姚传彪,马汉杰,董慧,等. 基于 MacBERT-GP 的中文医学命名实体识别方法[J]. 智能计算机与应用,2025,15(2):190-197.
DOI:10.20169/j.issn.2095-2163.250229

基于 MacBERT-GP 的中文医学命名实体识别方法

姚传彪[1],马汉杰[2],董 慧[3],许永恩[3],李少华[1],李东倪[4]
(1 浙江理工大学 信息科学与工程学院,杭州 310018;2 浙江理工大学 计算机科学与技术学院,杭州 310018;
3 杭州码全信息科技有限公司,杭州 311199;4 杭州电子科技大学 计算机学院,杭州 310018)

摘 要:医学命名实体识别作为医学信息提取的基础任务,在构建医学知识图谱、解决医学问题和自动分析病例等方面具有重要意义。相较于一般领域的命名实体识别,医学命名实体识别面临中文分词复杂性和医学领域术语丰富性等挑战,同时医学命名实体更为复杂、容易嵌套。为了提升现有命名实体识别模型的准确性,解决训练过程中高质量标注数据匮乏的问题,针对医学命名实体识别,提出一种基于 MacBERT-GP 的中文医学命名实体识别方法。在 CBLUE CMeEE-V2 医疗命名实体识别数据集和 CCKS2019 电子病历数据集上的实验结果,充分验证了该方法的有效性。相较于经典的 BERT-BiLSTM-CRF 方法,所提出的方法在 *F*1 值上分别提升了 6.24% 和 4.95%。

关键词:MacBERT;大语言模型;全局指针;特征融合;嵌套实体
中图分类号:TP311.1 **文献标志码**:A **文章编号**:2095-2163(2025)02-0190-08

Chinese Medical Named Entity Recognition based on MacBERT-GP

YAO Chuanbiao[1], MA Hanjie[2], DONG Hui[3], XU Yongen[3], LI Shaohua[1], LI Dongni[4]
(1 School of Information Science & Engineering, Zhejiang Sci-Tech University, Hangzhou 310018, China; 2 School of Computer Science & Technology, Zhejiang Sci-Tech University, Hangzhou 310018, China; 3 Hangzhou Codvision Technology Co., Ltd., Hangzhou 311199, China; 4 School of Computer, Hangzhou Dianzi University, Hangzhou 310018, China)

Abstract: Medical Named Entity Recognition (NER), as a fundamental task in medical information extraction, holds significant importance in building medical knowledge graphs, addressing medical issues, and automating case analysis. Compared to general domain NER, medical NER faces challenges such as the complexity of Chinese word segmentation and the richness of medical terminology. Additionally, medical named entities tend to be more complex and can be nested. In order to enhance the accuracy of existing NER models and address the issue of limited high-quality annotated data during training, a Chinese medical NER approach based on MacBERT-GP is proposed specifically for medical named entity recognition. Experimental results on the CBLUE CMeEE-V2 medical NER dataset and the CCKS2019 electronic medical record dataset have thoroughly validated the effectiveness of this method. Compared to the classic BERT-BiLSTM-CRF method, the proposed approach has shown improvements of 6.42% and 4.95% in *F*1 scores, respectively.

Key words: MacBERT; large language model; Global pointer; feature fusion; nested entity

0 引 言

命名实体是指在具有相似属性的其他项集合中,能够被清楚地识别出一个特定项的单词或短语[1]。命名实体识别(Named Entity Recognition, NER)是自然语言处理研究领域的基础性工作之一,其任务是提取非结构化文本中具有特定含义的命名实体。命名实体能否被准确识别,对于自然语言处

基金项目:浙江省重点研发项目(2021C01163);杭州市重大科技创新项目(2022AIZD0145);浙江省“尖兵”“领雁”研发攻关计划(2023C01041)。

作者简介:姚传彪(1999—),男,硕士研究生,主要研究方向:自然语言处理;董 慧(1994—),女,硕士,主要研究方向:计算机视觉;许永恩(1993—),男,硕士,主要研究方向:音视频编解码;李少华(1998—),男,硕士研究生,主要研究方向:计算机视觉;李东倪(2000—),女,硕士研究生,主要研究方向:能耗优化。
通信作者:马汉杰(1982—),男,博士,副教授,主要研究方向:视频图像传输与处理,机器视觉,情感计算等。Email:mahanjie@zstu.edu.cn。

收稿日期:2023-10-02

理上层任务(如:信息检索、自动问答、信息抽取、知识图谱构建)都有着重要的影响。对于分词、词性标注等底层序列标注任务,也存在着相互影响。在如今这个数据量飞速增长的数字时代,从海量数据中快速筛选有用信息,准确获取命名实体是关键性的一步。

目前,医疗信息化水平不断提升,随之产生了海量的医疗数据,然而这些数据很多是非结构化的文本数据,无法直接进行分析和利用。如何通过 NER 技术进行医疗文本结构化,已成为重要的研究课题[2]。

