色的表现，充分彰显了基于机器阅读理解框架的优越性。

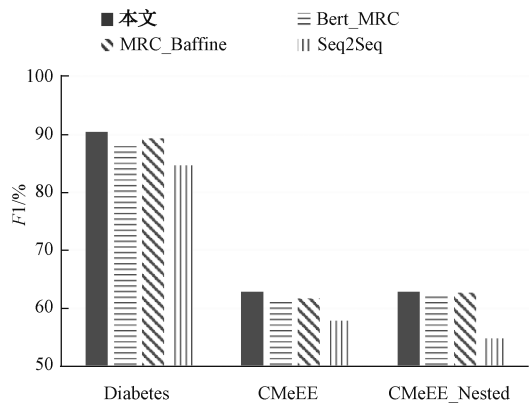

图 3 模型在数据集上的 $F1$ 值结果

在 CMeEE 和 CMeEE_Nested 两个数据集中，Bert_MRC 作为一种新兴的嵌套 NER 模型，$F1$ 值分别达到 61.55%和 62.04%。MRC_Baffine 模型在召回率和 $F1$ 分数上略优于 BERT_MRC 模型，提高了 0.39%、0.08%；1.16%、0.61%，但其精确度仍有待提高。MRC-PBM 模型通过联合 Biffine 预测器和 MLP 预测器并加入实体相对位置信息，帮助模型更加准确的识别嵌套实体并确定它们的边界。这使得精确度和 $F1$ 值明显提高，分别提高了 2.87%、1.21%；1.04%、0.79%，进一步证明了 MRC-PBM 模型在识别嵌套实体方面的优越性。

2.5 消融实验

为了验证所提出模型中关键组件的有效性，本文在 Diabetes 数据集上进行了消融实验，实验结果如表 8 所示。

表 8 消融实验结果 (%)

模型	精确度	召回率	$F1$ 值
本文的模型	90.51	90.51	90.51
移除 Hybrid-PBM	89.02	88.49	88.75
移除 实体相对位置信息	89.77	90.50	90.13
移除 Biffine 组件	91.25	89.33	90.28

首先，本文移除 Hybrid-PBM 模块，发现模型分别在精确度、召回率和 $F1$ 值方面下降了 1.49%、2.02%和 1.76%。实验结果表明，Hybrid-PBM 模块对模型的提升

是非常有帮助的。其次，本文移除 Hybrid-PBM 模块中相对位置信息组件，发现模型在精确度方面下降了 0.74%，这也导致了整体性能下降了 0.38%。实验表明实体相对位置信息组件能够提高模型预测的准确度。最后，本文移除 Hybrid-PBM 模块中 Biffine 组件，召回率大幅度下降（1.18%），性能下降 0.23%。实验表明 Biffine 组件能帮助查找出更多的实体，从而提高模型的性能。

3 结 论

为了解决嵌套命名实体识别时目标实体定位不完整或不准确，导致实体遗漏或误判的问题，本文提出了一种中文电子病历嵌套命名实体识别模型 MRC-PBM。该模型将 NER 任务转化为 MRC 问答任务，通过集成实体相关的先验信息到查询语句中，并使用预训练模型 MC_BERT 增强语境表示，Hybrid-PBM 预测器提高提取嵌套实体的能力，从而有效提取中文电子病历中的平面和嵌套实体。实验结果表明，该模型在 Diabetes 和 CMeEE 语料上优于主流的神经网络模型。下一步可以利用知识图谱技术将先验知识更好地集成到所提出的模型中，以提高模型在准确率、召回率以及 $F1$ 值上的性能表现。

参 考 文 献

[1] LEI J，TANG B，LU X，et al. A comprehensive study of named entity recognition in Chinese clinical text[J]. Journal of the American Medical Informatics Association，2014，21（5）：808-814.

[2] 张汝佳，代璐，王邦，等. 基于深度学习的中文命名实体识别最新研究进展综述[J]. 中文信息学报，2022，36（6）：20-35.

[3] 杨红梅，李琳，杨日东. 基于双向 LSTM 神经网络电子病历命名实体的识别模型[J]. 中国组织工程研究，2018，22（20）：3237-3242.

