

融合多级别粒度特征的中文医疗命名实体识别

代子男^{1,2}, 张捷^{1,2}, 陈冲冲^{1,2}, 谌章义^{1,2}, 陈付龙^{1,2}

(1. 安徽师范大学计算机与信息学院, 安徽 芜湖 241002; 2. 安徽省医疗大数据智能系统工程研究中心, 安徽 芜湖 241002)

摘要: 中文医疗命名实体识别(NER)旨在从医疗文本中识别具有特定意义的实体,如疾病、药物、症状及身体部位等多种类型的医疗实体。这一任务可为临床辅助决策、医疗信息整合和病案管理等方面提供有力支持。现有的中文医疗 NER 研究尚未充分考虑医疗文本的复杂结构,存在专业术语繁多、嵌入信息单一、语义信息利用不足等问题。为此,提出一种融合多级别粒度特征的中文医疗文本 NER 模型。该模型首先利用 BERT 预训练模型生成文本的字嵌入表示,并设计了一维和二维卷积神经网络(CNN)提取字符的字形与笔画特征,同时通过外部词库引入词级特征,以增强对词与实体边界的信息表达。此外,模型还加入句子级特征以捕获全局语义特征。通过交叉注意力机制将上述多级别的粒度特征进行迭代融合,得到包含深层语义信息的嵌入表示,最后利用条件随机场(CRF)输出实体识别结果。在 CCKS2017 和 CCKS2019 数据集上的实验结果表明,该模型 F1 值达到 92.88% 和 87.86%,相较于当前主流模型展现了更优异的识别性能。

关键词: 命名实体识别; 医疗文本; 特征融合; 特征增强; 交叉注意力

中图分类号: TP391

文献标志码: A

DOI: 10.19678/j.issn.1000-3428.0070675

Chinese Medical Named Entity Recognition Integrating Multi-Level Granular Features

DAI Zinan^{1,2}, ZHANG Jie^{1,2}, CHEN Chongchong^{1,2}, CHEN Zhangyi^{1,2}, CHEN Fulong^{1,2}

(1. School of Computer and Information, Anhui Normal University, Wuhu 241002, Anhui, China;

2. Anhui Engineering Research Center of Medical Big Data Intelligent System, Wuhu 241002, Anhui, China)

【Abstract】 Chinese medical Named Entity Recognition (NER) aims to identify entities with specific meanings from medical texts, such as diseases, drugs, symptoms, and anatomical parts. This task provides robust support for clinical decision making, medical information integration, and medical record management. However, existing research on Chinese medical NER has not fully addressed the complexity of medical texts, which are characterized by an abundance of specialized terminology, limited embedding diversity, and insufficient utilization of semantic information. To address these issues, this paper proposes a Chinese medical NER model that integrates multigranularity features. The model first employs a BERT pretrained model to generate character embeddings for the text. It then uses both one-dimensional and two-dimensional convolutional neural networks to extract character shape and stroke features, while external lexicons are incorporated to introduce word-level features, enhancing the representation of word and entity boundaries. Additionally, sentence-level features are included to capture global semantic information. A cross-attention mechanism is utilized to iteratively fuse these multigranular features, resulting in embeddings enriched with deep semantic information. Finally, Conditional Random Fields (CRF) are used to output the entity recognition results. Experimental results on the CCKS2017 and CCKS2019 datasets demonstrate that the proposed model achieves F1 values of 92.88% and 87.86%, respectively, outperforming mainstream models in recognition performance.

【Key words】 Named Entity Recognition(NER); medical text; feature fusion; feature enhancement; cross-attention

0 引言

随着互联网和医疗系统的迅速发展,医疗信息化不断加速推进。海量医学文本,如电子病历、医学

研究文献和药物说明书等,均以电子文本形式记录。然而,这些非结构化的医学文本存在大量专业术语和复杂语义结构,其信息可读性较差,且缺乏统一的结构化表达,难以分析和处理,可用性低^[1],导致数

基金项目: 安徽省科技创新攻坚计划(202423k09020050);安徽高校协同创新项目(GXXT-2022-054);芜湖市科技计划(2022yf54);芜湖市重点研发与成果转化项目(2023yf117)。