近年来,基于深度学习的 NER 模型占据主导地位,并取得了最先进的结果。与基于特征的方法相比,深度学习有利于自动发现隐藏特征[3]。深度学习的关键优势在于表征学习的能力,以及向量表示和神经处理所赋予的语义组合能力。这允许机器被输入原始数据,并自动发现分类或检测所需的潜在表示和处理[4]。典型的深度学习 NER 方法,包括双向长短期记忆网络后接条件随机场(Bidirectional Long Short-Term Memory with Conditional Random Field,BiLSTM-CRF)[5]和基于变换器的双向编码器(Bidirectional Encoder Representation from Transformers,BERT)[6],识别效果均明显优于传统识别方法。但是,基于深度学习的 NER 需要学习大量的已标注数据,而在医疗领域,精确的标注数据难以获取,主要是因为标注过程需要许多领域内的专家,耗时而且昂贵。如何在样本不足情境下提升命名实体识别的效果,已经成为 NER 领域的热门研究方向之一。

关于 NER 问题,研究人员提出了很多方法。Global Pointer[7]方法是一种结构精简高效的实体识别解码方法,是一种基于序列(*Span*)的识别方法,其将输入文本中所有子串分别打分,输出标签,能够适用于存在嵌套情况下的命名实体识别。Global Pointer 采用 Bert 的最后一层输出作为表征,进行实体的解码。然而预训练语言模型无法在单一的 Transformer 层学习到充分的语义信息,因此对 Global Pointer 可以进行更进一步的优化,将其实体判别打分方法从单一特征优化为多层特征。进而在命名实体识别中可以取得更好的性能。

此外,虽然深度学习已经在很多领域的 NER 任务中取得了较高的准确性,但是模型的性能高低,往往取决于训练数据的规模与质量。对于特定领域,例如医学领域,由于标签标注的难度较大,往往缺乏大规模高质量的数据集。因此,数据增强方法被引入到命名实体识别任务中,用来有效的扩充医学数据集,并且进一步提升模型的泛化性。数据增强方法最早应用于计算机视觉领域,通过扩充图像的多样性和鲁棒性来训练更优秀的模型。在自然语言处理领域中,传统数据增强技术主要通过语言翻译的方法来实现。由于语言具有差异性,语言翻译方法在中文文本上的效果不尽人意,但也有研究人员提出了基于中文的数据增强方法(Easy Data Augmentation,EDA),其中包括同义词替换等方法,在一些较小数据集上取得了一些效果。但是,由于领域的差异,通过这种方法部分增强的句子容易产生语义逻辑混乱,进而破坏了句子原本的语义信息,尤其在医学领域等特定领域的数据集,EDA 难以识别专业词汇,无法很好的扩展出与原句匹配的质量优秀的数据集。基于这种情况,本文提出了基于大语言模型的数据增强方法。利用大模型丰富的语义知识,替换数据集中的信息,扩展出新的数据集,从而提高模型的识别性能。

本文提出一种基于 MacBERT-GP 的中文医学命名实体识别方法,以减少现有的深度 NER 模型对大规模标注数据集的依赖。MacBERT[8]是一种通过用相似的单词 mask,减轻了预训练和微调阶段两者之间差距的 BERT 变体。该方法在利用多特征融合改进 Global Pointer 网络的基础上,利用大模型数据增强方法扩充数据集,最后在 CBLUE CMeEE-V2 医疗命名实体识别数据集和 CCKS2019 数据集上,评估了该方法的实用性和可行性。

1　相关工作

“命名实体”(Named Entity, NE)一词最早出现在第六届消息理解会议(Message Understanding Conference, MUC-6)上,用于定义对信息单元的识别。关于 NER 技术,大致可以将其分为 4 类:基于规则的方法、无监督学习方法、基于特征的监督学习方法和基于深度学习的方法。

随着人工智能技术和现代医疗的快速发展,两者之间的技术融合愈加深入,命名实体识别在医学中发挥越来越重要的作用,主要的研究方法包括传统方法和深度学习方法。其中,传统方法包括基于词典和规则的方法,以及基于机器学习的方法。

早期的命名实体识别主要采用基于规则和词典的方法。基于规则的命名实体识别依赖于手动定义的规则,通常是基于人工经验制定的,可以基于词

性、词性标签、上下文等特征来识别实体。规则方法简单易懂,可以根据任务和语料库的特点进行个性化定制。然而,其过于依赖人工设计的规则,可能在处理新的未知数据时表现不佳,需要大量的人工投入。基于词典的命名实体识别方法使用预先构建的词典或知识库来标识实体。这些词典包含已知的实体词汇和名称。在识别过程中,如果词汇出现在词典中,就将其标记为相应的实体类型。词典方法简单高效,对于词典中包含的实体类型表现出色,但无法处理未在词典中出现的新实体,也不能识别词典未包含的实体类型。为了克服规则方法和词典方法各自的限制,有时会将其结合起来使用。例如,可以使用规则方法来识别常见的实体类型,同时使用词典方法处理特定的实体,或者对规则和词典方法的输出进行后处理和整合,这种综合方法可以更好地应对多样性的实体识别需求。

