基于边界平滑的中文医疗嵌套命名实体识别方法

刘　纳[1,2], 吴克东[1,2], 刘　磊[1,2], 季　喆[1,2], 周雪雨[1,2]

(1.北方民族大学 计算机科学与工程学院,宁夏 银川 750021;2.北方民族大学 图像图形智能处理国家民委重点实验室,宁夏 银川 750021)

摘　要: 医疗语料普遍存在多层次、多粒度语义与实体重叠的问题,现有方法易出现边界预测过度置信与边界不确定性建模不足,难以有效建模实体间的嵌套关系。因此,提高对实体边界的预测能力成为解决该问题的关键。针对此问题,提出了一种基于边界平滑方法的中文医疗嵌套命名实体识别模型,结合改进的跨度编码策略以优化识别效果。模型通过 RoBERTa-wwm-ext-large 获取词级语义表示,结合 BiLSTM 建模长距离依赖;识别层采用全局指针统一定位实体起止边界,结合旋转位置编码显式编码相对位置信息,并通过双仿射解码器强化首尾交互完成跨度级判别;训练阶段引入边界平滑正则,对标注及其邻域跨度按距离分配软标签,以抑制硬边界噪声与过度置信,提升边界校准与召回能力。实验结果表明,模型在 CMeEE、CMeEE-V2 和 CLUENER2020 数据集上的 $F1$ 值均取得了显著提升,验证了该方法能够有效缓解中文医疗文本中的边界不确定性与嵌套干扰,具备较好的准确性与泛化能力。

关键词: 嵌套命名实体识别; 中文医疗文本; 边界预测; 边界平滑; 预训练语言模型

中图分类号: TP391.1　　**文献标志码:** A　　**doi:** 10.13705/j.issn.1671-6833.2026.02.014

命名实体识别(named entity recognition, NER)是自然语言处理中的核心任务,旨在从非结构化文本中提取具有语义类别的实体。传统方法多将其视为序列标注问题,即平面命名实体识别(flat NER, FNER),假设每个词或字符仅对应一个实体标签。然而,该假设难以覆盖现实语料中的复杂结构,尤其在医学领域,实体常存在包含或重叠关系,FNER 在多粒度语义建模上存在明显不足。为此,嵌套命名实体识别(nested NER, NNER)被提出,专注于识别嵌套结构,成为当前 NER 研究的重要方向。

在医学文本中,嵌套实体尤为常见,涵盖疾病、药物、治疗手段等多种概念。其结构复杂、术语密集,使 NNER 在医疗信息抽取与语义理解中具有关键作用[1]。准确识别嵌套实体不仅有助于提升临床决策效率,也为医学知识图谱构建与智能问答等下游任务提供重要支撑。

近年来,深度学习在生物医学 NER 中表现突出。BiLSTM-CRF[2]通过上下文建模与序列解码提升标注一致性,成为早期主流框架。2024 年,Hua 等[3]引入多标签局部度量与词性特征,缓解样本不均衡,拓展了多标签嵌套实体识别新思路。2025 年,郑国风等[4]将提示调优与对比学习结合,在小样本医疗语料上实现性能提升,但在嵌套边界区分上仍显不足。范锦涛等[5]构建融合对比学习的边界建模方法,利用双编码与标签语义参数化增强表示空间,有效提升了嵌套实体的识别能力。

预训练语言模型的引入使 NER 性能得到进一步提升,微调策略为下游任务提供更精准的上下文表征。2024 年,刘昕等[6]结合 MacBERT 上下文建模能力与 R-Drop 正则化以缓解训练与推理阶段不一致问题,但受限于小规模且标注不一致的领域数据,对复杂实体结构的识别仍然有限。2025 年,Yan 等[7]通过融合实体关联与门控上下文感知策略,以 RoBERTa-wwm-ext 强化全局信息一致性,在中文医疗嵌套实体识别任务中取得显著性能提升,但对大规模标注数据依赖较强,限制了低资源下的泛化性。

收稿日期: 2025-12-09;**修订日期:** 2026-02-07

基金项目: 国家自然科学基金资助项目(62162001);宁夏重点研发计划引才专项项目(2024BEH04020);北方民族大学校级科研项目(2024XYZJK01);北方民族大学研究生创新项目(YCX24361)

作者简介: 刘纳(1986—),女,宁夏银川人,北方民族大学讲师,博士,主要从事数据挖掘与自然语言处理技术研究,E-mail:liuna@nun.edu.cn。

传统序列标注方法在刻画嵌套层级关系时存在显著局限。近年来,NNER 被建模为跨度分类任务,即对候选跨度进行实体类别判定。Yu 等[8]提出 Biaffine 模型,引入双仿射机制强化首尾词元交互,并通过解码器对各跨度整体评分,从全局视角高效定位潜在实体,奠定了后续研究基础。全局指针方法[9]采用端到端建模方式联合判定实体位置与类别标签,显著提升了处理效率。

受中文词界不显著影响,边界判定尤为关键。2024 年,闫璟辉等[10]结合嵌套规则与改进的多头选择机制,将嵌套识别任务分解为边界识别与首尾关系判定的两个子任务,通过联合训练提升了中文医疗嵌套实体的边界利用效率 2025 年,杨采薇等[11]在范锦涛等[5]基础上引入多尺度差分卷积,增强了复杂结构下的边界感知能力,但在医疗语料与深层嵌套条件下的边缘信息捕获仍显不足。

