面向淋巴水肿疾病的电子病历命名实体识别应用研究*

汤昊宬[1] 苏万春[2] 冀秀元[1] 信建峰[2] 夏 松[2] 孙宇光[2] 徐 毅[1] 沈文彬[2]

([1] 中国科学院自动化研究所 北京 100190 [2] 首都医科大学附属北京世纪坛医院 北京 100038)

〔摘要〕 **目的/意义** 探讨人工智能技术应用于淋巴水肿患者电子病历非结构化文本数据的关键实体识别问题。**方法/过程** 阐述样本稀缺背景下模型微调训练的解决方案，选取首都医科大学附属北京世纪坛医院淋巴外科既往收治患者 594 例为研究对象，依据临床医生标注的 15 种关键实体类别，微调 GlobalPointer 模型的预测层，借助其全局指针识别嵌套和非嵌套的关键实体。分析实验结果的准确性和临床应用可行性。**结果/结论** 微调后模型总体精准率、召回率和 Macro_*F*1 均值分别为 0.795、0.641 和 0.697，为淋巴水肿电子病历数据精准挖掘奠定基础。

〔关键词〕 淋巴水肿；电子病历；命名实体识别；自然语言处理；医学

〔中图分类号〕R-058 〔文献标识码〕A 〔DOI〕10.3969/j.issn.1673-6036.2024.02.009

Study on the Application of Named Entity Recognition in Electronic Medical Records for Lymphedema Disease

TANG Haocheng[1], *SU Wanchun*[2], *JI Xiuyuan*[1], *XIN Jianfeng*[2], *XIA Song*[2], *SUN Yuguang*[2], *XU Yi*[1], *SHEN Wenbin*[2]

[1] *Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China*; [2] *Beijing Shijitan Hospital, Capital Medical University, Beijing 100038, China*

〔**Abstract**〕 **Purpose/Significance** The paper discusses the application of artificial intelligence technology to the key entity recognition of unstructured text data in the electronic medical records of lymphedema patients. **Method/Process** It expounds the solution of model fine-tuning training under the background of sample scarcity, a total of 594 patients admitted to the department of lymphatic surgery of Beijing Shijitan Hospital, Capital Medical University are selected as the research objects. The prediction layer of the GlobalPointer model is fine-tuned according to 15 key entity categories labeled by clinicians, nested and non-nested key entities are identified with its global pointer. The accuracy of the experimental results and the feasibility of clinical application are analyzed. **Result/Conclusion** After fine-tuning, the average accuracy rate, recall rate and Macro_*F*1 of the model are 0.795, 0.641 and 0.697, respectively, which lay a foundation for accurate mining of lymphedema EMR data.

〔**Keywords**〕 lymphedema; electronic medical records; named entity recognition; natural language processing; medicine

〔修回日期〕 2023-07-24

〔作者简介〕 汤昊宬，工程师，发表论文 5 篇；通信作者：徐毅，沈文彬。

〔基金项目〕 科技创新 2030——“新一代人工智能”重大项目（项目编号：2020AAA0105005）；北京市科学技术委员会项目（项目编号：Z191100007619049）。

1 引言

淋巴水肿主要表现为局部体液滞留和组织肿胀，是全球致残率最高的疾病之一，严重危害人体健康，及时准确的诊断是阻断疾病恶化、提升术后康复痊愈率的关键。伴随着人工智能技术的飞速发展，疾病相关数据驱动的精准医学研究为此提供了行之有效的解决方案。研究者基于文本数据[1]、图像数据[2]在临床疾病辅助诊断领域已取得显著效果。患者电子病历[3]是医务人员借助医疗信息系统对临床治疗经过的记录，包括患者检查、诊断和治疗过程等重要医疗信息，通常以半结构化或非结构化形式存储，是构建智能化诊疗分析系统的数据基础。但是电子病历记录具有明显的子语言特性[3]，例如包含大量专业术语和行业习惯用语、表达模式化、数字和单位混合（如 6.0～8.0 mmol/L）、句子语法结构不完整等，数据噪声显著，呈异质性分布，尤其是针对同种疾病，不同医生遵循不同标准或习惯书写病历，存在一词多义和多词一义等不规范的现象，并且相较于英文语料缺乏明显的边界分隔符，词频分布呈现厚尾效应，严重影响双向编码器表征（bidirectional encoder representations from transformer, BERT）[4]等序列化语义分析技术的使用。因此，电子病历文本数据挖掘往往需要人工提取关键信息，依赖于高年资临床医生的精细标注，标注过程耗时费力，电子病历标注语料稀缺，尤其体现在亚专业学科。由此可见，针对淋巴水肿电子病历文本数据的智能化预处理或信息提取尤为重要。