由于词典和规则方法的局限性,逐渐提出了基于机器学习的方法,主要包括隐马尔可夫链[9](Hidden Markov Model, HMM)、支持向量机[10](Support Vector Machine, SVM)和条件随机场模型[11](Conditional Random Fields, CRF)等,基于机器学习的方法,在命名实体识别中具有一定的灵活性和泛化能力,可以适应不同类型的实体和不同领域的数据。然而,其对标注数据的依赖较大,并且需要设计合适的特征来表示文本。

随着人工智能的快速发展,基于深度学习的神经网络逐渐成为解决命名实体问题的主流方法。例如:卷积神经网络[12](Convolutional Neural Networks, CNN)、循环神经网络[13](Recurrent Neural Network, RNN),其可以自动学习更丰富的特征表示,避免了手动设计特征的繁琐过程,并在许多任务中取得了更好的性能。基于 RNN-CRF[14]的方法在中文命名实体识别任务中取得了很好的效果。Dong 等[15]首次将基于字符嵌入的双向长短期记忆网络和条件随机场模型相结合,并将其作为中文 NER 任务,并且将字符的偏旁作为序列输入 BiLSTM 进行编码,在不依靠手工设计语言特征的情况下,取得了较好的结果。在医学命名实体识别领域,许多研究都采用 BiLSTM-CRF 模型作为核心框架,对医学数据进行实体识别。

2 模型框架

2.1 任务定义

传统的 NER 标注方法,将 NER 视为序列标注任务,但是这种模式具有一定的局限性,对于单个 Token 只能输出一种标记,无法识别嵌套实体。本文将 NER 视为跨度(*span*)分类任务,给定文本序列 $x=[x_1,\cdots,x_i,\cdots,x_n]$,其中 x_i 表示序列中的第 i 个字符,设 $S=[S_1,S_2,\cdots,S_n]$ 是句子中可能出现的所有 *span*。$span(i,j)$ 为序列 x 的第 i 个到第 j 个元素组成的连续子串,其中 i 和 j 分别是头索引和尾索引。NER 的目标是识别所有 $s \in E$,其中 E 是实体类型集。

2.2 模型框架

本文采用预训练模型 MacBERT 作为基础框架,并接入 Global Pointer 模块。首先通过在 MacBERT 模型中进行运算,得到字级别的 Token 向量;然后利用实体矩阵计算 *span* 得分,得到最终的实体分数。本文所用模型架构如图 1 所示。

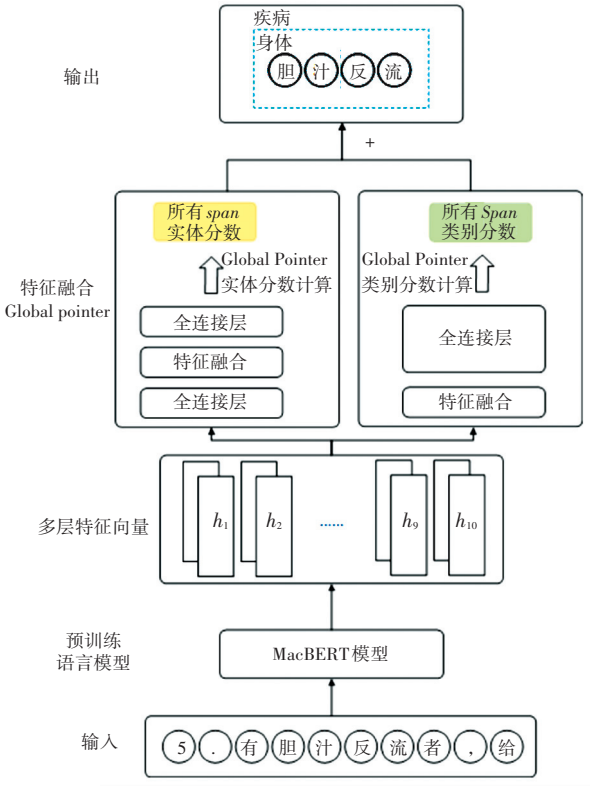

图 1 本文模型框架图

Fig. 1 Model framework diagram

在本文中,采用 MacBERT 作为预训练的语言模型。输入的文本序列经过预训练模型中的 Embedding 层转换为向量表示,并将这些向量输入到模型的自注意力模块中进行学习。通过这一过程,得到了最终的向量表示序列 $\boldsymbol{H}$,如下式所示:

$$E_{\text{input}}=\{E_0,\cdots E_i,\cdots E_n\}=\text{embeddings}(x_1,\cdots x_i,\cdots x_n) \tag{1}$$

$$\boldsymbol{H}=\{h_1,h_2,\cdots,h_n\}=\text{MacBERT}(E_{\text{input}}) \tag{2}$$

