2025 年 第 25 卷 第 11 期
2025, 25(11):04656-10
科 学 技 术 与 工 程
Science Technology and Engineering
ISSN 1671—1815
CN 11—4688/T

DOI:10.12404/j.issn.1671-1815.2403519
引用格式:吴天宇, 郭冬冬, 李文桥, 等. 基于 MacBERT 与全局指针网络的中文电子病历命名实体识别[J]. 科学技术与工程, 2025, 25(11): 4656-4665.
Wu Tianyu, Guo Dongdong, Li Wenqiao, et al. Named entity recognition for chinese electronic medical records using MacBERT and global pointer network[J]. Science Technology and Engineering, 2025, 25(11): 4656-4665.

基于 MacBERT 与全局指针网络的中文电子病历命名实体识别

吴天宇, 郭冬冬*, 李文桥, 李子康, 苗琳
(北京信息科技大学计算机学院, 北京 100101)

摘 要 针对现有序列标注方法不能有效解决中文电子病历嵌套实体识别问题,提出一种基于 MacBERT 与全局指针网络的中文电子病历命名实体识别模型。首先通过 MacBERT-large 预训练模型将文本转换为结合语境信息的动态向量,然后使用 FGM (fast gradient method)方法生成对抗样本添加至原有向量并一同输入 BiLSTM (bi-directional long short-term memory)网络获取上下文特征,并通过引入注意力机制增强长距离语义特征获取,最后利用全局指针网络模型同时考虑头部和尾部的特征信息进行解码以获得更好的医学嵌套实体预测效果。实验结果表明,本文模型相较于识别效果较好的主流模型全局指针网络模型在 CCKS2019 以及两个版本的 CMeEE 中文电子病历数据集上 F_1 分别提高了 1.8%、1.37%、1.72%,证明了模型的有效性。

关键词 命名实体识别; 中文电子病历; 全局指针网络; 注意力机制
中图法分类号 TP391; **文献标志码** A

Named Entity Recognition for Chinese Electronic Medical Records Using MacBERT and Global Pointer Network

WU Tian-yu, GUO Dong-dong*, LI Wen-qiao, LI Zi-kang, MIAO Lin
(Computer School, Beijing Information Science and Technology University, Beijing 100101, China)

[**Abstract**] Addressing the limitation of existing sequence labeling approaches in effectively recognizing nested entities within Chinese electronic health records (EHRs), a novel named entity recognition model that integrates MacBERT and a global pointer network was proposed. Initially, the MacBERT-large pre-trained model transformed the text into context-sensitive dynamic vectors. Subsequently, the fast gradient method (FGM) was employed to generate adversarial samples, which were incorporated into the original vectors and fed into a BiLSTM (bi-directional long short-term memory) network to capture contextual features. To enhance the capture of long-distance semantic features, an attention mechanism was introduced. Finally, a global pointer network model was leveraged to decode simultaneously considering both head and tail feature information, thereby achieving superior prediction performance for medical nested entities. Experimental results demonstrate that compared to the state-of-the-art global pointer model, the proposed model achieves an improvement of 1.8%, 1.37%, and 1.72% in F_1-score on the CCKS2019 dataset and two versions of the CMeEE Chinese EHR dataset, respectively, validating the effectiveness of the proposed approach.

[**Keywords**] named entity recognition; Chinese electronic medical record; global pointer network; attention mechanism

随着医疗技术的不断进步和医疗服务的日益完善,大量的医疗记录不断积累,给医疗管理和研究带来了极大的挑战。为了更加高效、便捷地管理这些记录,电子病历应运而生并逐渐得到广泛应用。电子病历的出现,不仅极大地提高了医疗工作的效率,也为医疗研究提供了更加便捷的数据来源。电子病历中通常蕴含着丰富的医学知识,当前,如何从非结构化公开的中文电子病历文本中抽取出结构化的医学知识,已成为医学信息抽取领域的研究热点之一[1-2]。

命名实体识别(named entity recognition, NER)可以从文本中识别出具有特定意义的实体,并将它们分类为预定义的类别,如人名、地名、组织机构名、日期、时间等[3]。中文电子病历命名实体识别

收稿日期: 2024-05-13 **修订日期**: 2024-08-01
基金项目: 国家重点研发计划(2021YFB2600600);北京信息科技大学校级科研项目(2023XJJ15,2023XJJ17)
第一作者: 吴天宇(2000—),男,汉族,北京丰台人,硕士研究生。研究方向:知识图谱。E-mail:767101839@qq.com。
* **通信作者**: 郭冬冬(1990—),男,汉族,山西大同人,博士。研究方向:自然语言处理。E-mail:dongdongguo@bistu.edu.cn。

旨在利用规则模板、机器学习以及深度学习等方法从非结构化的电子病历中抽取出不涉及患者隐私的、结构化的医学知识信息,如药物名、手术名、症状名等,并利用这些数据为医学信息抽取、医学知识图谱构建等工作提供坚实的支撑[4],从而进一步推动医疗信息化。

然而中文电子病历命名实体识别不同于通用领域的实体识别,其所包含的医学术语专业性强、结构复杂,同时含有大量嵌套医学实体[5],例如,针对中文电子病历文本"对有心脏损害者应进行床边动态心电监护",其中存在疾病名称实体"心脏损害"、治疗方案实体"床边动态心电监护",而"心脏损害"实体中又嵌套身体部位实体"心脏",现有序列标注方法通常不能同时识别两个实体。同时电子病历通常涉及患者隐私,需要专业人士进行数据标注,现有公开电子病历数据集少且存在较多的标注噪声。多种原因导致了电子病历信息抽取难度增大,中文电子病历的命名实体识别面临着巨大的挑战。

中文电子病历作为医疗信息化的重要组成部分,其命名实体识别的准确性直接影响到医疗数据的挖掘和利用。因此,为了提升中文电子病历命名实体识别模型性能以及解决中文电子病历中嵌套实体处理难的问题,对中文电子病历领域命名实体识别方法进行研究,使用深度学习方法提出一种基于 MacBERT 与全局指针网络的中文电子病历命名实体识别模型,在解决扁平医学实体识别任务的同时,能够更有效地处理嵌套医学实体的识别问题。此外通过引入注意力机制和对抗训练方法提升模型的鲁棒性与特征提取能力,进一步提高医疗命名实体的识别性能。本文模型为解决中文电子病历命名实体识别任务提供新的思路,同时有助于推动医疗信息化的发展。

1　相关工作

早期针对于命名实体识别研究主要采用两种方法,一是基于事先定义的规则模板,二是基于机器学习技术。基于规则模板的方法通常依赖医学领域专家根据数据分布特征制定固定的规则,然后进行实体匹配,制定规则的过程耗时,容易产生错误,同时可移植性差[6]。基于机器学习的方法通常利用机器自动提取文本特征,将命名实体识别视为序列标注任务,对输入序列的每个元素进行标记或打标签,然而这种方法需要大量的特征工程且训练出的模型泛化能力较弱。