尽管现有方法在 NNER 任务中取得了显著进展,但在中文医疗场景下仍面临诸多挑战。首先,实体边界模糊与多粒度表达导致候选跨度空间庞大,边界判定不稳,削弱模型泛化能力。其次,现有跨度方法虽具覆盖与效率优势,但对边界不确定性与过度置信的建模不足,难以应对复杂嵌套结构。此外,医学术语复杂、标注一致性不足及类别分布失衡引入标签噪声,影响模型鲁棒性与校准性。

边界平滑方法旨在缓解噪声与过度置信,提升模型对复杂边界的泛化与校准能力。Szegedy 等[12]于 2016 年率先将标签平滑作为正则化策略引入图像分类任务,有效抑制过度置信并改善模型泛化,为后续研究提供了新思路。2022 年,Zhu 等[13]首次将边界平滑引入 NER 任务,将类别平滑扩展至实体边界的概率分布,以降低边界预测的过度置信并改善模型校准性。2023 年,Shen 等[14]以边界去噪扩散建模,从噪声跨度逐步复原清晰实体,增强了模型对边界不确定性的鲁棒性。2024 年,Deng 等[15]结合句法信息与边界平滑,有效缓解过度置信与边界分割误差,显著提升性能。2025 年,Gao 等[16]将标签平滑与联合边界预测耦合,在训练中引入平滑交叉熵以减轻硬边界噪声与过拟合,提升长实体与嵌套结构的边界判定与识别精度。

为应对上述挑战,本文提出了一种基于边界平滑的中文医疗嵌套命名实体识别方法——GPBBS(Global Pointer with Biaffine and Boundary Smoothing)。该方法基于跨度预测定位实体起止位置,并引入边界平滑机制,通过对实体及其邻域赋予递减权重,缓解过度置信与边界偏差,形成边界判别与跨度分类的互补机制,为分类器提供更有效的分类信息,提升模型对实体边界不确定性的容忍度与复杂场景下的识别鲁棒性。

1 中文医疗嵌套命名实体识别模型

为了增强中文医疗命名实体识别任务中对嵌套结构与实体边界的建模能力,本文提出一种融合预训练表征、上下文编码、全局边界建模与平滑优化的端到端识别框架。模型由表示层、编码层与识别层构成,分别用于获取词级语义表示、建模上下文依赖关系及实现实体跨度的精确识别。

输入序列经 RoBERTa-wwm-ext-large 获取词级上下文表征,结合 BiLSTM 捕捉双向依赖特征。识别阶段采用全局指针模块对所有候选跨度进行全局评分,并结合双仿射解码器对起止边界对进行实体类别判定,强化边界定位与类别建模能力。最后,引入边界平滑策略优化标签分布与边界预测精度,以缓解边界标注不确定性导致的训练偏差。模型结构如图 1 所示。

1.1 表示模块:RoBERTa-wwm-ext-large

鉴于中文以词为基本语义单元,且医学文本中

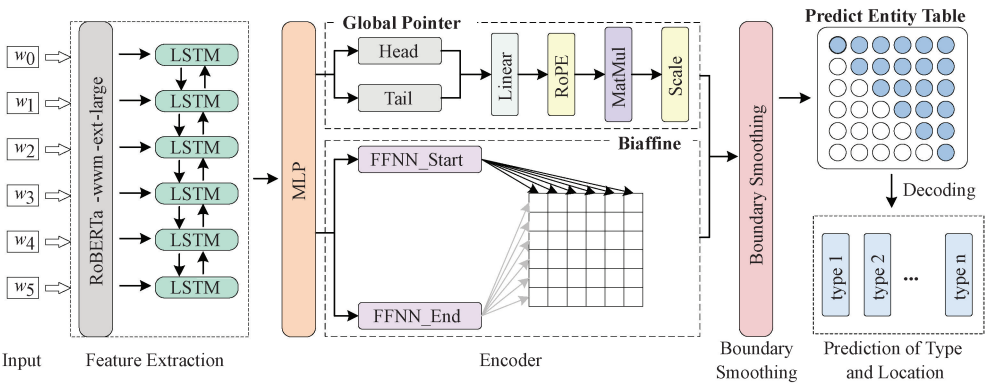

图 1 GPBBS 模型结构图

Figure 1 Diagram of the GPBBS model

多字术语与嵌套边界尤为常见,本文在表示层采用 RoBERTa-wwm-ext-large 作为表示模型。相较 BERT 的字符级子词掩码策略,RoBERTa-wwm-ext-large 在预训练阶段引入全词掩码(whole word masking, WWM)与动态掩码机制,使同一词单元的所有子字符同步掩码,并在不同训练轮次动态调整掩码位置,从而强化词级语义建模,减少字符级掩码引发的语义信息丢失。其训练策略扩大了语料规模与训练轮次,移除了“下一句预测”任务以简化训练目标,强化了对多字术语的连续建模能力,增强了对复杂实体边界的表达能力。