命名实体识别（named entity recognition, NER）技术可以从文本中检测关键实体的范围和语义类别，是目前从非结构化文本数据中进行信息抽取的关键技术之一[5]。在电子病历数据中，实体重叠是相当普遍的现象，见图 1。“左下肢”与“淋巴水肿”首尾不相交，为非嵌套实体，而“手术后淋巴水肿”包含更细粒度的“淋巴水肿”实体，为嵌套实体。如果忽略嵌套实体，则无法捕获底层文本中更细粒度的语义信息。针对该问题，基于超图[6]、序列标注[7-8]和区域设置[9]的方法存在计算复杂度高、错误级联、准确率低等问题。而 GlobalPointer 模型[10]无需复杂的特征工程，采用全局指针在中文嵌套实体识别任务中取得了最优效果。因此，本研究利用 GlobalPointer 模型和模型微调方法实现少量标注样本背景下的淋巴水肿电子病历命名实体识别模型训练，并选取基准模型进行比较，建立高质量电子病历标注文本语料库，构建人工智能技术辅助淋巴水肿疾病精准诊断、分期研究和应用的关键数据基础。

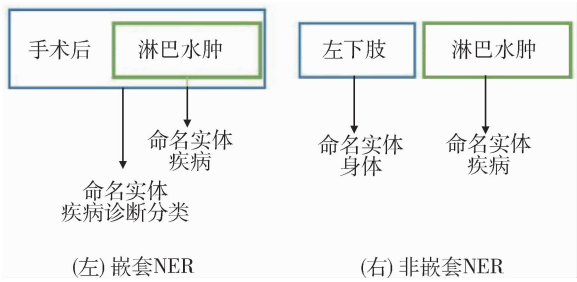

图 1 命名实体识别任务分类

2 模型介绍

2.1 预训练语言模型

GlobalPointer 模型以预训练语言模型为编码器提取文本特征。BERT 是预训练语言模型之一，由多层编码器堆叠而成，采用完全自注意力机制，计算每个词与其他所有词的关联，在自然语言处理领域取得了显著效果，但其时间和空间复杂度与序列长度为二次方关系 $O(n^2)$，可以处理的最大序列长度为 512 字符，长文本处理能力受限。BigBird 模型[11]是另一种预训练语言模型，同样由多层编码器堆叠构成，但区别于 BERT 普通的多头注意力机制，其采用稀疏的多头注意力机制，将时间和空间复杂度降低为线性 $O(n)$，运行效率更高，可以处理的最大序列长度为 4 096 字符，是 BERT 的 8 倍，适用于本研究中的长文本电子病历。因此，采用 BigBird 模型作为 GlobalPointer 模型的编码器。注意力值计算方式如下：

$$\mathrm{ATTN}_D(X)_i = x_i + \sum_{h=1}^{H}\sigma(Q_h(x_i)K_h(X^T_{N(i)}) \times V_h(X_{N(i)}) \quad (1)$$

其中，Q_h 和 K_h 分别是查询函数和键函数，V_h 是

值函数，σ 是评分函数，H 表示头数（Head），$N(i)$ 表示所有需要计算的词。

2.2 GlobalPointer 模型

传统嵌套实体识别方法设计两个模块分别识别实体的头、尾位置，未考虑实体片段的内在关系，GlobalPointer 模型构造文本长度的方形矩阵，同时考虑首、尾位置，通过行和列索引位置来判断文本片段是否为一个实体，更具全局性，见图 2。第 1 行第 3 列属于病程类型的实体“5 年前”，赋予标签 1，其余部分为 0。此外，方形矩阵的数量与实体类别数量相同，每一个方形矩阵用来判别一种实体类别。命名实体识别任务方向为从前向后，如要判别“5 年前”是否为实体，无需考虑“前年 5”是否为实体的情况。基于此特性，矩阵左下三角为空白，无需赋予标签，训练时亦无需计算损失。图中每个小方框代表 1 个待识别的实体，对于长度为 n 的文本，若仅需要识别一种实体，则有 $n(n+1)/2$ 个不同的连续片段（待识别实体），因此，研究任务可转化为从中选择 a 个实体的多标签分类问题。