随着各种算法和模型不断更迭,目前深度学习神经网络方法在医学信息处理中占据主导地位,尤其在实体识别任务上表现出显著优势,其主要研究方法包括卷积神经网络(convolutional neural networks,CNN)、循环神经网络(recurrent neural networks,RNN)等,以及改进 RNN 模型的长短期记忆网络(long-short term memory,LSTM)等模型[7]。Yin 等[8]提出基于字符偏旁的 BiLSTM-CRF(bi-directional long short-term memory-conditional random field)命名实体识别模型,并在基础上引入注意力机制,最终在中文电子病历数据集 CCKS2017 上取得了 93.00% 的 F_1 值。

基于深度学习的方法通常使用词向量嵌入模型进行编码,近期学者们发现将 BERT[9] 预训练模型及其各种变形如 RoBERTa、RoBERTa-wwm-ext、ALBERT 以及 MacBERT 等运用到电子病历实体识别模型的嵌入中,有助于增强语义表示。陈琛等[10]利 BERT 预训练模型代替 word2vec 生成词向量,并与 BiLSTM-CRF 模型结合,实验表明该模型对中文电子病历命名实体识别效果好于 BiLSTM-CRF。孔令巍等[11]在基线模型中引入对抗训练样本,最终在电子病历数据集 CCKS2021 上的精准率、召回率以及 F_1 值相比于基线模型均有所提升。陈娜等[12]在 BERT-BiGRU-CRF 模型的基础上引入注意力机制,强化了长距离文本语义特征获取,在多为扁平实体的电子病历数据集 CCKS2019 上相较于 BiLSTM-CRF 等基准模型取得了较好的识别效果。李洋等[13]结合对抗训练方法和 BERT 嵌入技术提出了一种命名实体识别模型,解决了复合材料检测领域的数据规模较小且专业名词多、边界混淆等问题。蒋丽媛等[14]使用笔画组成编码器获取汉字字形特征并与 BERT-BiLSTM-CRF 模型结合,在 Resume 数据集上取得了较好的识别效果。赵珍珍等[15]提出一种融合词信息与图注意力的命名实体识别模型,该模型使用学习了医疗知识的 MedBERT 作为嵌入层,在引入词向量嵌入的同时通过图注意力机制增强模型学习医学文本上下文关系的能力,最终在两个医学数据集上均取得了较好的识别水平。陆鑫涛等[16]提出一种融入拼音与词性特征的电子病历命名实体识别方法,在使用 BERT 进行嵌入的基础上引入中文拼音特征,并提取词性特征对拼音特征的不确定性加以约束,最后通过缩放点积注意力模块将三种类型特征进行融合,在电子病历数据集 CCKS2018、CCKS2019 与通用领域数据集 Weibo 上,F_1 分别达到了 98.66、87.25、73.41。

综上所述,目前针对命名实体识别的现有研究取得了显著进展,研究者们通过设计更复杂的神经

网络模型、利用更多的特征工程、优化算法、调整模型结构等方式,不断提升实体识别的准确率和效率。但其中大多数针对中文电子病历领域实体识别的研究主要采用序列标注方法集中解决扁平医学实体的识别,而对于电子病历文本中嵌套医学实体识别的研究较少,且效果不理想。因此,针对上述问题,同时为了提升中文电子病历领域命名实体识别准确率,提出一种结合 MacBERT 与全局指针网络的命名实体识别方法 MABAGP(MacBERT-adversarial-BiLSTM-attention-global pointer),解中文电子病历文本中扁平医学实体识别的同时更好地解决嵌套医学实体识别问题。

2 基于 MacBERT 与全局指针网络的中文电子病历命名实体识别模型

提出一种基于 MacBERT-large[17]、对抗训练(adversarial training,AT)[18]、双向长短期记忆网络(bi-directional long short-term memory,BiLSTM)、注意力机制以及全局指针网络(global pointer)[19]的中文电子病历命名实体识别模型 MABAGP,模型框架如图 1 所示。

MABAGP 模型的具体识别流程如下:首先将待识别的电子病历文本输入模型的嵌入层,通过使用 n-gram 与全词掩码随机替换策略的中文预训练语言模型 MacBERT-large 转换为高维度的动态向量表示,然后通过 FGM(fast gradient method)对抗训练算法针对原始向量添加扰动以生成对抗样本以提升模型的鲁棒性和泛化能力,并将原始向量与对抗向量一同输入双向长短期记忆网络中进行序列编码,再利用多头注意力机制增强长距离语义特征获取,最后在解码层通过全局指针网络模型同时考虑实体起始和终止位置的特征信息输出解码实体信息矩阵,在预测扁平实体的同时更好的预测医学嵌套实体,完成对中文电子病历文本的命名实体识别。

2.1 MacBERT-large

BERT 使用 MLM(masked language model)随机掩码策略,将输入序列中的一部分字随机地 MASK 掉,并要求模型根据上下文来预测这些被 MASK 的字的标识符。这使得模型能够双向地理解文本,并生成更具上下文意义的向量表示。但在中文信息处理中,词汇相比单个字可能包含了更多的语义信息,单纯对字进行 MASK 可能导致一些具有重要意义的信息缺失。RoBERTa-wwm-ext 在 BERT 的基础上进行改进,采用更大的训练集、参数、训练轮次,并使用全词掩码策略(whole word masking,WWM)代替 MLM,在保留了整个词语的完整性的同时缓解