其中，h_i 是字符 x_i 对应的向量表示。

Global Pointer 是一种基于全局归一化方法的命名实体识别(NER)技术，其具备出色的性能并能无差别地识别嵌套和非嵌套实体。在非嵌套命名实体识别(Flat NER)任务中，其性能与 CRF 相媲美，而在嵌套命名实体识别(Nested NER)任务中，CRF 通常表现不佳，因为无法胜任嵌套情况。

综上所述，Global Pointer 不仅能够提取嵌套实体，而且性能卓越。传统的 Pointer Network 在实体识别任务中一般使用两个模块分别识别实体的开头和结尾，而这种方法无法保证训练和预测的一致性。然而，Global Pointer 经过优化，将实体的开头和结尾视为一个整体进行判别，因此具备更全面的视角。从命名实体识别任务出发，许多模型将其视为序列标注任务，因此对于同一字无法输出多个标签，也无法解决嵌套问题。Global Pointer 将文本序列中的所有连续子序列都视为候选实体片段，对于长度为 n 的文本序列，就有 $n(n+1)/2$ 个候选实体片段。这些子序列包含了所有可能需要识别的实体，本文的任务是从中找出真正的实体，因此命名实体识别问题被转化为“$n(n+1)/2$ 选 k”的多标签分类问题。如果需要识别 m 种不同类型的命名实体，则任务就演化为 m 个“$n(n+1)/2$ 选 k”的多标签分类问题。

设输入长度为 n 的文本 X，经过预训练模型编码后，可以得到向量序列$\boldsymbol{H}=\{h_1,h_2,\cdots,h_n\}$，通过下式变换：

$$q_{i,\alpha}=W_{q,\alpha}h_i+b_{q,\alpha} \tag{3}$$

$$k_{i,\alpha}=W_{k,\alpha}h_i+b_{k,\alpha} \tag{4}$$

可以得到向量序列 $[q_{1,\alpha},q_{2,\alpha},\cdots,q_{n,\alpha}]$ 和 $[k_{1,\alpha},k_{2,\alpha},\cdots,k_{n,\alpha}]$，作为识别第 α 种类型实体所使用的向量序列。下面定义 $s_\alpha(i,j)=q_{i,\alpha}^T k_{j,\alpha}$，作为 $span(i,j)$ 类型为 α 的实体的打分。即使用 $q_{i,\alpha}$ 与 $k_{i,\alpha}$ 的内积作为 $span(i,j)$ 是类型为 α 的实体的打分，其中 $span(i,j)$ 为序列 x 的第 i 个到第 j 个元素组成的连续子串。

在此基础上，由于 $W_{q,\alpha},W_{k,\alpha}\in R^{v\times d}$，其中 v 为预训练模型 MacBERT 的隐藏层维度，一般为 1 024，d 为可调节的超参数，设定为 128。因此，每增加一个新的实体类型，参数需要增加 $2vd$。相同条件下，传统的序列标记法的参数增加约 $2v$，可见一般指针网络的参数量远远大于 CRF。为了减少参数，实体识别任务可以分解为“抽取”和“分类”两个步骤，在“抽取”阶段判断 $s[i:j]$ 是否是实体，仅使用一组实体评分矩阵即 $(W_q h_i)^{\mathrm{T}}(W_k h_j)$ 为评分矩阵；对于“分类”过程，则使用特征拼接后再与全连接层运算来实现分类(即 $w_\alpha^{\mathrm{T}}[h_i:h_j]$)，将两项组合就可以得到最终的该序列是否为实体 α 的分数计算方式。

$$s_\alpha(i,j)=(W_q h_i)^{\mathrm{T}}(W_k h_j)+w_\alpha^{\mathrm{T}}[h_i:h_j] \tag{5}$$

2.3　特征融合全局指针网络

原始的全局指针网络，是利用预训练语言模型的最后一层输出，来进行实体和类别的判别计算。由于 MacBERT 的输出有 24 层，除了最后的一层输出之外，其他隐藏层的输出也包含着许多语义信息，高层注意力层更能捕获句子级别的语义和语境信息，这些层在捕获语义和上下文信息方面更强大。因此本文对指针网络做了进一步优化，提出了特征融合的全局指针网络(Global Pointer Network with Feature Fusion，GP-FF)，丰富和扩展 Global Pointer 打分基准，融合了 MacBERT 多层表征，从而提高 Global Pointer 的打分精度。使用 MacBERT 对文本序列处理得到字向量的计算过程如下：

$$H^i=\{h_1^i,h_2^i,\cdots h_n^i\}=\mathrm{MacBERT}(E_{\mathrm{input}}) \tag{6}$$

$$H^{\mathrm{new}}=\{h_1^{\mathrm{new}},h_2^{\mathrm{new}},\cdots h_n^{\mathrm{new}}\}=\{[wh_1:wh_1^i],[wh_2:wh_2^i],\cdots[wh_n:wh_n^i]\} \tag{7}$$