[4] 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，2022，46：100511.

[5] CHEN P，ZHANG M，YU X，et al. Named entity recognition of Chinese electronic medical records based on a hybrid neural network and medical MC-BERT[J]. BMC Medical Informatics and Decision Making，2022，22（1）：1-13.

[6] 李博，康晓东，张华丽，等. 采用 Transformer-CRF 的中文电子病历命名实体识别[J]. 计算机工程与应用，2020，56（5）：153-159.

[7] CHIU J P C，NICHOLS E. Named entity recognition with bidirectional LSTM-CNNs[J]. Transactions of the Association for Computational Linguistics，2016，4：357-370.

[8] 顾佼佼，翟一琛，姬嗣愚，等. 基于 BERT 和知识蒸馏的航空维修领域命名实体识别[J]. 电子测量技术，2023，46（3）：19-24.

[9] JU M，MIWA M，ANANIADOU S. A neural layered model for nested named entity recognition[C]. Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies，2018：1446-1459.

[10] JIANG X，HE K，HE J，et al. A new entity extraction method based on machine reading comprehension[J]. Computer Science，2021，DOI：10.48550/arXiv.2108.06444.

[11] YU J，BOHNET B，POESIO M. Named entity recognition as dependency parsing[C]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics，2020：6470-6476.

[12] LI J，FEI H，LIU J，et al. Unified named entity recognition as word-word relation classification[C]. Proceedings of the AAAI Conference on Artificial Intelligence，2022，36（10）：10965-10973.

[13] DEVLIN J，CHANG M W，LEE K，et al. Bert：Pre-training of deep bidirectional transformers for language understanding[C]. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies，2019：4171-4186.

[14] ZHANG N，JIA Q，YIN K，et al. Conceptualized representation learning for chinese biomedical text mining[J]. Computer Science，2020，DOI：10.48550/arXiv.2008.10813.

[15] 魏玮，吕游，齐欣宇，等. 基于 CNN-LSTM-AM 动态集成模型的电站风机状态预测方法[J]. 仪器仪表学报，2023，44(4)：19-27.

[16] 王瑞峰，李扬. 基于 1DCNN-BiLSTM 组合模型的 S700K 转辙机故障诊断[J]. 电子测量与仪器学报，2022，36(11)：193-200.

[17] KONG J，ZHANG L，JIANG M，et al. Incorporating multi-level CNN and attention mechanism for Chinese clinical named entity recognition[J]. Journal of Biomedical Informatics，2021，116：103737.

[18] HENDRYCKS D，GIMPEL K. Gaussian error linear units（GELUs）[J]. Computer Science，2016，DOI：10.48550/arXiv.1606.08415.

[19] SU J，MURTADHA A，PAN S，et al. Global pointer：Novel efficient span-based approach for named entity recognition[J]. Computer Science，2022，DOI：10.48550/arXiv.2208.03054.

[20] ZHANG N, CHEN M, BI Z, et al. Cblue: A Chinese biomedical language understanding evaluation benchmark[C]. Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics,2022: 7888-7915.

[21] SHIBUYA T, HOVY E. Nested named entity recognition via second-best sequence learning and decoding[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 605-620.

[22] YU J, BOHNET B, POESIO M. Named entity recognition as dependency parsing[C]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics,2020: 6470-6476.

[23] LI X, FENG J, MENG Y, et al. A unified MRC framework for named entity recognition [C]. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020: 5849-5859.

作 者 简 介

周佳伦，硕士研究生，主要研究方向为大数据分析技术。

E-mail: 202108540021130@ctgu.edu.cn

李琳宇，硕士研究生，主要研究方向为大数据分析技术。

E-mail: 202108120021010@ctgu.edu.cn

马洪彬，硕士研究生，主要研究方向为大数据分析技术。

E-mail: 202108120021015@ctgu.edu.cn

姜艳静(通信作者)，副教授，主要研究方向为健康医疗大数据分析。

E-mail: jiangyanjing@ctgu.edu.cn