如表 1 所示,以“可致手掌多汗和体温下降”为例,“手”和“掌”共同构成一个高频出现的词,且具有独立完整语义。然而,字符级掩码易将“掌”视为独立单位,破坏“手掌”一词的整体语义;WWM 则将“手掌”整体同步掩码,保持词汇边界与语境一致性,增强模型对词级语义信息的感知能力,提升模型在中文医疗场景中对多字实体的表示质量。

表 1　掩码策略示例

Table 1　Masking strategy examples

掩码策略	处理后文本示例
Original Text	可致手掌多汗和体温下降
BERT	可致手[M]多汗和体[M]下降
RoBERTa-wwm-ext-large	可致[M][M]多汗和[M][M]下降

1.2　编码模块:BiLSTM

为增强对中文医学文本中复杂句法结构与长跨度实体依赖的建模能力,在表示层之后引入 BiLSTM 进行上下文编码。设输入序列长度为 n,表示层输出为 $\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$,其中 $\boldsymbol{x}_t$ 表示第 t 个位置的词级上下文表示。BiLSTM 由前向 LSTM 与后向 LSTM 组成,分别产生前向隐藏状态 $\overrightarrow{\boldsymbol{h}_t}$ 与后向隐藏状态 $\overleftarrow{\boldsymbol{h}_t}$。在任意时间步 $t\in\{1,2,\cdots,n\}$ 的最终隐藏状态定义为二者的特征维拼接,如公式(1)所示。

$$\boldsymbol{h}_t=[\overrightarrow{\boldsymbol{h}_t};\overleftarrow{\boldsymbol{h}_t}]。\tag{1}$$

将全序列的隐藏状态按时间步处理得到编码后的向量序列表示为 $\boldsymbol{H}=\{\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n\}$,为实体识别模块提供更具判别性的特征表示。BiLSTM 的门控机制可自适应调节信息流以缓解梯度消失,从而提升对长距离依赖与实体边界的建模效果。

1.3　识别模块:Global Pointer

使用两个独立模块分别预测实体的起始位置和结束位置,易导致训练与推理阶段不一致问题,进而影响模型的泛化能力。为提升识别一致性与稳定性,本文采用全局指针机制(global pointer, GP)[9]作为识别模块,通过全局归一化对实体边界进行整体判别,以增强对嵌套实体的建模鲁棒性。

对于任意给定句子,GP 在类别维度上构造上三角跨度评分矩阵,枚举所有满足 $i\leqslant j$ 的候选跨度 $s[i:j]$。每个矩阵对应一个实体类别,矩阵中的每个单元表示一个潜在实体跨度的置信度得分,单元 (i,j) 表示从位置 i 至 j 的连续子序列属于该类别实体的概率强度。以输入文本“可致手掌多汗和体温下降”为例,GP 构造了两个上三角矩阵,分别用于识别“身体部位”(bod)类实体和“临床症状”(sym)类实体(如图 2 所示)。

图 2　GP 策略示例

Figure 2　An example of GP strategy

在表征层面,对句子中的任意位置 i 与 j,经 BiLSTM 特征编码获取上下文表示,并通过全连接层得到每个位置的查询向量 $\boldsymbol{q}_{i,a}$ 与键向量 $\boldsymbol{k}_{j,a}$(其中 α 表示实体类别),分别组成对应的查询向量序列 $\boldsymbol{Q}=\{\boldsymbol{q}_{1,a},\boldsymbol{q}_{2,a},\cdots,\boldsymbol{q}_{n,a}\}$ 与键向量序列 $\boldsymbol{K}=\{\boldsymbol{k}_{1,a},\boldsymbol{k}_{2,a},\cdots,\boldsymbol{k}_{n,a}\}$。基于这些向量,类别 α 下跨度 $s[i:j]$ 的打分函数如公式(2)所示。

$$S_a(i,j)=\boldsymbol{q}_{i,a}^{\mathrm{T}}\boldsymbol{k}_{j,a}。\tag{2}$$

式中:$S_a(i,j)$ 表示从位置 i 到 j 的连续子序列属于类别 α 实体的置信度得分;$\boldsymbol{q}_{i,a}$ 与 $\boldsymbol{k}_{j,a}$ 分别对应起始与终止位置的表示。

为显式编码相对位置信息,在打分函数中采用旋转位置编码对起止向量进行位置相关的二维旋转变换,使打分与相对位置直接关联,优化实体边界表示。改进后的打分函数如公式(3)所示。

$$S_a(i,j)=(\boldsymbol{R}_i\boldsymbol{q}_{i,a})^{\mathrm{T}}(\boldsymbol{R}_j\boldsymbol{k}_{j,a})=\boldsymbol{q}_{i,a}^{\mathrm{T}}\boldsymbol{R}_{j-i}\boldsymbol{k}_{j,a}。\tag{3}$$

式中:$\boldsymbol{R}_t$ 为满足 $\boldsymbol{R}_i^{\mathrm{T}}\boldsymbol{R}_j=\boldsymbol{R}_{j-i}$ 的旋转变换矩阵。