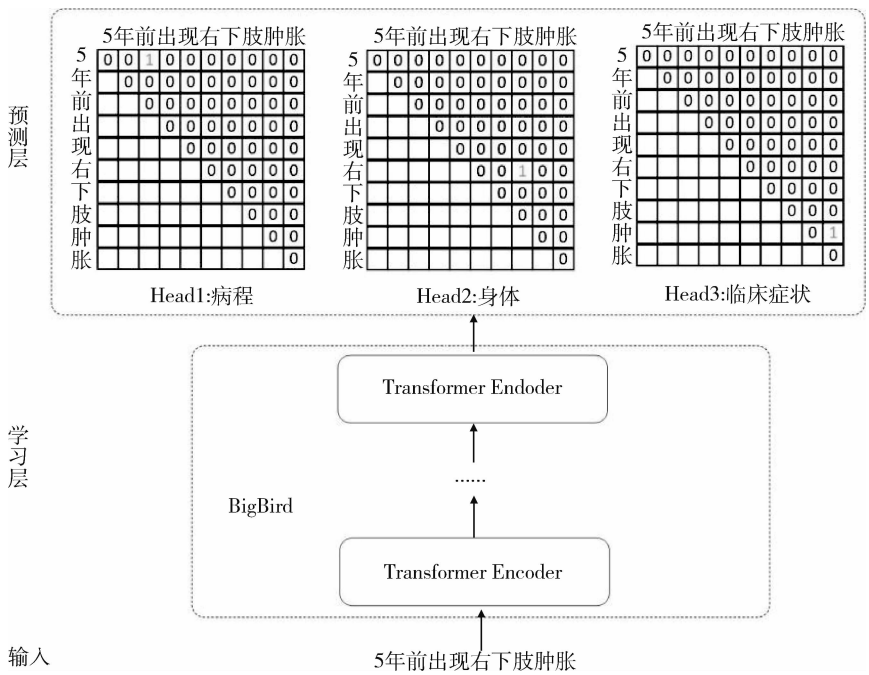

图 2 GlobalPointer 模型示例

GlobalPointer 模型由学习层和预测层两部分组成，学习层由 BigBird 编码器构成，输入文本 $X = [x_1, x_2, \cdots, x_n]$ 经过预训练语言模型 BigBird 编码得到语义表示 $H = [h_1, h_2, \cdots, h_n]$，其中：

$$h_1, h_2, \cdots, h_n = \mathrm{PLM}(x_1, x_2, \cdots, x_n) \quad (2)$$

令 $s[i:j]$ 表示文本的片段序列，i 表示开始位置索引，j 表示结束位置索引，H 经过前馈层变换后得到用于识别 α 类型实体的向量表示 $q_{i,\alpha}$（开始位置，矩阵中的行）和 $k_{j,\alpha}$（结束位置，矩阵中的列）：

$$q_{i,\alpha} = W_{q,\alpha} h_i + b_{q,\alpha} \quad (3)$$

$$k_{j,\alpha} = W_{k,\alpha} h_j + b_{k,\alpha} \quad (4)$$

GlobalPointer 模型利用稀疏的多头注意力机制将每一个头视为一种实体类型识别任务，从而实现多个实体类型的识别任务，将 q 和 k 的内积作为最后的打分（舍去注意力机制的值部分），$s_\alpha(i,j) = q_{i,\alpha}^T k_{j,\alpha}$ 表示文本第 i 个元素到第 j 个元素组成的连续片段属于 α 类型实体的得分。在此基础上，采用基于变换矩阵原理的旋转位置编码引入相对位置信息，用位置关系来限制实体长度，提升模型对实体长度的敏感性。例如，输入文本为“下肢核磁示右下肢继发性淋巴水肿”，对于识别“身体”类型的实体，真正实体为“下肢”“右下肢”，而模型的可能预测结果为“下肢核磁示右下肢”，引入相对位置信息后，有利于分辨出真正的实体：

$$R_i^T R_j = R_{j-i} \tag{5}$$

$$s_\alpha(i,j) = q_{i,\alpha}^T R_{j-i} k_{j,\alpha} \tag{6}$$

由于电子病历文本长度 n 较长，$n(n+1)/2$ 个待识别实体中包含的真正实体（标签为 1）数量往往占比较小，会带来极其严重的类别不均衡问题。采用多标签分类的损失函数解决此问题：

$$\text{Loss} = \log\left(1 + \sum_{(q,k)\in P_\alpha} e^{-s_\alpha(q,k)}\right) + \log\left(1 + \sum_{(q,k)\in Q_\alpha} e^{s_\alpha(q,k)}\right) \tag{7}$$