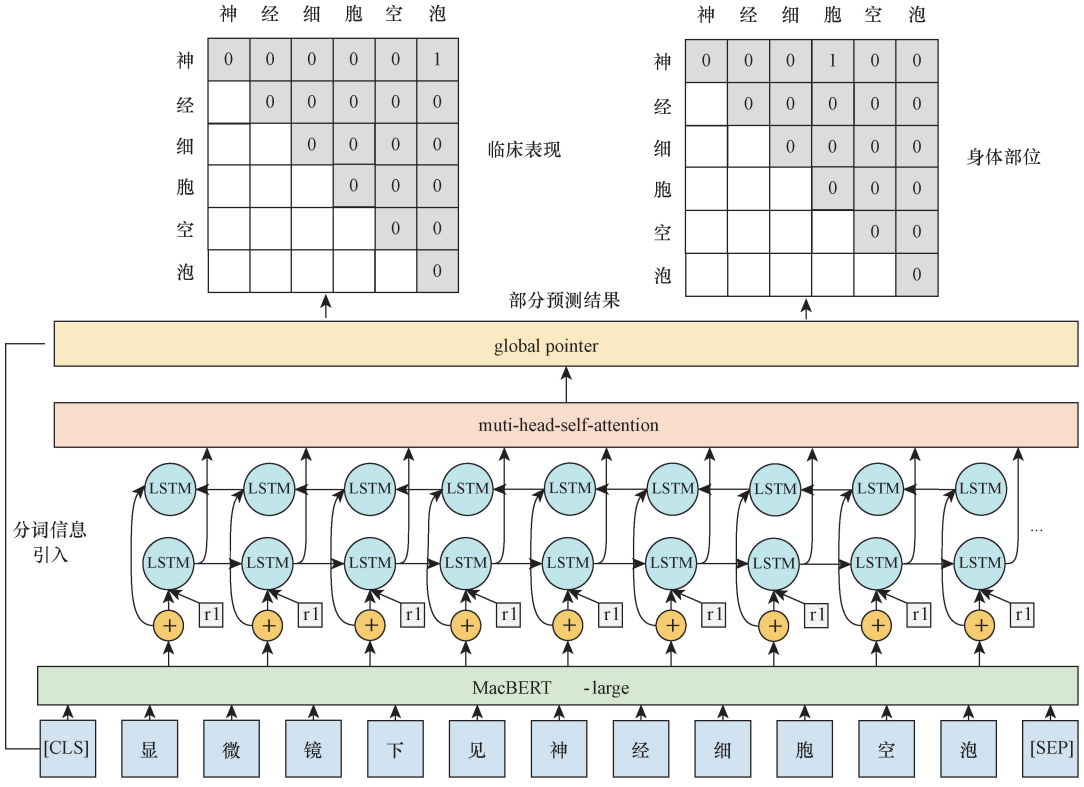

图 1 基于 MacBERT 与全局指针网络的中文电子病历命名实体识别模型结构图

Fig. 1 Structure diagram of Chinese electronic medical record named entity recognition model based on MacBERT and global pointer network

了信息丢失的问题,但全词掩码方法虽然对预训练阶段是有帮助的,能加强预训练的阶段模型效果,但是对下游任务提升不明显。MacBERT 是由哈工大讯飞联合实验室提出的新型预训练语言模型,使用纠错型掩码策略(masked language model as correction,MAC),其在全词掩码策略的基础上,利用 N-gram 方式选择待掩码的标记,其中 1-gram 至 4-gram 的概率分别为 40%、30%、20%、10%,再利用相似词进行代替 MASK 掩码标记,有效解决了预训练阶段有 MASK 标记,下游任务无 MASK 标记,这种上下游任务不一致的问题,从而提高了模型的泛化能力。MLM、WWM、MAC 三种不同掩码策略分别如图 2 所示。MacBERT 在中文信息处理任务上表现优秀,超越了 BERT、RoBERTa、ERNIE 等其他预训练模型。同时大量现有研究表明,大模型的识别效果通常要比小模型好,因此本文中使用更大规模语料库训练、模型结构更复杂的 MacBERT-large 作为模型的编码嵌入层。

MacBERT-large 首先对输入的中文电子病历文本进行分词操作,之后输入嵌入层。如图 3 所示,嵌入层包括三种嵌入表示,其中词嵌入用于表示词本身的信息特征,段嵌入用于判断句子的先后顺序以获取句子特征,位置嵌入则是用于编码和学习词在句中的位置信息特征,[CLS]和[SEP]分别代表句子开始与结束的标志,最后将 3 种嵌入方法获得的向量相加,得到输入中文电子病历文本的特征向量,作为后续模块的输入。

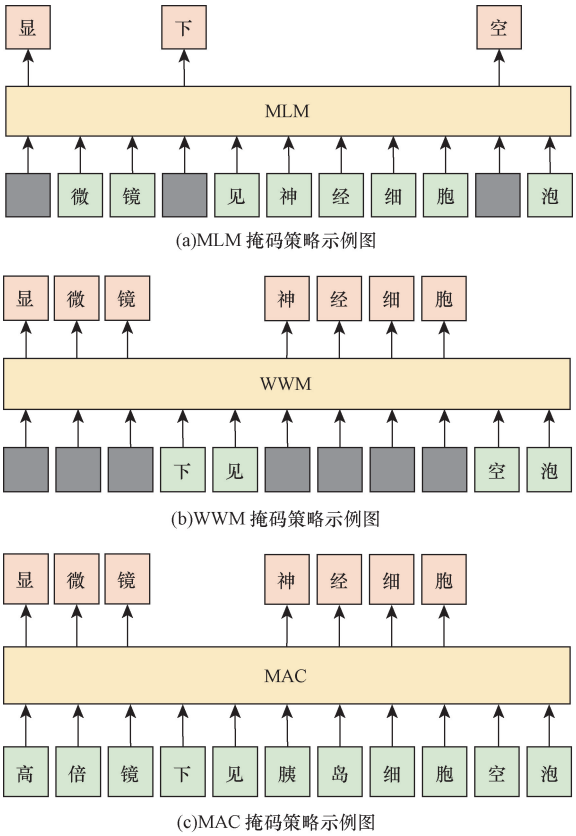

图 2　不同掩码策略对比图

Fig. 2　Comparison chart of different mask strategies

2.2　对抗训练

对抗训练是一种有效的深度学习正则化技术,其核心原理是使模型同时接受正样本和对抗样本的训练,其中对抗样本是通过对原始样本进行微小但有针对性的扰动而生成的。中文电子病历文本数据中通常存在各种形式的噪声和干扰,例如语义模糊、存在歧义和标注错误等。通过添加对抗训练,迫使模型在学习中更好地理解数据的分布,提高对于噪声和干扰的抵抗能力,以应对输入数据的微小变化。对抗训练可以简化为

$$\max p \geqslant (\boldsymbol{y} \mid \boldsymbol{x} + \Delta\boldsymbol{x}, \boldsymbol{\theta}) \tag{1}$$

式(1)中:$\boldsymbol{x}$ 代表输入;$\Delta\boldsymbol{x}$ 代表对抗扰动;$\boldsymbol{\theta}$ 为模型参数;$\boldsymbol{y}$ 为真实标签;为增加扰动后预测真实标签的概率,其中 $\Delta\boldsymbol{x}$ 在一定扰动空间内进行扰动。

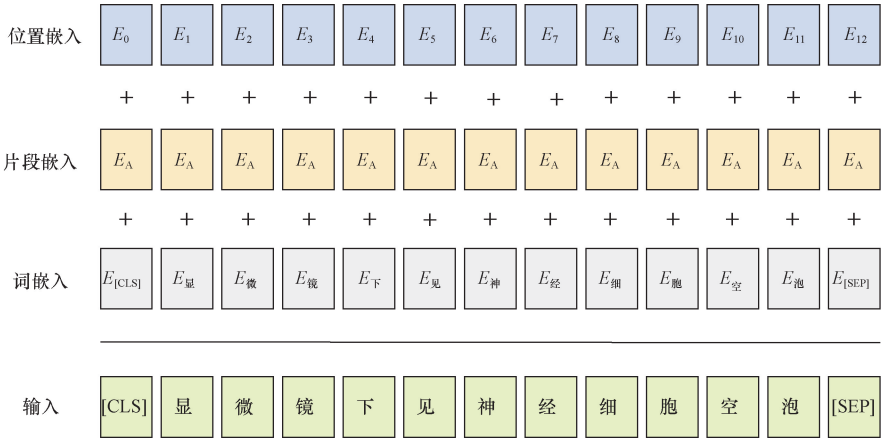

图 3　MacBERT 输入示例图

Fig. 3　MacBERT input example diagram

GoodFellow 等[20]提出 FGM 方法,以输入向量 $\boldsymbol{x}=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_t]$ 为例,首先复制预训练阶段的词汇向量,计算 $\boldsymbol{x}$ 的梯度并进行标准化处理,得到扰动值 $\Delta\boldsymbol{x}$,具体计算公式为

$$\Delta\boldsymbol{x}=\varepsilon\frac{\boldsymbol{g}}{\|\boldsymbol{g}\|_2} \tag{2}$$