该方法在不增加额外参数的前提下,增强了对实体边界与跨度相对位置的建模能力,并通过全局归一化提升实体识别精度。

1.4　解码模块:Biaffine Decoder

在 BiLSTM 提取的序列表示基础上,采用双仿射解码器(Biaffine Decoder)[8]进行实体边界预测。表示序列经两个前馈网络进行仿射变换后,得到起

始位置表示 $\boldsymbol{h}_i^s$ 与终止位置表示 $\boldsymbol{h}_t^e$。对于包含 c 个类别(含“非实体”)的实体识别任务,跨度 $s[i:j]$ 的类别评分向量 $\boldsymbol{r}_{ij}$ 如公式(4)所示。

$$\boldsymbol{r}_{ij} = (\boldsymbol{h}_i^s)^{\mathrm{T}}\boldsymbol{U}\boldsymbol{h}_j^e + \boldsymbol{W}(\boldsymbol{h}_i^s \oplus \boldsymbol{h}_j^e \oplus \boldsymbol{w}_{j-i}) + \boldsymbol{b}。\quad (4)$$

式中:$\boldsymbol{w}_{j-i}$ 为跨度宽度 $(j-i)$ 的可学习宽度嵌入;$\boldsymbol{U}$、$\boldsymbol{W}$、$\boldsymbol{b}$ 为可学习参数,$\oplus$ 表示向量拼接。

对 $\boldsymbol{r}_{ij}$ 在类别维进行归一化,得到最终类别概率 $\hat{\boldsymbol{y}}_{ij}$,如公式(5)所示。

$$\hat{\boldsymbol{y}}_{ij} = softmax(\boldsymbol{r}_{ij})。\quad (5)$$

若真实标签 $\boldsymbol{y}_{ij}$ 对应的类别为某一实体类,则在该类维度取值为 1,否则为 0。模型在所有候选跨度上通过标准交叉熵损失进行优化,如公式(6)所示。

$$\mathcal{L}_{CE} = -\sum_{0\leqslant i\leqslant j<n} \boldsymbol{y}_{ij}^{\mathrm{T}} log(\hat{\boldsymbol{y}}_{ij})。\quad (6)$$

在推理阶段,首先过滤预测结果中判定为“非实体”的跨度,然后按预测置信度对剩余跨度进行排序,并对边界冲突采用高置信度优先的抑制策略,保留一致的实体集合。

1.5 边界平滑(boundary smoothing)

为缓解 NNER 模型在标签预测过程中的过度置信问题,本文在空间特征解码后引入边界平滑策略,以显式降低模型对实体边界的置信度,提升预测结果在边界不确定条件下的鲁棒性。

传统方法通常假设实体边界具有确定性,即标注实体的边界概率为 1,邻近未标注跨度概率为 0,该策略被称为硬边界假设。如图 3(a)[13] 所示,在一个包含两个实体的示例句子中,其真实标注 $\boldsymbol{y}_{ij}$ 基于硬边界假设,仅对上三角区域中满足准确跨度的单元赋予非零概率。然而,实体边界常因上下文变化或标注主观性而模糊,尤其在中文中更为显著,硬边界假设未能有效建模这一模糊性,导致训练时对边界的过度置信,影响模型的泛化能力。

对此,本文引入边界平滑机制,通过重新分配实体边界的标签概率,使相邻区域获得一定概率,减少模型对严格边界约束的过度依赖。在设定平滑参数 ε 后,标注跨度 $s[i:j]$ 的原始边界概率为 $1-\varepsilon$,剩余概率 ε 被均匀分配给其周围跨度。在平滑范围 D 内,与标注实体的曼哈顿距离的所有邻近跨度均等共享概率 ε/D。最终,未分配概率的跨度被归类为“非实体”类别,从而形成平滑边界。图 3(b)展示了经边界平滑处理后标注实体边界的概率分布。

为优化模型的训练过程,本文采用引入边界平滑正则化的交叉熵损失函数,如公式(7)所示。

$$\mathcal{L}_{BS} = -\sum_{0\leqslant i\leqslant j<n} \tilde{\boldsymbol{y}}_{ij}^{\mathrm{T}} log(\hat{\boldsymbol{y}}_{ij})。\quad (7)$$

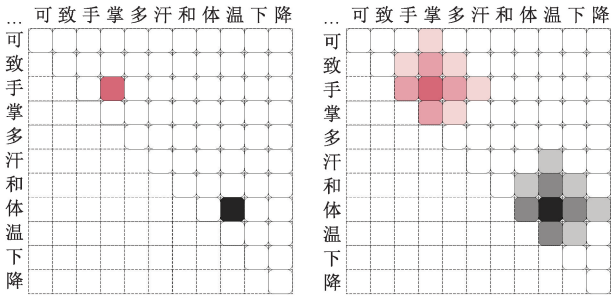

图 3 边界平滑策略

Figure 3 Boundary smoothing strategy

式中:$\tilde{\boldsymbol{y}}_{ij}$ 为平滑后的目标分布;$\hat{\boldsymbol{y}}_{ij}$ 为模型在类别维的预测分布。

通过边界平滑方法,模型能够在处理复杂边界时提升鲁棒性,并有效减少预测不确定性。

2 实验分析

2.1 数据集

本实验在 CMeEE 和 CMeEE-V2 中文医疗命名实体识别数据集上进行评估,并在中文通用领域数据集 CLUENER2020 上验证模型的泛化能力。数据集均来自阿里云天池实验室平台,具体信息如下。