其中，H^i 表示 MacBERT 中的各层输出，H^{new} 表示多特征融合后的最终特征。

本文中首先提取 MacBERT 倒数第二层隐藏层高层信息也就是第 23 层的隐藏特征与最后的输出隐藏特征，通过一层参数共享的线性层变换，然后进行特征的拼接融合。其中，$w\in R^{v\times d}$，与第 23 层特征融合后的 H^{new} 计算过程如下：

$$H^{\mathrm{new}}=\{h_1^{\mathrm{new}},h_2^{\mathrm{new}},\cdots h_n^{\mathrm{new}}\}=\{[wh_1:wh_1^{23}],[wh_2:wh_2^{23}],\cdots[wh_n:wh_n^{23}]\} \tag{8}$$

使用融合后的多层特征代替单一隐藏层特征进行实体打分计算，得到优化后的指针网络计算分数 $s_\alpha(i,j)$ 为：

$$s_\alpha(i,j)=(W_q h_i^{\mathrm{new}})^{\mathrm{T}}(W_k h_j^{\mathrm{new}})+w_\alpha^{\mathrm{T}}[h_i:h_j] \tag{9}$$

同时，除了优化实体打分部分外，对于实体的类别进行分类打分时也可以采用隐藏特征融合的方法。将两个隐藏特征拼接融合得到新的隐藏特征 $h_i^{'}$：

$$h_i^{'}=w_1[h_i:h_i^{23}] \tag{10}$$

其中，$w_1\in R^{2v\times v}$ 用于与拼接后的隐层特征计算，从而融合两个特征向量。最终的指针网络计算分数 $s_\alpha(i,j)$ 为：

$$s_\alpha(i,j)=(W_q h_i^{\mathrm{new}})^{\mathrm{T}}(W_k h_j^{\mathrm{new}})+w_\alpha^{\mathrm{T}}[h_i^{'}:h_j^{'}] \tag{11}$$

2.4　数据增强训练

语言大模型[16-17](Large Language Model)在上

下文学习方面表现出了令人印象深刻的能力,仅以少数特定任务示例作为演示,已经能够为新的测试输入生成令人印象深刻的结果。在语境学习的框架下,这些语言大模型在包括机器翻译在内的各种自然语言处理(NLP)任务中取得了显著的成功,充分证明了语言模型已经具备深度语义理解的能力,并蕴含着丰富的知识。尽管在一些方面取得了进展,但语言大模型在命名实体识别(NER)任务方面的性能仍远低于有监督学习的基准模型,这是因为NER 任务与语言大模型之间存在根本差异。NER 任务本质上是一项语义任务,要求模型为句子中的每个标记分配正确的实体类型标签,而语言大模型则主要针对文本生成任务进行训练。语义标注任务和文本生成模型之间的这种本质差异导致语言大模型在处理 NER 任务时性能不佳,但语言大模型所蕴含的丰富知识和经验不能忽视。基于这一观点,本文提出了一种基于语言大模型的数据增强策略(LLM Data Augmentation, LLMDA),如图 2 所示。

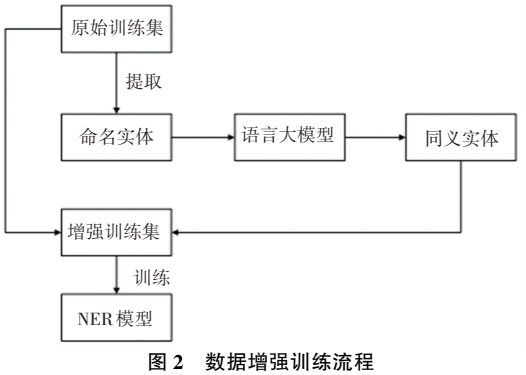

图 2 数据增强训练流程

Fig. 2 Data augmentation training workflow diagram

首先,从标注数据中提取命名实体;然后在获取的命名实体的基础上,结合提示信息将其输入到大模型中,以获取同义实体;最后整合原始语料,将命名实体替换为同义实体,从而生成增强的数据集。