式(2)中:ε 为一个缩放因子;$\boldsymbol{g}$ 为梯度,计算公式为

$$\boldsymbol{g}=\nabla_x L(\boldsymbol{v}_t,\boldsymbol{y},\boldsymbol{\theta}) \tag{3}$$

式(3)中:$L(\boldsymbol{v}_t,\boldsymbol{y},\boldsymbol{\theta})$ 是单个样本的损失。然后将扰动值与原始向量 $\boldsymbol{x}$ 相加,生成新的向量,并使用新的向量计算梯度,并将其与原始梯度相加,最后根据当前梯度来更新参数。以下是 FGM 对抗训练算法的简要描述。

对于数据集中的 $\boldsymbol{x}$:

(1)计算 $\boldsymbol{x}$ 的前向传播损失,并反向传播以获取梯度。

(2)通过嵌入矩阵的梯度算出 $\Delta\boldsymbol{x}$,并与 $\boldsymbol{x}$ 相加得到 $\boldsymbol{x}+\Delta\boldsymbol{x}$。

(3)计算 $\boldsymbol{x}+\Delta\boldsymbol{x}$ 的前向损失并反向传播以获取梯度,然后累加到步骤(1)中的梯度上。

(4)将 embedding 重置为步骤(1)时的状态。

(5)根据步骤(3)中计算得到的梯度来更新模型的参数。

2.3 双向长短期记忆网络(BiLSTM)

使用双向长短期记忆网络作为字向量的特征提取模块来捕获医学文本中的上下文语义信息。LSTM 是对循环神经网络进行改进后的模型,相较于传统的 RNN,LSTM 通过引入门控机制实现了对输入数据的有选择性存储与遗忘。门控机制赋予了 LSTM 网络更强大的记忆和学习能力,使其在处理长序列数据时表现更为优秀,同时很好地解决了循环神经网络的梯度爆炸问题。其中 LSTM 的单元结构如图 4 所示。

LSTM 按时间步从左到右处理输入序列,每个时间步的隐藏状态 $\boldsymbol{h}_t$ 和单元状态 $\boldsymbol{c}_t$ 的表达式为

$$\boldsymbol{i}_t=\sigma(\boldsymbol{W}_{ix}\boldsymbol{x}_t+\boldsymbol{W}_{ih}\boldsymbol{h}_{t-1}+\boldsymbol{b}_i) \tag{4}$$

$$\boldsymbol{f}_t=\sigma(\boldsymbol{W}_{fx}\boldsymbol{x}_t+\boldsymbol{W}_{fh}\boldsymbol{h}_{t-1}+\boldsymbol{b}_f) \tag{5}$$

$$\boldsymbol{o}_t=\sigma(\boldsymbol{W}_{ox}\boldsymbol{x}_t+\boldsymbol{W}_{oh}\boldsymbol{h}_{t-1}+\boldsymbol{b}_o) \tag{6}$$

$$\tilde{\boldsymbol{c}}_t=\tanh(\boldsymbol{W}_{cx}\boldsymbol{x}_t+\boldsymbol{W}_{ch}\boldsymbol{h}_{t-1}+\boldsymbol{b}_c) \tag{7}$$

$$\boldsymbol{c}_t=\boldsymbol{f}_t\odot\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\odot\tilde{\boldsymbol{c}}_t \tag{8}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\odot\tanh(\boldsymbol{c}_t) \tag{9}$$

式中:$\boldsymbol{i}_t$、$\boldsymbol{f}_t$、$\boldsymbol{o}_t$ 和 $\tilde{\boldsymbol{c}}_t$ 分别为输入门、遗忘门、输出门和表示当前信息的候选状态,$\boldsymbol{W}$ 为相应的权重;σ 和 tanh 分别为 Sigmoid 和双曲正切函数;$\odot$ 表示逐元素乘积。

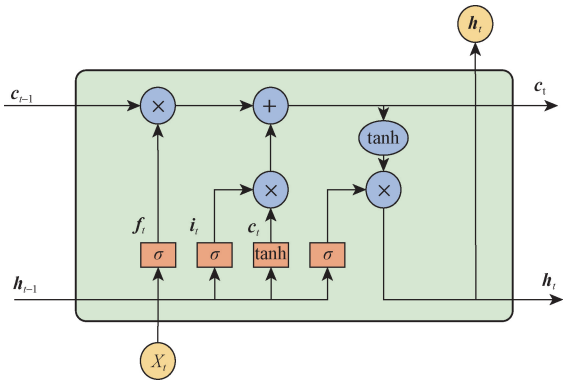

图 4 LSTM 单元结构图

Fig. 4 LSTM unit structure diagram

BiLSTM 则由两个方向相反的 LSTM 层组成,一个从序列的起始处开始处理(正向),另一个从序列的末尾处开始处理(逆向),通过同时考虑输入序列的过去和未来信息来捕捉序列中的长期依赖关系。在医学文本中,一个实例的含义通常与其前后的文本都密切相关。BiLSTM 能够捕获这种双向的依赖关系,从而更准确地识别出医学文本中的医疗实例。BiLSTM 在 t 时刻的输出 $\boldsymbol{h}_t$ 变化表达式为

$$\overrightarrow{\boldsymbol{h}_t}=\mathrm{LSTM}(\boldsymbol{x}_t,\overrightarrow{\boldsymbol{h}_t}) \tag{10}$$

$$\overleftarrow{\boldsymbol{h}_t}=\mathrm{LSTM}(\boldsymbol{x}_t,\overleftarrow{\boldsymbol{h}_t}) \tag{11}$$

$$\boldsymbol{h}_t=\left[\overrightarrow{\boldsymbol{h}_t},\ \overleftarrow{\boldsymbol{h}_t}\right] \tag{12}$$

式中:$\boldsymbol{x}_t$ 代表输入;$\overrightarrow{\boldsymbol{h}_t}$ 代表正向的 LSTM 的隐藏状态;$\overleftarrow{\boldsymbol{h}_t}$ 代表逆向的 LSTM 的隐藏状态;$\boldsymbol{h}_t$ 表示 BiLSTM 最终的输出,由两个方向的隐藏状态拼接而成。