其中，P_α 表示 α 类型实体的首、尾集合，Q_α 表示非实体或者非 α 类型实体的首、尾集合，因此，损失函数的优化方向为属于 α 实体的 $s_\alpha(q,k)$ 得分增大，非 α 实体的 $s_\alpha(q,k)$ 得分减小。

3 实验设置

3.1 数据介绍

实验数据来自医院脱敏数据，见表 1。利用 Doccano 工具进行数据标注，临床医生确定的实体类别以及统计的实体数量，见表 2。实体数量分布不均衡，例如“临床症状”实体类别包含 29 342 个实体，而“微生物”实体类别只包含 2 个实体。共有 19 名淋巴外科专业的医生参与数据标注任务，其中主任医师 1 人，副主任医师 3 人，主治医师 5 人，住院医师 10 人。学历学位分布方面，12 人为博士学位，5 人为硕士学位，2 人为本科学位。数据标注流程为：先由高年资医生制定数据标注标准和质量控制规范，并标注 300 例示例数据；然后经过培训的低年资医生以“双人标注，双人核查”的方式标注剩余数据。对标注不一致的数据，由高年资医生进行最终决策，保证数据标注的准确性和规范性。

表 1 淋巴水肿电子病历文本数据统计

统计项	值
总数（例）	594
实体类别数量（个）	15
最小文本长度（字符）	1 574
最大文本长度（字符）	3 225
平均文本长度（字符）	2 100
男性	73
女性	521

续表 1

统计项	值
最大年龄（岁）	80
最小年龄（岁）	2
平均年龄（岁）	51

表 2 各类实体类别包含实体数量

实体类别	实体统计数（个）	不同实体名称数（个）
疾病诱因	9 328	532
病程	4 051	856
身体	27 089	1 256
临床症状	29 342	2 547
科室	1 975	194
医疗程序	5 335	1 046
频次	824	204
否认	26 220	112
疾病	23 047	1 383
医学检查项目	3 068	415
药物	488	205
疾病诊断分类	4 209	2 546
医疗设备	99	18
当前的	11	6
微生物	2	2
合计	135 088	11 322

3.2 实验参数设置

实验中模型的超参数包括训练批次（epoch）、学习率（learning rate）、文本最大长度（max_len）、批量大小（batch_size）。由于显存限制，文本最大长度设置为 2 800 字符，超出部分将截断，本研究数据截断占比为 1%。批量大小设置为 2，学习率一般为 e-5 级别，对常用的 2e-5、3e-5 和 5e-5 利用网格搜索法进行实验，结果表明学习率设定为 5e-5、训练批次设定为 25 时模型效果最优。

3.3 模型微调过程

借助预训练语言模型，GlobalPointer 模型已经在普通带嵌套命名实体识别任务中取得了最优效果，因此，本研究主要进行垂直领域微调学习，根据少量医生标注样本数据实现最终模型的快速学习。微调训练过程如下。输入：模型初始化参数

θ，学习率 λ。输出：更新后的参数。初始化模型的参数 θ，学习率为 λ。数据预处理：电子病历文本经过 BigBird 编码器后，得到语义向量 $H = [h_1, h_2, \cdots, h_n]$，作为新的输入 X。计算损失函数 $L_E(\theta) = -y\log p(y \mid x;\theta)$，$p(y \mid x)$ 表示预测标签为 y 的概率。则 $\theta = \theta - \lambda \nabla L_E(\theta)$。

3.4 基准模型选取

为验证本研究方法的适用性，选取 BERT - MRC 模型[12]进行比较。BERT - MRC 是一种基于机器阅读理解（machine reading comprehension，MRC）的命名实体识别模型，通过构建问句的方式引入实体类别相关先验信息，再与文本内容共同作为模型输入。随后，模型通过两个多分类任务从文本内容中抽取问句答案，分别预测答案的开始和结束位置，即实体在文本中的起止位置，从而完成命名实体识别任务。这种方法在多个中英文数据集的命名实体识别任务中表现优异，可作为基准模型与 GlobalPointer 模型进行预测效果的比较。