通过 LLM 数据增强,可以有效的引入新实体,而且极大程度的保留了语句的原始信息,不会破坏模型已学习的语义信息,并且能够增强模型的鲁棒性和准确率。

假设,句子中存在 n 个实体关键词 $X=\{X_1,X_2,\cdots,X_n\}$,每一个实体 X_i 在通过大模型数据增强后,获取的同义实体的数量为 m_i 条,需要在原始数据集中的 n 个实体中挑选 j 个实体进行替换,每次替换可从 m 条同义实体中选取一个实体,遍历所有关键词后的方案数 T 的公式如下:

$$T=C_n^j\cdot\prod_{i=1}^{j}m_i \tag{12}$$

数据增强示例见表 1。

表 1 数据增强示例

Table 1 Data augmentation example

医学文本	句子中的实体	LLM 计算返回同义实体
持续 1~4 天,常有低热、头痛、乏力、咽痛、腹痛、烦躁等。	低热	微热、轻微发热
	头痛	头疼、头晕
	头	脑袋、颅部
	乏力	无力、疲惫、无精打采
	咽痛	喉咙痛,喉咙疼痛、咽喉痛、咽疼
	咽	喉咙、咽喉
	腹痛	腹部疼痛、腹部不适、腹部隐痛
	腹	肚子、腹部
	烦躁	焦躁、烦闷、烦恼

2.5 损失函数

本文模型采用多标签分类交叉熵损失[18],基于单目标多分类交叉熵损失改进实现。多标签分类交叉熵损失的优势在于没有类别不均衡现象,因为其不是将多标签分类变成多个二分类问题,而是转化成目标类别得分与非目标类别得分的两两比较,并且借助于 logSumExp 的良好性质,自动平衡了每一项的权重。相比于精调权重等方式,该损失函数可以在不需要过多调参的情况下达到较优的效果。其表达公式如下:

$$\log\Big(1+\sum_{(i,j)\in T_\alpha}e^{-s_\alpha(i,j)}\Big)+\log\Big(1+\sum_{(i,j)\in O_\alpha}e^{s_\alpha(i,j)}\Big) \tag{13}$$

其中,T_α 是该样本的所有类型为 α 的实体的首尾集合,O_α 是该样本的所有非实体或者实体类型非 α 的实体的首尾集合,其中只有当 $i\leqslant j$ 时,实体才有意义,所以只需考虑 $i\leqslant j$ 的组合,即:

$$\Omega=\{(i,j)\mid 1\leqslant i\leqslant j\leqslant n\}$$

$$T_\alpha=\{(i,j)\mid span_{[i,j]}\text{ 是类型为 }\alpha\text{ 的实体}\}$$

$$O_\alpha=\Omega-T_\alpha$$

其中,Ω 是所有符合 $i\leqslant j$ 的候选序列,在进行命名实体识别模型训练时,所有满足 $s_\alpha(i,j)>0$ 的片段 $span_{[i,j]}$ 都视为实体类型为 α 的实体输出。

3 实验

3.1 数据集

实验使用 CMeEE-V2[19] 中文医学命名实体识别数据集和 CCKS2019 数据集见表 2。CMeEE-V2 数据集包含疾病(dis)、临床表现(sym)、药物(dru)、医疗设备(equ)、医疗程序(pro)、身体(bod)、医学检验项目(ite)、微生物类(mic)、科室(dep)9 大类,数据集采用 *span* 标注的形式存在嵌套实体,其中训练集 15 000 条,验证集 5 000 条,测

试集 3 000 条，相比 CMeEE 原始版本，CMeEE-V2 的更新主要是修复了原始数据中的部分标注错误，提升了语料质量。CCKS2019 数据集的文本来自电子病历，包含疾病与诊断、影像检查、实验室检查、手术、药物和解剖 6 类实体。

表 2　实验数据集的统计

Table 2　Statistics of experimental datasets

数据集	标签	训练集	验证集	测试集	总样本
CMeEE-V2[19]	疾病、临床表现、药物、医疗设备，医疗程序、身体、医学检验项目、微生物类、科室	15 000	5 000	3 000	23 000
CCKS2019	解剖部位、疾病和诊断、药物、手术、影像检查、实验室检验	800	200	379	1 379

3.2　评价指标

模型评估采用准确率（*Precision*）、召回率（*Recall*）和 *F*1 分数（*F*1 - Score）作为评价指标，其表达形式如下：

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

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

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{16}$$

其中，*TP* 表示正确识别当前实体类别的样本数量；*FP* 表示错误识别当前类别的样本数量；*FN* 表示本属于当前类别但没有被识别到的样本数量；*TP* + *FP* 则表示被识别为当前类别的所有样本数量；*TP* + *FN* 则表示被标注为当前类别的所有样本数量。

3.3　参数设置

在实验中，采用 chinese_macbert_large 预训练模型，并将批量大小设置为 8；使用 Adam 优化器，学习率设为 2e-5，并将预热超参数设置为 0.1，在数据集上进行了 10 轮训练。实验参数设置见表 3。

表 3　实验参数设置

Table 3　Experimental parameter configuration

超参数	取值
Epoch	10
Batch size	8
Learning Rate	2e-5
Dropout	0.1
Optimizer	Adam
Warm up	0.1