2.4 注意力机制

由于中文电子病历文本中长句较为普遍,BiLSTM 在提取文本特征时难以有效捕获长距离依赖关系,以及难以获得句中各个字符对于识别的重要性,导致对于电子病历语义理解缺失,进而影响医学命名实体识别的准确性。因此本文借助注意力机制对 BiLSTM 模块进行增强,通过注意力分配机制使模型能够更多的学习与医学实体紧密相关的特征,同时减少学习与医学实体无关的特征,并在此基础上加强长距离语义特征获取,以达到更好的医学实体识别效果。

设 BiLSTM 层的输出为 $\boldsymbol{H}=\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n\}$,其中 $\boldsymbol{h}_i$ 为 BiLSTM 的第 i 个隐藏状态,n 为序列长度。首先,计算注意力权重 $\boldsymbol{e}_i$、$\boldsymbol{e}_i$ 为当前信息与上下文信息相关度的注意力分数向量,表达式为

$$\boldsymbol{e}_i=\tanh(\boldsymbol{W}_i\boldsymbol{h}_i+\boldsymbol{b}_i) \tag{13}$$

式(13)中:$\boldsymbol{W}_i$ 为可学习的权重矩阵;$\boldsymbol{b}_i$ 为偏置向量,然后使用 softmax 函数对注意力分数向量 $\boldsymbol{e}_i$ 进行归一化处理,得到注意力权重向量 $\boldsymbol{\alpha}_i$,表达式为

$$\boldsymbol{\alpha}_i = \frac{\mathrm{e}^{e_i}}{\sum_{j=1}^{n}\mathrm{e}^{e_j}} \tag{14}$$

最后使用注意力权重 $\boldsymbol{\alpha}$ 对 BiLSTM 的输出进行加权求和,得到注意力机制模块的输出,表达式为

$$\mathrm{Attention}(\boldsymbol{H}) = \sum_{i=1}^{n}\boldsymbol{\alpha}_i\boldsymbol{h}_i \tag{15}$$

2.5　全局指针网络

中文电子病历文本数据在经过 MacBERT-large 编码、对抗训练添加扰动、双向长短期记忆网络以及注意力机制提取文本特征后,还需要进行解码以输出预测实体。条件随机场模型(conditional random field,CRF)是命名实体识别领域常用的解码器,但是针对中文电子病历中存在大量嵌套医疗实体问题未能有效的解决。因此本文提出使用全局指针网络作为模型的解码层,其同时考虑头部和尾部的特征信息进行解码,在正确预测医学非嵌套实体的基础上获得更好的医学嵌套实体预测效果。

全局指针网络的整体思想为使用类似 Attention 的打分机制,将多个医学实体类型的识别视为 Mutihead 机制,其中每一个 head 负责一种医学实体类型的识别。设长度为 n 的中文电子病历文本经过编码得到 $\boldsymbol{x}=[\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_t]$,通过 $\boldsymbol{q}_{i,\alpha}=\boldsymbol{W}_{q,\alpha}\boldsymbol{v}_i+\boldsymbol{b}_{q,\alpha}$ 和 $\boldsymbol{k}_{i,\alpha}=\boldsymbol{W}_{k,\alpha}\boldsymbol{v}_i+\boldsymbol{b}_{k,\alpha}$ 变化得到序列向量序列 $\boldsymbol{q}_\alpha=[\boldsymbol{q}_{1,\alpha},\boldsymbol{q}_{2,\alpha},\cdots,\boldsymbol{q}_{n,\alpha}]$ 和 $\boldsymbol{k}_\alpha=[\boldsymbol{k}_{1,\alpha},\boldsymbol{k}_{2,\alpha},\cdots,\boldsymbol{k}_{n,\alpha}]$,用来识别第 α 种类型实体,相应的可以定义打分函数,表达式为

$$s_\alpha(i,j)=\boldsymbol{q}_{i,\alpha}^{\mathrm{T}}\boldsymbol{k}_{j,\alpha} \tag{16}$$

式(16)中:$s_\alpha(i,j)$ 表示从 i 到 j 的连续电子病历片段是一个类型为 α 的实体的打分。在此基础上,为了充分利用相对位置信息,引入旋转位置编码(rotary positional encoding,RoPE),表达式为

$$s_\alpha(i,j)=(\boldsymbol{R}_i\boldsymbol{q}_{i,\alpha})^{\mathrm{T}}(\boldsymbol{R}_j\boldsymbol{k}_{j,\alpha})=\boldsymbol{q}_{i,\alpha}^{\mathrm{T}}\boldsymbol{R}_i^{\mathrm{T}}\boldsymbol{R}_j\boldsymbol{k}_{j,\alpha} \tag{17}$$

$$\boldsymbol{q}_{i,\alpha}^{\mathrm{T}}\boldsymbol{R}_i^{\mathrm{T}}\boldsymbol{R}_j\boldsymbol{k}_{j,\alpha}=\boldsymbol{q}_{i,\alpha}^{\mathrm{T}}\boldsymbol{R}_{j-i}\boldsymbol{k}_{j,\alpha} \tag{18}$$

式中:$\boldsymbol{R}_i^{\mathrm{T}}\boldsymbol{R}_j=\boldsymbol{R}_{j-i}$ 。同时为了解决参数量倍增问题,采用矩阵共享方法共用一个打分矩阵 $(\boldsymbol{W}_q\boldsymbol{h}_i)^{\mathrm{T}}(\boldsymbol{W}_k\boldsymbol{h}_j)$ 对于同一实体类型的实体进行识别,该打分矩阵在新增实体类型也可继续沿用,打分函数可以优化为

$$s_\alpha(i,j)=(\boldsymbol{W}_q\boldsymbol{h}_i)^{\mathrm{T}}(\boldsymbol{W}_k\boldsymbol{h}_j)+\boldsymbol{w}_\alpha^{\mathrm{T}}[\boldsymbol{h}_i;\boldsymbol{h}_j] \tag{19}$$

还可用 $[\boldsymbol{q}_i;\boldsymbol{k}_i]$ 来替代 h_i 以进一步减少参数量,表达式为

$$s_\alpha(i,j)=\boldsymbol{q}_i^{\mathrm{T}}\boldsymbol{k}_j+\boldsymbol{w}_\alpha^{\mathrm{T}}[\boldsymbol{q}_i;\boldsymbol{k}_i;\boldsymbol{q}_j;\boldsymbol{k}_j] \tag{20}$$

在拥有考虑位置信息以及优化后的打分函数 $s_\alpha(i,j)$ 后,采用单目标多分类交叉熵损失函数 explogsum 实现权重的自动平衡,表达式为

$$L=\lg\left[1+\sum_{(i,j)\in P\alpha}\mathrm{e}^{-s_\alpha(i,j)}\right]+\lg\left[1+\sum_{(i,j)\in Q_\alpha}\mathrm{e}^{s_\alpha(i,j)}\right] \tag{21}$$