3.5 模型评估指标

采用 5 折交叉验证方法，将数据集分成 5 个子集，每次使用其中 4 个子集作为训练集，剩余的 1 个子集作为测试集，评价模型的预测能力。评估指标包括精准率（precision）、召回率（recall）和 Macro_$F1$ 分数，并计算均值和方差：

$$\text{precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{8}$$

$$\text{recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{9}$$

$$F1 = 2 \times \frac{\text{precision} \times \text{recall}}{\text{precision} + \text{recall}} \tag{10}$$

$$\text{Macro\_}F1 = \frac{1}{n} \sum_{i=1}^{n} F1^i \tag{11}$$

其中，TP 表示实际为正样本且预测为正样本的个数，FP 表示实际为负样本但预测为正样本的个数，TN 表示实际为负样本且预测为负样本的个数，FN 表示实际为正样本但预测为负样本的个数，精准率表示全部正样本的预测结果中正确预测所占比例，召回率表示全部正样本中正确预测所占比例，Macro_$F1$ 分数是精准率和召回率的调和平均值。此外，采用箱线图四分位数反映数据分布特征，并判断是否存在异常值。下限表示为 Q1 - 1.5（Q3 - Q1），下四分位数表示为 Q1，中四分位数表示为 Q2，上四分位数表示为 Q3，上限表示为 Q3 + 1.5（Q3 - Q1），异常值为低于下限或超过上限的值。

4 实验结果分析

4.1 准确性分析

数据集共 15 种实体类别，其中“微生物”“当前的”包含实体数量极少（分别为 2 个、11 个），予以剔除。因此，本研究评估含有 13 种实体类别的命名实体识别 GlobalPointer 模型效果，并与基准模型 BERT - MRC 进行对比。总体实验结果，见表 3，GlobalPointer 模型方差较小，并且没有异常值（“-”表示没有异常值），与 BERT - MRC 模型相比，在 Macro_$F1$ 分数方面可以提升约 8 个百分点，展现了实体识别总体结果最佳。

表 3 总体实验结果

评估指标	GlobalPointer			BERT - MRC		
	精准率	召回率	Macro_$F1$	精准率	召回率	Macro_$F1$
均值	0.795	0.641	0.697	0.617	0.621	0.619
方差	0.000 0	0.000 3	0.000 1	0.000 1	0.000 1	0.000 1
下限	0.785	0.579	0.667	0.589	0.593	0.603
下四分位数	0.792	0.623	0.688	0.612	0.615	0.616
中四分位数	0.795	0.647	0.702	0.615	0.619	0.617
上四分位数	0.797	0.652	0.702	0.627	0.630	0.625
上限	0.805	0.696	0.723	0.650	0.653	0.639
异常值	-	-	-	-	-	-

GlobalPointer 模型针对每个实体类别的分类效果，见表 4。“医疗设备”实体数量相较于其他实体类别过少，仅包含 99 个实体，虽然精准率高，但召回率过低，待样本数量增加后，Macro_*F*1 分数会有所提升。此外，“临床症状”“医疗程序”等实体类别包含不同名称的实体数量较多，对实验结果造成一定干扰，待数据标注规范更新统一后，命名实体识别模型效果可得到进一步提升。

表 4 GlobalPointer 模型 13 种实体类别实验结果

实体类型	均值			方差		
	精准率	召回率	Macro_*F*1	精准率	召回率	Macro_*F*1
疾病诱因	0.767	0.661	0.709	0.000 3	0.001 5	0.000 3
病程	0.833	0.805	0.819	0.000 5	0.000 0	0.000 1
身体	0.749	0.603	0.668	0.000 1	0.000 6	0.000 2
临床症状	0.706	0.513	0.592	0.001 1	0.001 4	0.000 5
科室	0.887	0.883	0.886	0.000 2	0.000 2	0.000 1
医疗程序	0.673	0.568	0.613	0.001 8	0.002 0	0.000 3
频次	0.613	0.538	0.614	0.000 3	0.002 6	0.002 1
否认	0.911	0.838	0.873	0.000 3	0.000 3	0.000 0
疾病	0.759	0.692	0.723	0.000 3	0.001 4	0.000 2
医学检查项目	0.755	0.670	0.710	0.001 0	0.000 2	0.000 2
药物	0.841	0.651	0.734	0.000 8	0.001 3	0.000 9
疾病诊断分类	0.722	0.607	0.659	0.000 9	0.000 6	0.000 7
医疗设备	1.000	0.318	0.456	0.000 0	0.005 8	0.006 4