1. CMeEE 数据集

CMeEE(Chinese Medical Entity Extraction)为中文医疗信息处理挑战榜 CBLUE(Chinese Biomedical Language Understanding Evaluation)[17] 评测中的医学命名实体识别子任务数据集,包含训练集 15 000 条、验证集 5 000 条和测试集 3 000 条,涵盖身体(bod)、疾病(dis)、症状(sym)等 9 类医学实体。

2. CMeEE-V2 数据集

CMeEE-V2 数据集[17] 在 CMeEE 原始版本基础上进行修正,纠正了部分标注错误,并将实体下标位置由左闭右闭标记法调整为左闭右开标记法,提升了标注规范性与数据质量。

3. CLUENER2020 数据集

CLUENER2020 数据集[18] 由清华大学基于其文本分类数据集 THUCTC 构建,进行细粒度实体标注,包含训练集 10 748 条、验证集 1 343 条和测试集 1 345 条,涵盖地址、书名、公司等 10 类通用实体。

2.2 实验基准与评估指标

为评估所提方法的有效性,本文选取多种命名实体识别模型作为基线模型,涵盖当前主流建模范式,包括序列标注方法(BiLSTM-CRF[2]),层级建模方法(Pyramid[19]),基于跨度的识别方法(Biaffine[8]、GlobalPointer[9] 及其变体 Efficient GlobalPointer[9]),预训练集成模型(PT-CL[4]),以及边界

增强方法(Diffusionner[14]、SLRBC[20]、BANER[21])。

精确率 P(Precision)、召回率 R(Recall)和 $F1$ 分数(F1-score)是通用的评价指标,三者能够综合反映模型的整体性能,因此,本文评估模型的性能主要以三者作为参照,计算方法如公式(8)-(10)所示。

$$P = \frac{TP}{TP + FP}; \tag{8}$$

$$R = \frac{TP}{TP + FN}; \tag{9}$$

$$F1 = \frac{2 \times P \times R}{P + R}。 \tag{10}$$

式中:TP 表示正类预测为正样本的数量,FP 表示负类预测为正样本的数量,FN 表示正类预测为负样本的数量。仅当模型准确预测实体的起始和结束下标且实体类型精准匹配时计为预测正确。

2.3 实验结果与分析

2.3.1 领域数据集中的性能对比

在中文医疗命名实体识别数据集 CMeEE、CMeEE-V2 和中文通用领域命名实体识别数据集 CLUENER2020 中,通过与基线方法进行对比实验,验证了本文提出的 GPBBS 模型在中文嵌套命名实体识别任务中具有更好的性能表现。各数据集均采用 RoBERTa-wwm-ext-large 作为预训练模型,该模型在生物医学语料上进行全实体掩码和全跨度掩码策略的预训练,有助于提升中文医疗 NER 任务中的表现。实验结果如表 2 所示。

实验结果表明,本文提出的 GPBBS 模型在多个数据集上均优于其他主流基线方法,并在整体 $F1$ 值上取得最优性能。在 CMeEE 数据集上,相较传统序列标注方法 BiLSTM-CRF,$F1$ 值提升 10.69 个百分点;相较具备嵌套实体识别能力的 Pyramid、Biaffine、GlobalPointer、Efficient GlobalPointer 与 Diffusionner,$F1$ 值分别提升 3.00、3.90、1.77、1.15 和 0.31 个百分点;与面向平面实体识别的 SLRBC、PT-CL 与 BANER 对比时,$F1$ 值分别提升 1.25、0.55 和 0.14 个百分点。

在 CMeEE-V2 数据集上,性能提升与上述趋势一致。相较当前最优嵌套基线,$F1$ 值提升 0.52 个百分点,验证了其对高标注一致性数据的适应能力。该结果表明,预训练模型的全词掩码策略能有效增强医疗语境下语义表达,而边界平滑机制在该一致性语料上仍能产生稳定增益,体现出对边界偏差与标签噪声的鲁棒性。

在 CLUENER2020 数据集上,本文模型取得 81.18%的 $F1$ 值,相较 BiLSTM-CRF、Diffusionner 与 BANER,分别提升了 11.18、0.88 与 0.24 个百分点,表现出良好的跨域泛化能力。

GPBBS 模型在 NNER 任务中具有较强的召回能力与良好的泛化表现,但在部分对比中精确率略低。如 CMeEE 数据集中,模型的精确率较 Efficient GlobalPointer 与 Diffusionner 分别低 0.32 与 0.04 个百分点,这与模型采用的全局边界建模与边界平滑策略有关,通过对邻近边界进行概率软分配,模型在处理边界模糊或实体重叠样本时更具包容性,显著提升召回率,但同时引入近边界的误检(FP),影响精确率。而其他方法在边界预测上更为保守,虽有助于提升预测精度,但召回能力较为有限。

CLUENER2020 数据集中亦呈现出类似趋势,精确率较 Biaffine 与 SLRBC 分别低 0.73 与 0.89 个百分点。全局指针的统一评分机制与双仿射解码器的首尾交互能力有助于在中文语境中捕获边界模糊的真实实体,边界平滑进一步抑制硬边界导致的过度置信,有效降低了漏检率,在提升召回率的同时带