3.4　消融实验

为验证模型不同模块的有效性，本文设计了模型消融实验。实验结果见表 4。

表 4　消融实验结果

Table 4　Lesion experiment results　%

模型	CMeEE-V2			CCKS2019		
	Precision	*Recall*	*F*1	*Precision*	*Recall*	*F*1
Bert-BiLSTM-CRF	71.76	66.83	69.21	73.84	75.31	74.57
MacBERT-GP	73.67	74.94	74.30	80.06	69.91	74.64
MacBERT-GP-LLMDA	74.70	74.23	74.46	81.09	70.36	75.34
MacBERT-GP-FF	75.08	75.00	75.04	80.87	70.66	75.42
本文模型	**75.74**	**75.18**	**75.45**	**81.23**	**77.89**	**79.52**

参与消融实验模型包括：

（1）Bert-BiLSTM-CRF 是 NER 领域的经典模型，在诸多任务上取得了优秀的识别效果。

（2）MacBERT-GlobalPointer 在模型 1 的基础上，使用 MacBERT 作为预训练语言模型，以新颖的 MLM 作为校正预训练任务，减轻了预训练和微调的差异。

（3）MacBERT-GlobalPointer-LLMDA 在模型 2 的基础上，使用大模型扩展数据进行了数据增强。

（4）MacBERT -GP-FF 在模型 2 的基础上，优化了 Global Pointer 方法，使用特征融合的 Global Pointer，充分利用不同 Transformer 层的输出，丰富

Global Pointer 语义信息。

(5)本文模型为结合模型 3 和模型 4 得到的最终模型。

从表 4 中可以看出,本文模型相较于其他模型,对于命名实体识别的性能更加优异,相较于 BERT-BiLSTM-CRF 模型,在 CMeEE-V2 数据集上的 *F*1 分数提高了 6.24%,验证了使用大模型数据增强和特征融合 Global Pointer 方法的实体识别模型的有效性。将 Bert-BiLSTM-CRF 模型和 MacBERT-GlobalPointer 模型进行对比可以看出,由于在 BERT 训练时的掩蔽策略是固定的,即随机掩蔽一定比例的输入词语。而 MacBERT 采用了使用类似的单词来达到屏蔽目的,从而增加了数据的多样性,有助于模型更好地理解上下文。BERT 中使用了固定的词表大小,而 MacBERT 在训练过程中使用全字掩码技术,这允许模型处理更多的词汇,尤其是在一些低频词汇的处理上。实验结果证明 MacBERT 模型更适用于中文医疗实体识别任务。

此外,使用大语言模型增强数据训练的 MacBERT-Global Pointer 模型,其 *F*1 分数高于未使用大语言模型增强数据的 MacBERT-GlobalPointer 模型,这是因为使用增强后的数据扩展了数据的广度,丰富了语义信息,从而提高了模型的性能。使用了特征融合 Global Pointer 的 MacBERT-Global Pointer 模型,结果优于未使用特征融合打分 Global Pointer 的模型,表明使用特征融合 Global Pointer 方法对模型性能具有一定提升效果。原始的 Global Pointer 仅仅基于一层表征输出构建实体矩阵判别,但是实际上在 Bert 及其改进模型 MacBERT 中,每一层都学习到一定的语言学信息,越深层的输出,学习到的语言学信息越多,优化后的 Global Pointer 通过多层表征融合,将融合后的最终输出作为构建实体判别矩阵的表征,可以全面的利用 MacBERT 中的特征。实验结果表明,特征融合的 Global Pointer 能够有效提升模型的识别效果。

3.5 对比实验

本文在 CMeEE-V2 和 CCKS2019 数据集上进行的实验结果见表 5。

表 5 本文模型与其他 NER 方法性能比较

Table 5 Performance comparison of this model with other NER methods %

模型	CMeEE-V2			CCKS2019		
	Precision	*Recall*	*F*1	*Precision*	*Recall*	*F*1
Bert-BiLSTM-CRF	71.76	66.83	69.21	73.84	75.31	74.53
RoBERTa-BiLSTM-CRF	73.64	75.22	74.42	74.62	74.38	74.50
MedBERT-BiLSTM-CRF	74.82	75.24	75.03	78.26	77.62	77.94
RoBERTa-GP	73.89	74.23	74.06	79.93	70.28	74.80
本文模型	**75.74**	**75.18**	**75.45**	**81.23**	**78.86**	**79.52**