4.2 案例分析

以某份淋巴水肿电子病历的命名实体识别结果为例进行分析，展示现病史、既往史、体格检查和出院诊断的标注情况，见图 3。左图为模型标注结果，右图为真实标签情况。针对“出院诊断”标注部分，模型不仅能够识别出“手术后淋巴水肿”这一“疾病诊断分类”类型的实体，同时可以识别出更细粒度的“疾病”类型实体“淋巴水肿”，模型具备识别出“XX 淋巴水肿”的能力，可以较好地解决实体嵌套问题，提升命名实体识别效果。但标注模型仍存在一定缺陷。例如，虽然将“宫颈癌根治术”正确识别为“疾病诱因”实体类别，却又赋予该实体“医疗程序”的错误标签，一定程度上说明针对某些实体，模型区分实体类型之间的差别能力较差。此外，模型存在一定的漏标（如未能识别“放疗”这一疾病诱因实体）问题，有待进一步提升。

图 3 模型标注部分效果示例

5 结语

本研究主要开展淋巴水肿疾病患者电子病历文本命名实体识别应用研究。利用医生专业领域知识确定了 13 种常见实体类别，涵盖疾病病史、症状、

诊断、治疗、评估等方面。基于少量医生标注的电子病历数据，针对电子病历文本数据实体嵌套特性，采用 GlobalPointer 模型，以及以自然语言理解大模型为基础的预训练－微调模型学习范式，实现领域快速自适应学习。实验结果表明 GlobalPointer 模型对淋巴水肿患者电子病历命名实体识别任务有效，这为真实临床病历数据构建和预处理奠定基础；数据和算法均填补了智能化方法在淋巴疾病领域的应用空白。本研究采用的数据来自电子病历中的非结构化文本内容，医学专业名词表达不统一、数据记录习惯不一致，产生了一词多义和多词一义的问题。因此，如果能经预处理实现命名实体归一化，排除噪声干扰，则可以进一步提升模型表现。

个性化精准医疗是疾病诊疗的必然需求，应在模型研发时融入更多领域知识，识别多种类型文档蕴含的重要实体。与此同时，实体之间的关系也影响关键实体识别，应结合知识图谱相关技术，深入挖掘实体和实体之间的多种关系，识别其与细分领域疾病的关联关系，共同构建数据基础。此外，从淋巴外科智能化诊疗技术发展远景来看，引入数据规范和模型应用验证的标准是推动技术研发和临床应用转化协同发展的必经之路。

利益声明：所有作者均声明不存在利益冲突。

参考文献

1 李文锋，朱威，王晓玲，等. Text2DT：面向临床诊疗文本的决策规则抽取技术［J］. 医学信息学杂志，2022，43（12）：16－22.

2 李甜，李晓东，刘敬禹. 人工智能辅助诊断肺结节的临床价值研究［J］. 中国全科医学，2020，23（7）：828－831，836.

3 杨锦锋，于秋滨，关毅，等. 电子病历命名实体识别和实体关系抽取研究综述［J］. 自动化学报，2014，40（8）：1537－1562.

4 DEVLIN J，CHANG M W，LEE K，et al. BERT：pre－training of deep bidirectional transformers for language understanding［EB/OL］.［2023－05－19］. https：//arxiv. org/abs/1810. 04805.

5 沈蓉蓉，夏帅帅，晏峻峰. 命名实体识别在中医药领域研究进展［J］. 医学信息学杂志，2023，44（1）：47－53.

6 WANG B，LU W. Neural segmental hypergraphs for overlapping mention recognition［EB/OL］.［2023－05－19］. https：//arxiv. org/abs/1810. 01817.

7 YU J，BOHNET B，POESIO M. Named entity recognition as dependency parsing［EB/OL］.［2023－05－19］. https：//arxiv. org/abs/2005. 07150.

8 JU M，MIWA M，ANANIADOU S. A neural layered model for nested named entity recognition［C］. New Orleans，Louisiana：The 2018 Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies，2018.

9 SOHRAB M G，MIWA M. Deep exhaustive model for nested named entity recognition［C］. Brussels：The 2018 Conference on Empirical Methods in Natural Language Processing，2018.

10 SU J，MURTADHA A，PAN S，et al. GlobalPointer：novel efficient span－based approach for named entity recognition［EB/OL］.［2023－05－19］. https：//arxiv. org/abs/2208. 03054.

11 ZAHEER M，GURUGANESH G，DUBEY K A，et al. BigBird：transformers for longer sequences［J］. Advances in neural information processing systems，2020，33（1）：17283－17297.

12 LI X，FENG J，MENG Y，et al. A unified MRC framework for named entity recognition［EB/OL］.［2023－05－19］. https：//arxiv. org/abs/1910. 11476.