表 2 CMeEE、CMeEE-V2 与 CLUENER2020 数据集实验结果

Table 2 The experimental results on CMeEE, CMeEE-V2, and CLUENER2020 datasets 单位:%

Method	CMeEE			CMeEE-V2			CLUENER2020		
	P	R	$F1$	P	R	$F1$	P	R	$F1$
BiLSTM-CRF[2]	60.39	51.35	55.50	65.40	61.39	62.70	71.06	68.97	70.00
Pyramid[19]	62.82	63.50	63.19	70.93	73.24	71.98	77.24	80.46	78.82
Biaffine[8]	64.17	61.29	62.29	72.64	72.94	72.34	81.39	71.95	76.13
GlobalPointer[9]	65.17	64.74	64.42	72.94	72.34	72.64	78.74	80.59	79.66
Efficient GlobalPointer[9]	**66.45**	64.69	65.04	74.14	73.30	73.53	79.66	81.09	80.35
SLRBC[20]	63.73	66.25	64.94	71.54	72.59	72.06	**81.55**	80.51	81.03
Diffusionner[14]	66.17	66.60	65.88	73.95	74.50	74.22	80.10	80.50	80.30
PT-CL[4]	65.50	65.78	65.64	74.26	74.07	74.16	80.47	81.06	80.76
BANER[21]	65.78	66.33	66.05	74.24	74.79	74.51	80.68	81.21	80.94
GPBBS(本文)	66.13	**67.10**	**66.19**	**74.48**	**75.25**	**74.74**	80.66	**81.99**	**81.18**

来更优的整体性能。

而在标注一致性更高、边界模糊程度较低的CMeEE-V2数据集中,模型在精确率与召回率之间取得良好平衡,验证了该模型在不同数据质量下的适应能力与鲁棒性。

结合以上分析,本文模型能够有效表征中文嵌套实体特征,在提升召回表现的同时保持了较高的整体性能,适用于存在实体重叠或边界不确定性较强的医疗文本场景。

2.3.2 预训练模型的比较分析

预训练模型的选择对于医疗NER任务的识别效果至关重要。为了验证不同预训练模型的融合效果,本文比较了七种预训练模型——BERT、ERNIE、XLNet、RoBERTa-wwm-ext-large、MacBERT、MCBERT和BioBERT的性能,这些模型的预训练权重均来自Hugging Face,并在CMeEE数据集上进行了微调。实验结果如图4所示。

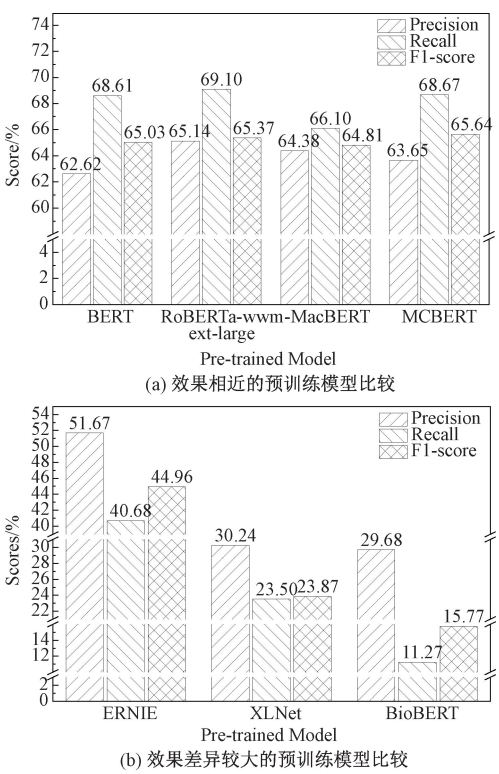

图4 不同预训练模型的评估结果对比

Figure 4 Evaluation results of different pre-trained models

实验结果表明,RoBERTa-wwm-ext-large在整体性能上优于其他模型,精确率和召回率分别达到65.14%和69.10%。其优势源于预训练阶段引入的全词掩码与动态掩码机制,较BERT的静态掩码策略能更有效地捕捉上下文信息并增强对长距离依赖的建模能力。

MacBERT通过同义词替换的掩码方式增强中文语义建模能力,在医疗细粒度实体识别任务中具有良好适应性。MCBERT基于中文医学语料进行预训练,在处理医学术语和复杂句法结构方面表现优异,$F1$值达65.64%,在术语识别和复杂结构建模方面优于BERT,但其精确率略低,受限于预训练语料的覆盖范围。

ERNIE和XLNet虽在大规模通用数据集上进行了预训练,在其他任务中表现较好,但由于缺乏医学领域知识,数据集存在较大差异,导致这两种模型在该数据集上的表现较差。尽管BioBERT以生物医学语料为基础进行预训练,但其预训练数据与CMeEE在语言分布与任务特性上存在偏差,导致模型难以有效识别医学文本中的实体,导致其效果远不如MCBERT明显。