相比于基线模型,使用 RoBERTa 模型作为编码器生成字向量,本文方法的 *F*1 值均有不同程度的提升,进一步验证了本文所提出模型的有效性。使用 RoBERTa 模型[20]作为编码器生成字向量,后接 BiLSTM-CRF 结构,在预训练过程中的语义特征相比于 Bert 有所区别,主要是使用了全词掩码以及基于更大的原始语料进行训练;后接 Global Pointer 进行实体矩阵运算,在训练过程中使用全词掩码,因而能够获得词级别的表征。MedBERT[21]在 6.5 亿字符中文临床自然语言文本语料上基于 Bert 模型预训练得到。

4 结束语

本文针对医学命名实体识别任务中专业标注数据稀少,Bert 模型无法获取词级信息,以及 Global Pointer 模型表征单一的问题,提出了基于 MacBERT-GP 的中文医学命名实体识别方法。该模型将优化了 Global Pointer 实体矩阵计算方法,使用多层语义输出构建实体矩阵解析,同时利用大语言模型进行数据增强,扩充了数据集,经在 CMeEE-V2 命名实体数据集和 CCKS2019 数据集上的试验,验证了本文方法的有效性。此外,本文提出的方法不局限于医学实体命名识别任务,大模型丰富的语言能力,可以扩展到各个领域,在未来的工作中,希望能够探索该方法在其他领域命名实体识别任务中的表现。

参考文献

[1] SHARNAGAT R. Named entity recognition: A literature survey [D]. Center For Indian Language Technology, 2014: 1-27.

[2] 李丹, 徐童, 郑毅, 等. 部首感知的中文医疗命名实体识别[J]. 中文信息学报, 2020, 34(12): 54-64.

[3] LI J, SUN A, HAN J, et al. A survey on deep learning for named

entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(1): 50-70.
[4] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436-444.
[5] WU G, TANG G, WANG Z, et al. An attention-based BiLSTM-CRF model for Chinese clinic named entity recognition[J]. IEEE Access, 2019(7):113942-113949.
[6] LIU X, HERSCH G L, KHALIL I, et al. Clinical trial information extraction with BERT[C]//Proceedings of the 9th International Conference on Healthcare Informatics. Piscataway, NJ:IEEE, 2021: 505-506.
[7] SU J, MURTADHA A, PAN S, et al. Global pointer: Novel efficient span-based approach for named entity recognition[J]. arXiv preprint arXiv,2208.03054, 2022.
[8] CUI Y, CHE W, LIU T, et al. Revisiting pre-trained models for Chinese natural language processing[J]. arXiv preprint arXiv, 2004.13922, 2020.
[9] MORWAL S, JAHAN N, CHOPRA D. Named entity recognition using hidden Markov model[J]. International Journal on Natural Language Computing, 2012, 4(1): 15-23.
[10] JU Z, WANG J, ZHU F. Named entity recognition from biomedical text using SVM[C]//Proceedings of the 5th International Conference on Bioinformatics and Biomedical Engineering. Piscataway,NJ:IEEE, 2011: 1-4.
[11] SONG S, ZHANG N, HUANG H. Named entity recognition based on conditional random fields[J]. Cluster Computing, 2019, 22(3): 5195-5206.
[12] GUI T, MA R, ZHANG Q, et al. CNN-Based Chinese NER with Lexicon Rethinking[C]//Proceedings of the International Joint Conference on Artificial Intelligence. IJCAI, 2019: 4982-4988.
[13] CHOWDHURY S, DONG X, QIAN L, et al. A multitask bi-directional RNN model for named entity recognition on Chinese electronic medical records[J]. BMC Bioinformatics, 2018, 19(17): 75-84.
[14] OUYANG E, LI Y, JIN L, et al. Exploring n-gram character presentation in bidirectional RNN-CRF for Chinese clinical named entity recognition[C]//Proceedings of the CEUR Workshop. CEUR, 2017: 37-42
[15] DONG C, ZHANG J, ZONG C, et al. Character-based LSTM-CRF with radical-level features for Chinese named entity recognition[C]//Proceedinfgs of the 24th International Conference on Computer Processing of Oriental Languages. 2016: 239-250.
[16] BROWN T, MANN B, RYDER N, et al. Language models are few-shot learners[C]// Proceedings of Advances in Neural Information Processing Systems. 2020: 1877-1901.
[17] OUYANG L, WU J, JIANG X, et al. Training language models to follow instructions with human feedback[C]// Proceedings of Advances in Neural Information Processing Systems. 2022: 27730-27744.
[18] SU J, ZHU M, MURTADHA A, et al. Zlpr: A novel loss for multi-label classification[J]. arXiv preprint arXiv,2208.02955, 2022.
[19] ZHANG N, CHEN M, BI Z, et al. Cblue: A Chinese biomedical language understanding evaluation benchmark[J]. arXiv preprint arXiv,2106.08087, 2021.
[20] LIU Y, OTT M, GOYAL N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv, 1907.11692, 2019.
[21] RASMY L, XIANG Y, XIE Z, et al. Med-BERT: Pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction[J]. NPJ Digital Medicine, 2021, 4(1): 86.