式(21)中:P_α 为电子病历样本所有类型为 α 的实体首位集合;Q_α 为电子病历样本的所有非实体或者类型非 α 的实体的首位集合,同时只考虑 $i\leqslant j$ 的组合。最后在解码时,满足 $s_\alpha(i,j)>0$ 的电子病历文本片段会输出类型为 α 的医学实体,由此完成对中文电子病历中命名实体的识别。

3　实验与分析

3.1　实验数据

为了验证本文所提模型对于中文电子病历命名实体识别的有效性,选取 CCKS2019、CMeEE[21]、CMeEE-V2 这 3 个数据集并设计相关实验并进行分析。CCKS2019 数据集来源于 2019 年全国知识图谱与语义计算大会,由医渡云医学根据真实的患者病历分布人工编辑而成,包含 1 000 条训练样本与 379 条测试样本。CMeEE 数据集来源于中文医疗信息处理评测基准 CBLUE[22],包含 15 000 条训练集、5 000条验证集与 3 000 条测试集。CMeEE-V2 是对 CMeEE 的补充修正版本,主要更新是修复了原始数据中的部分标注错误,提升了语料质量,数据集的分布情况没有改变。其中 CMeEE 与 CMeEE-V2 数据集均包含了医疗嵌套实体,用于验证本文提出的方法的有效性。本文所用数据集的分布情况及所用评价指标如表 1 所示。

表 1　数据集分布情况介绍

Table 1　Introduction to dataset distribution

数据集名称	实体类型	训练集	验证集	测试集	评价指标
CCKS2019	解剖部位、手术、疾病和诊断、药物、实验室检验、影像检查	1 000	200	179	F_1
CMeEE	疾病名称、临床表现、医疗程序、医疗设备、药物名称、医学检验项目、身体部位、科室微生物类	15 000	5 000	3 000	F_1
CMeEE-V2	疾病名称、临床表现、医疗程序、医疗设备、药物名称、医学检验项目、身体部位、科室微生物类	15 000	5 000	3 000	F_1

3.2 评价指标

本文使用 F_1 值作为主要评价指标在三个数据集上来测试命名实体识别模型的性能与效果,其中精准率 P(precision)、召回率 R(recall)作为辅助评价指标,F_1 值的计算同时兼顾了准确率和召回率。评价指标的具体计算方法为

$$P = \frac{n_p}{n_p + n_t} \tag{22}$$

$$R = \frac{n_p}{n_c} \tag{23}$$

$$F_1 = \frac{2PR}{P + R} \times 100\% \tag{24}$$

式中:n_p 为正确识别出的电子病历实体数量;n_t 为识别错误的电子病历实体数量;n_c 为待识别语料中所有电子病历实体数量。

3.3 实验设置

本文中所采用的实验环境设置如表 2 所示,实验参数设置如表 3 所示。

表 2 实验环境设置表

Table 2 Experimental environment setting table

实验环境名称	配置
操作系统	Windows10
编程语言	Python3.8
深度学习框架	Pytorch1.9.0
GPU	RTX 3090

表 3 实验参数设置表

Table 3 Experimental parameter setting table

参数名称	参数值
epoch	10
学习率	2×10^{-5}
LSTM 隐层维度	512
dropout	0.1
优化器	Adam
batch_size	16
最大序列长度	256

3.4 对比实验分析

3.4.1 解码方法对比

为了对比基于全局指针网络的解码方法与 CRF 解码方法在模型中发挥的作用,本文中分别在无嵌套实体中文电子病历数据集 CCKS2019 与有嵌套实体电子病历数据集 CMeEE 上进行实验。

如表 4 所示,在本文提出的模型其他模块保持不变的基础上,基于全局指针网络的解码方法相比于 CRF 在无嵌套实体电子病历数据集 CCKS2019 上 F_1 提高了 0.33%,在有嵌套实体电子病历数据集 CMeEE 上 F_1 提高了 1.11%。全局指针网络模型从全局的角度出发,同时考虑医疗实体的起始和终止位置,是将首尾看成一个整体来预测医疗非嵌套实体与嵌套实体,其对于非嵌套实体的识别效果可以媲美甚至略优于 CRF,对于嵌套实体的识别效果相比 CRF 有较大提升,且其训练和预测过程都是并行的,因此本文提出使用全局指针网络模型作为模型的解码方法是有效的。

表 4 解码方法对比实验

Table 4 Comparison experiment of decoding methods

序号	模型	F_1/%	
		CCKS2019	CMeEE
1	CRF	80.14	65.09
2	Global Pointer	80.47	66.21

为了对比全局指针网络模型与 CRF 模型在真实电子病历数据样本上识别医疗嵌套实体的能力,本文分别从不包含嵌套实体的中文电子病历数据集与包含嵌套实体的中文电子病历数据集数据集中随机选取测试数据并进行预测,结果如表 5 所示。可以看到针对只包含扁平医疗实体的电子病历文本“患者 3 月余前于我院诊断为直肠癌”,CRF 解码方法与 Global Pointer 解码方法均可以准确识别出“直肠癌”为疾病和诊断实体,且与真实结果一致。针对同时包含扁平医疗实体和嵌套医疗实体的电子病历文本“显微镜下见神经细胞空泡形成”,CRF 解码方法可以识别出“显微镜”为医疗设备实体,“神经细胞空泡”为临床表现实体,缺少对于嵌套身体物质实体“神经细胞”的识别,而 GlobalPointer 解码方法可以准确识别出以上三个类型的实体,且与真实结果一致,由此证明了全局指针网络模型作为解码方式的有效性与识别嵌套医疗实体的能力。

3.4.2 预训练模型对比

为了探究不同预训练模型编码方法对于电子病历命名实体识别模型性能的影响,实验选取了命名实体识别领域当中常用的预训练模型在中文电子病历数据集 CMeEE-V2 上进行对比实验分析。

表 5 不同解码方法识别结果示例

Table 5 Examples of recognition results using different decoding methods

文本	CRF	Global Pointer	真实结果
患者 3 月余前于我院诊断为直肠癌	直肠癌(疾病和诊断)	直肠癌(疾病和诊断)	直肠癌(疾病和诊断)
显微镜下见神经细胞空泡形成	显微镜(医疗设备) 神经细胞空泡(临床表现)	显微镜(医疗设备) 神经细胞(身体物质) 神经细胞空泡(临床表现)	显微镜(医疗设备) 神经细胞(身体物质) 神经细胞空泡(临床表现)

表 6　预训练模型对比

Table 6　Comparison of pre trained models

序号	模型	F_1/%
1	BERT	73.44
2	RoBERTa	73.66
3	RoBERTa-wwm-ext	73.75
4	MacBERT	73.86
5	RoBERTa-wwm-ext-large	74.69
6	MacBERT-large(MABAGP)	74.84

如表 6 所示:

(1)RoBERTa-wwm-ext 模型采用全词掩码策略获得词级别的向量,在 CMeEE-V2 数据集上取得了 73.75% 的 F_1,优于仅获得字级别的向量模型 RoBERTa,而在随机全词掩码策略的基础上采用 n-gram 与同义词替代方法的 MacBERT 模型消除了预训练阶段与下游任务不一致的问题,取得的 F_1 分数均优于上述两种模型,其原理也更适用于专业性强的中文电子病历数据。

(2)RoBERTa 模型相比原始 BERT 模型采用了动态字向量机制、更大的预训练数据集、更长的训练时间以及更多的预训练轮次以取得了更好的识别效果。RoBERTa-wwm-ext-large 和 MacBERT-large 采用 24 层 Transformer 编码器而含有更大的参数量,相比于 base 版 F_1 分别提升了 0.94% 与 0.98%。实验结果表明,训练数据越多、模型结构越复杂、参数量越多,预训练模型处理电子病历文本的效果越好,因此本文使用 F_1 最高 MacBERT-large 作为模型的嵌入层,模型整体的识别效果达到最优。

3.4.3　模型整体对比

为了验证本文提出的中文电子病历命名实体识别模型的有效性,实验选取了命名实体识别领域常用的基准模型与近期主流模型分别在 3 个不同的中文电子病历数据集上进行对比,选取的模型分别是:BiLSTM-CRF、BERT-BiLSTM-CRF、Global Pointer[19]、Deep Biaffine[23]、W2NER[24]。

如表 7 所示,BiLSTM-CRF 使用双向 LSTM 网络对文本进行序列建模,提取上下文包含的信息并使用 CRF 进行解码,而 BERT-BiLSTM-CRF 在 BiLSTM-CRF 的基础上使用 BERT 预训练模型作为字向量的嵌入方法以代替随机初始化方法,在 3 个数据集上的 F_1 分别提升了 3.26%、4.52%、3.46%,可以作为本文对比的基准模型。Deep Biaffine 利用双仿射结构识别实体间的依存关系,间接可以对实体进行识别,相较于基准模型有略微提升。Global Pointer 利用全局归一化的思路来进行命名实体识别,使用类似 Attention 的打分机制作为最后的标注矩阵,同时考虑头部和尾部的特征信息,并在此基础上引入了旋转式位置编码,在包含嵌套实体的数据集上较基准模型提升明显。W2NER 将命名实体识别任务转换为预测词对之间的关系类别,能够统一处理扁平实体、重叠实体和非连续实体三种命名实体识别任务但其在 CCKS2019 数据集上的 F_1 较基准模型下降 0.57%,在 CMeEE 和 CMeEE-V2 分别提升 0.11% 与 0.57%,原因是在面对多为长文本的中文电子病历数据时,W2NER 可能预测出了除扁平实体以外其他类型的实体,导致了 F_1 分数不理想。本文提出的中文电子病历命名实体识别模型 MABAGP 相比基准模型在 3 个数据集上分别提高了 2.07%、2.41%、4.92%;相比于识别效果最好的主流模型 Global Pointer 在 3 个数据集上分别提高了 1.8%、1.37%、1.72%,由此验证了本文所提出模型在中文电子病历命名实体识别领域中的有效性与实用性。

表 7　与主流模型对比实验

Table 7　Comparison experiment with mainstream models

序号	模型	F_1/%		
		CCKS2019	CMeEE	CMeEE-V2
1	BiLSTM + CRF	75.14	59.32	66.46
2	BERT-BiLSTM-CRF	78.40	63.80	69.92
3	Deep Biaffine[23]	78.51	63.89	71.57
4	W2NER[24]	77.83	63.91	70.49
5	Global Pointer[19]	78.67	64.84	73.12
6	MABAGP(本文方法)	80.47	66.21	74.84

3.5　消融实验分析

本文所提出的方法在多个评估数据集上均优于其他对比模型,同时又对比了不同解码方法以及不同预训练模型对于模型性能的影响。在此基础之上,为了进一步探究加入对抗训练(AT)与注意力机制(Attention)对于模型识别效果的影响,本文在多为扁平医疗实体的中文电子病历数据集 CCKS2019 上设计了消融实验。将 MacBERT-BiLSTM-GP 模型作为基准模型(Baseline),并分别引入对抗训练样本和注意力机制进行实验。

由表 8 可知:

(1)通过比较模型 baseline 和 Baseline + AT 可以得出,只增加对抗训练样本对模型识别效果有提升作用,F_1 较基准模型提升了 0.4%。针对电子病历数

表 8　消融实验

Table 8　Ablation experiment

序号	模型	F_1/%
1	Baseline	79.88
2	Baseline + AT	80.28
3	Baseline + Attention	80.32
4	Baseline + AT + Attention	80.47

据中存在的标注噪声问题,本文通过 FGM 方法生成对抗训练样本向量,然后以扰动形式添加至原始电子病历文本向量,迫使模型学习到更加鲁棒和泛化的特征表示,提高了其在面对未知数据时的性能,从而达到了更好的识别效果。

(2)通过比较模型 baseline 和 Baseline + Attention 可以得出,只增加注意力机制对模型识别效果有提升作用,F_1 较基准模型提升了约 0.45%。针对中文电子病历文本语句通常较长,BiLSTM 提取上下文特征时无法获得长距离的特征问题,通过引入注意力机制,在捕获长距离特征的基础上,为医疗命名实体相关的特征分配较多的注意力,无关的特征分配较少的注意力,进一步强化了当前信息与上下文信息之间潜在的语义关联性,从而提高了医疗命名实体的识别准确率。

(3)通过比较模型 baseline 和 Baseline + AT + Attention(即 MABAGP)可以得出,同时增加对抗训练与注意力机制至模型中,F_1 较基准模型提升了约 0.6%,证明了各个模块在模型中的有效性,模型的识别效果达到最优。

4 结论

为了改善中文电子病历命名实体识别存在标注噪声以及嵌套实体处理难等问题,提出了一种命名实体识别模型 MABAGP。该模型首先通过 MacBERT-large 预训练模型将输入的电子病历文本转换为结合语境信息的动态向量,然后使用 FGM 对抗训练方法生成对抗样本以在原始向量上添加扰动,再将原始样本与对抗样本共同输入 BiLSTM 网络中捕获电子病历包含的上下文语义信息,并通过引入注意力机制增强长距离语义特征获取,最后利用全局指针网络模型同时考虑头部和尾部的特征信息进行解码。本文中分别在 3 个中文电子病历数据集上通过对比实验与消融实验证明了模型的有效性,在解决医学嵌套实体处理难问题的基础上改善了中文电子病历命名实体识别的效果,但是还存在一定改进空间。

(1)可以考虑使用医学领域数据预训练的 BERT 模型以取得更好的效果。

(2)通过引入 FGM 对抗训练方法提高了识别效果,可以考虑使用其他新型对抗训练方法进行对比取优。

(3)将本文提出的模型应该于其他命名实体识别领域以验证模型的泛化能力。

参 考 文 献

[1] 杜晋华, 尹浩, 冯嵩. 中文电子病历命名实体识别的研究与进展[J]. 电子学报, 2022, 50(12): 3030-3053.
Du Jinhua, Yin Hao, Feng Song. Research and progress on named entity recognition of Chinese electronic medical records[J]. Journal of Electronics, 2022, 50(12): 3030-3053.