2.3.3 消融实验

为探讨GPBBS模型中各组件对嵌套命名实体识别性能的贡献,在不引入新的解码范式的前提下进行消融实验,验证了各模块对模型性能的影响。实验通过在CMeEE数据集上逐一移除关键模块,结果如表3所示。

表3 消融实验

Table 3 Ablation experiment

Method	P	R	$F1$
w/o BS	65.00	62.86	63.64
w/o GP	65.76	63.81	64.52
w/o Biaffine	65.14	65.30	64.80
GPBBS	66.13	67.10	66.19

实验结果表明,移除边界平滑模块(w/o BS)后,模型$F1$值显著下降,验证其在缓解标签噪声与提升边界判别能力方面的关键作用,尤其适用于噪声较大的医疗文本。去除全局指针(w/o GP)和双仿射解码器(w/o Biaffine)分别导致$F1$值降至64.52%和64.80%,表明这两个模块在实体边界建模与类别判别中的重要性。上述模块与边界平滑协同有效,显著提升了嵌套实体识别性能。

2.3.4 参数敏感性分析

边界平滑方法受平滑强度ε与平滑范围D的影响。为评估其对模型影响,本文在CMeEE数据集上测试了多组参数组合,结果如表4所示。

实验结果表明,模型在$\varepsilon \in \{0.1, 0.2, 0.3\}$与$D \in \{1, 2\}$设置下进行训练,各参数组合均优于未使用边界平滑的基线模型,验证了该机制在提升边界预测准确性方面的有效性。当适度增强平滑强度时,模型$F1$值显著提升;当$\varepsilon > 0.20$时,强度过大使边界范围扩展过宽,引入更多非目标区域,导致精

表 4　超参数(ε, D)敏感性分析

Table 4　Sensitivity analysis of hyperparameter (ε, D)

(ε, D)	P	R	$F1$
(0.10,1)	65.30	67.00	66.14
(0.10,2)	**66.13**	**67.10**	**66.19**
(0.20,1)	66.10	65.30	65.70
(0.20,2)	65.32	66.80	66.05
(0.30,1)	65.00	65.24	65.12
(0.30,2)	64.80	66.50	65.64

度下降,整体性能减弱。平滑范围 $D=1$ 时性能提升有限,而 $D=2$ 在多组 ε 下能够有效提升召回率,整体表现更优。当 D 继续增大时,性能趋于饱和,部分场景下略有下降,说明过大的 D 引入冗余上下文信息,干扰边界判定的稳定性。

参数组合 ($\varepsilon=0.10, D=2$) 在本任务中表现最优,设为默认组合。整体结果表明,边界平滑方法在合理参数范围内具备良好的鲁棒性,能够有效增强模型对嵌套实体边界的建模能力。

3　结论

针对嵌套实体边界模糊与过度置信问题,本文提出了一种结合边界平滑策略的改进型跨度建模方法。通过在训练阶段引入权重分配机制强化边界上下文连续性,并在推理阶段采用边界平滑机制优化实体预测,有效降低了远距噪声干扰与边界偏差。实验结果表明,该方法在医疗和通用领域均表现出较强的泛化能力与应用价值,能够有效完成中文嵌套命名实体识别任务。

未来工作将进一步在更多领域验证方法有效性,并探索引入对抗训练和远程监督学习等方法以提升模型在嵌套命名实体识别任务中的表现。

参考文献:

[1] Goyal N, Singh N. Named entity recognition and relationship extraction for biomedical text: a comprehensive survey, recent advancements, and future research directions [J]. Neurocomputing, 2025, 618: 129171.

[2] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stuoudsburg: ACL, 2016: 260-270.

[3] Hua Zaifeng, Chen Yifei. Local Metric NER: a new paradigm for named entity recognition from a multi-label perspective [J]. Knowledge-Based Systems, 2024, 305: 112686.

[4] Zheng Guofeng, Liu Na, Li Chen, et al. Chinese medical named entity recognition based on prompt tuning and contrastive learning[J/OL]. Computer Engineering and Applications, . https://link.cnki.net/urlid/11.2127.tp.20240923.1435.004. [郑国风, 刘纳, 李晨, 等. 基于提示调优与对比学习的中文医疗命名实体识别方法[J/OL]. 计算机工程与应用, . https://link.cnki.net/urlid/11.2127.tp.20240923.1435.004.]

[5] Fan Jintao, Chen Yanping, Yang Caiwei, et al. Nested named entity recognition by contrastive learning with boundary information[J]. Journal of Computer Applications, 2025, 45(10): 3111-3120. [范锦涛, 陈艳平, 杨采薇, 等. 结合边界信息的对比学习嵌套命名实体识别[J]. 计算机应用, 2025, 45(10): 3111-3120.]

[6] Liu Xin, Xu Hongzhen, Liu Aihua, et al. Geological named entity recognition based on MacBERT and R-drop [J]. Journal of Zhengzhou University (Engineering Science), 2024, 45(3): 89-95. [刘昕, 徐洪珍, 刘爱华, 等. 基于 MacBERT 和 R-Drop 的地质命名实体识别[J]. 郑州大学学报(工学版), 2024, 45(3): 89-95.]

[7] Yan Yang, Kang Yufeng, Huang Wenbo, et al. Chinese medical named entity recognition utilizing entity association and gate context awareness[J]. PLoS One, 2025, 20(2): e0319056.

[8] Yu Juntao, Bohnet B, Poesio M. Named entity recognition as dependency parsing[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stuoudsburg: ACL, 2020: 6470-6476.

[9] Su Jianlin, Murtadha A, Pan Shengfeng, et al. Global pointer: novel efficient span-based approach for named entity recognition[PP/OL]. (2022-08-05)[2025-04-10]. https://doi.org/10.48550/arXiv.2208.03054.

[10] Yan Jinghui, Zong Chengqing, Xu Jin'an. Nested entity recognition approach in Chinese medical text[J]. Journal of Software, 2024, 35(6): 2923-2935. [闫璟辉, 宗成庆, 徐金安. 中文医疗文本中的嵌套实体识别方法[J]. 软件学报, 2024, 35(6): 2923-2935.]

[11] Yang Caiwei, Chen Yanping, Qin Yongbin, et al. A multi-scale semantic convergence difference operator for named entity recognition[J]. Journal of Chinese Information Processing, 2025, 39(6): 99-109. [杨采薇, 陈艳平, 秦永彬, 等. 多尺度语义收敛差分算子的命名实体识别方法[J]. 中文信息学报, 2025, 39(6): 99-109.]

[12] Szegedy C, Vanhoucke V, Ioffe S, et al. Rethinking the inception architecture for computer vision[C]//Proceedings of the 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2016: 2818-2826.

[13] Zhu Enwei, Li Jinpeng. Boundary smoothing for named entity recognition[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stuoudsburg: ACL, 2022: 7096-7108.

[14] Shen Yongliang, Song Kaitao, Tan Xu, et al. DiffusionNER: boundary diffusion for named entity recognition[C]//Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stuoudsburg: ACL, 2023: 3875-3890.

[15] Deng Zhenrong, Huang Zheng, Wei Shiwei, et al. KCB-FLAT: enhancing Chinese named entity recognition with syntactic information and boundary smoothing techniques [J]. Mathematics, 2024, 12(17): 2714.

[16] Gao Kai, Zhou Jiahao, Chi Yunxian, et al. TourismNER: a Tourism Named Entity Recognition method based on entity boundary joint prediction[J]. Intelligent Systems with Applications, 2025, 25: 200475.

[17] Zhang Ningyu, Chen Mosha, Bi Zhen, et al. CBLUE: a Chinese biomedical language understanding evaluation benchmark[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Stuoudsburg: ACL, 2022: 7888-7915.

[18] Xu Liang, Hu Hai, Zhang Xuanwei, et al. CLUE: a Chinese language understanding evaluation benchmark [C]//Proceedings of the 28th International Conference on Computational Linguistics. Barcelona: International Committee on Computational Linguistics, 2020: 4762-4772.

[19] Wang Jue, Shou Lidan, Chen Ke, et al. Pyramid: a layered model for nested named entity recognition[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stuoudsburg: ACL, 2020: 5918-5928.

[20] Cui Xiaohui, Yang Yu, Li Dongmei, et al. Fusion of SoftLexicon and RoBERTa for purpose-driven electronic medical record named entity recognition[J]. Applied Sciences, 2023, 13(24): 13296.

[21] Guo Quanjiang, Dong Yihong, Tian Ling, et al. BANER: boundary-aware LLMs for few-shot named entity recognition[C]//Proceedings of the 31st International Conference on Computational Linguistics. Kerrville: Association for Computational Linguistics 2025: 10375-10389.

A Boundary Smoothing-based Method for Chinese Medical Nested Named Entity Recognition

LIU Na[1,2], WU Kedong[1,2], LIU Lei[1,2], JI Zhe[1,2], ZHOU Xueyu[1,2]

(1. College of Computer Science and Engineering, North Minzu University, Yinchuan 750021, China; 2. The Key Laboratory of Images and Graphics Intelligent Processing of State Ethnic Affairs Commission, North Minzu University, Yinchuan 750021, China)

Abstract: Medical corpora commonly exhibit multi-level and multi-granularity semantics with overlapping entities. Existing approaches tend to produce overconfident boundary predictions and insufficient modeling of boundary uncertainty, which hinders effective representation of nested relations among entities. Strengthening boundary prediction is therefore essential. A Chinese medical nested named entity recognition model based on boundary smoothing is developed, together with an improved span-encoding strategy to enhance recognition. The model uses RoBERTa-wwm-ext-large to obtain token-level representations and employs a BiLSTM to capture long-range dependencies. In the recognition layer, a GlobalPointer uniformly locates start and end boundaries, Rotary Position Embedding explicitly encodes relative positional information, and a biaffine decoder strengthens head-tail interactions for span-level discrimination. During training, boundary-smoothing regularization assigns soft labels to annotated spans and their neighboring spans according to distance, which suppresses hard-boundary noise and overconfidence and improves boundary calibration and recall. Experiments on CMeEE, CMeEE-V2, and CLUENER2020 show significant improvements in $F1$, confirming that the method effectively mitigates boundary uncertainty and nested interference in Chinese medical text, with strong accuracy and generalization.

Keywords: nested NER; Chinese medical text; boundary prediction; boundary smoothing; pre-trained language model