[2] 周冬冬. 中文电子病历命名实体识别研究[D]. 大庆: 东北石油大学, 2024.
Zhou Dongdong. Research on named entity recognition of Chinese electronic medical records [D]. Daqing: Northeast University of Petroleum, 2024.

[3] 赵继贵, 钱育蓉, 王魁, 等. 中文命名实体识别研究综述[J]. 计算机工程与应用, 2024, 60(1): 15-27.
Zhao Jigui, Qian Yurong, Wang Kui, et al. A review of research on Chinese named entity recognition[J]. Computer Engineering and Applications, 2024, 60(1): 15-27.

[4] Gao Y, Gu L, Wang Y, et al. Constructing a Chinese electronic medical record corpus for named entity recognition on resident admit notes[J]. BMC Medical Informatics and Decision Making, 2019, 19(2): 67-78.

[5] 吉旭瑞, 魏德健, 张俊忠, 等. 中文电子病历信息提取方法研究综述[J]. 计算机工程与科学, 2024, 46(2): 325-337.
Ji Xurui, Wei Dejian, Zhang Junzhong, et al. A review of research on information extraction methods for Chinese electronic medical records[J]. Computer Engineering and Science, 2024, 46(2): 325-337.

[6] 刘浩, 张建业, 吕张成, 等. 面向数控机床设计知识图谱构建的实体识别[J]. 科学技术与工程, 2023, 23(13): 5655-5661.
Liu Hao, Zhang Jianye, Lü Zhangcheng, et al. Entity recognition based on knowledge graph construction for CNC machine tool design [J]. Science Technology and Engineering, 2023, 23 (13): 5655-5661.

[7] 赵辉, 庞海婷, 冯珊珊, 等. 中文命名实体识别技术综述[J]. 长春工业大学学报, 2021, 42(5): 444-450.
Zhao Hui, Pang Haiting, Feng Shanshan, et al. Overview of Chinese named entity recognition technology[J]. Journal of Changchun University of Technology, 2021, 42(5): 444-450.

[8] Yin M W, Mou C J, Xiong K N, et al. Chinese clinical named entity recognition with radical-level feature and self-attention mechanism[J]. Journal of Biomedical Informatics, 2019, 98: 103289.

[9] Devlin J, Chang M W, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[J]. Association for Computational Linguistics, 2019, 54(2): 104-124.

[10] 陈琛, 吴芬琳. 基于 BERT 的电子病历命名实体识别[J]. 自动化与仪器仪表, 2021, 41(3): 173-176.
Chen Chen, Wu Fenlin. Named entity recognition of electronic medical records based on BERT[J]. Automation and Instrumentation, 2021, 41(3): 173-176.

[11] 孔令巍, 朱艳辉, 张旭, 等. 基于对抗训练的中文电子病历命名实体识别[J]. 湖南工业大学学报, 2022, 36(3): 36-43.
Kong Lingwei, Zhu Yanhui, Zhang Xu, et al. Chinese electronic medical record named entity recognition based on adversarial training[J]. Journal of Hunan University of Technology, 2022, 36 (3): 36-43.

[12] 陈娜, 孙艳秋, 燕燕. 结合注意力机制的 BERT-BiGRU-CRF 中文电子病历命名实体识别[J]. 小型微型计算机系统, 2023, 44(8): 1680-1685.
Chen Na, Sun Yanqiu, Yan Yan. BERT BiGRU-CRF Chinese

electronic medical record named entity recognition combined with attention mechanism[J]. Small Micro Computer Systems, 2023, 44(8): 1680-1685.

[13] 李洋, 蔡红珍, 邢林林, 等. 基于对抗迁移的复合材料检测领域命名实体识别[J]. 科学技术与工程, 2022, 22(30): 13370-13377.
Li Yang, Cai Hongzhen, Xing Linlin, et al. Named entity recognition in the field of composite material detection based on adversarial transfer[J]. Science Technology and Engineering, 2022, 22(30): 13370-13377.

[14] 蒋丽媛, 吴亚东, 王书航, 等. 融合笔画特征的命名实体识别方法[J]. 科学技术与工程, 2023, 23(17): 7436-7443.
Jiang Liyuan, Wu Yadong, Wang Shuhang, et al. A named entity recognition method based on fusion of stroke features[J]. Science Technology and Engineering, 2023, 23(17): 7436-7443.

[15] 赵珍珍, 董彦如, 刘静, 等. 融合词信息和图注意力的医学命名实体识别[J]. 计算机工程与应用, 2024, 60(11): 147-155.
Zhao Zhenzhen, Dong Yanru, Liu Jing, et al. Medical named entity recognition based on fusion of word information and graph attention[J]. Computer Engineering and Applications, 2024, 60(11): 147-155.

[16] 陆鑫涛, 孙丽萍, 凌晨, 等. 融入拼音与词性特征的中文电子病历命名实体识别[J/OL]. 小型微型计算机系统: 1-12[2024-03-06]. http://kns.cnki.net/kcms/detail/21.1106.TP.20240228.1116.013.html.
Lu Xintao, Sun Liping, Ling Chen, et al. Chinese electronic medical record named entity recognition incorporating pinyin and part ofspeech features[J/OL]. Mini microcomputer systems: 1-12[2024-03-06] http://kns.cnki.net/kcms/detail/21.1106.TP.20240228.1116.013.html.

[17] Cui Y M, Che W X, Liu T, et al. Revisiting pre-trained models for Chinese natural language processing[J]. Findings of the Association for Computational Linguistics. Online: EMNLP, 2020: 657-668.

[18] Szegedy C, Zaremba W, SutskeveR I, et al. Intriguing properties of neural networks[J/OL]. Computer Science, 2013. https://arxiv.org/abs/1312.6199.

[19] Su J L, Murtadha A, Pan S F, et al. Global pointer: novel efficient span-based approach for named entity recognition[J]. arXiv: 2208.03054, 2022.

[20] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[EB/OL]. http://arxiv.org/abs/1412.6572.

[21]]Hong Y Z, Wen X L, Kun L Z, et al. Building a pediatric medical corpus: word segmentation and named entity annotation[C]//Workshop on Chinese Lexical Semantics. Cham: Springer, 2020: 652-664.

[22] Zhang N, Chen M, Bi Z, et al. CBLUE: a Chinese biomedical language understanding evaluation benchmark[J]. arXiv preprint arXiv: 2106.08087, 2021.

[23] 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.

[24] Li J, Fei H, Liu J, et al. Unified named entity recognition as word-word relation classification[J]. Artificial Intelligence, 2022, 36(10): 10965-10973.