作者简介: 代子男,男,硕士研究生,主研方向为自然语言处理;张捷(通信作者),副教授、博士;陈冲冲,硕士研究生;谌章义,副教授、博士;陈付龙,教授、博士。

收稿日期: 2025-02-27

修回日期: 2025-04-07

E-mail: zjzj2526@163.com

据碎片化和信息孤岛问题日益严重,从而极大地制约了其在临床决策、科研支持和智能医疗中的应用。医生和研究人员在提取关键诊疗信息时往往面临传统信息抽取技术难以应对医学文本高度专业化、多层次语义结构的挑战,使得这些技术的潜力未能得到充分发挥。因此,如何准确高效地利用其中的有效信息至关重要。

命名实体识别(NER)作为信息抽取的核心技术,在医疗领域的应用尤为关键。以电子病历为例,对其进行 NER 可以有效识别电子病历中涵盖的疾病、症状、体征、诊疗手段及治疗效果等关键信息。这些关键信息不仅能显著节省医疗资源,还能提升医疗服务质量,并推动智能诊疗、医疗知识图谱构建和病程追踪等领域的快速发展。然而,与通用领域相比,中文医疗命名实体识别^[2]面临更复杂的挑战。医学文本包含大量专业术语,通用方法难以直接适用。现有方法多集中于单一粒度特征,如字符级或词级,难以全面捕捉局部和全局信息之间的关联关系。在处理复杂医疗文本时,缺乏对多种嵌入特征的有效融合与深层交互机制。

为解决上述问题,本文提出一种多粒度特征融合模型 MGFCN。该模型通过整合字符级、词级和句子级特征,结合深层特征交互机制,在全面捕捉文本特征的基础上,有效提升中文医疗命名实体识别的性能。该方法为缓解信息孤岛问题和促进智能医疗应用提供一种切实可行的技术方案和理论依据。

1 相关工作

目前,NER 常用方法主要分为 3 类^[3]:基于规则与词典的方法、基于机器学习的方法和基于深度学习的方法。此外,在中文医疗领域,也有针对命名实体识别的专门方法。

1.1 基于规则与词典的方法

基于规则和词典的方法通过人工编写专业词典和规则来识别命名实体^[4]。这种方法需要先收集整理相关专有名词,并建立词典作为匹配模板,通过实体构造规则提取相应实体。该方法实施难度较低,但识别效果依赖于词典的完备性和规则的可行性,同时受到各地自然语言处理技术发展不均衡的限制,使得词典构建面临较大困难。此外,随着专业术语不断更新和扩展,人工维护词典的成本较高,限制了该方法的适用性。

1.2 基于机器学习的方法

基于机器学习的方法通过标注数据集提取词

性、上下文语义等特征,并将这些特征输入条件随机场(CRF)^[5]、支持向量机^[6](SVM)和隐马尔可夫^[7](HMM)等监督学习模型进行训练,使模型能够自动识别文本中的实体。相比基于规则的方法,机器学习方法减少了人工构造规则的工作量,能够在一定程度上提升识别的泛化能力。然而,该方法高度依赖于手动标注的数据和特征工程,不仅需要大量高质量的标注数据,还需要人工设计特征,这一过程极为耗时,并在一定程度上限制了模型的性能表现。此外,由于不同专业领域文本包含复杂的语境信息和丰富的领域特征,传统的机器学习模型在深层语义信息的捕捉方面存在一定局限性。

1.3 基于深度学习的方法

近年来,随着深度学习的快速发展,基于神经网络的方法逐渐成为命名实体识别任务的主流。长短期记忆(LSTM)网络是最先广泛用于命名实体识别任务中的深度学习模型之一^[8-9],LSTM 模型可以处理长距离的依赖信息,但是单向结构无法对反向文本数据进行处理。因此,SCHUSTER 等^[10]在 1997 年提出了双向长短时记忆(BiLSTM)网络。此后,XU 等^[11]将 BiLSTM 与 CRF^[12]结合在通用领域取得显著的效果。研究人员在此基础上加入迁移学习,在人民日报数据集上取得了 91.53% 的准确率。为了提升特征捕捉能力,YIN 等^[13]提出基于字符偏旁等级和自注意力(Self-attention)机制的 BiLSTM+CRF 模型,并使用卷积神经网络(CNN)来提取字符的偏旁特征。ZHANG 等^[14]提出的 Lattice-LSTM 模型,在处理字符序列时引入词汇表来捕捉字符序列中潜在词汇信息,并设计门控单元以选择最佳匹配的词汇。MA 等^[15]借鉴 Lattice-LSTM 模型,利用词典匹配字符得到 B、M、E、S 标注的词汇信息,并将其融入对应的字信息中,以增强模型的识别能力。

随着大规模预训练语言模型的兴起,DEVILIN 等^[16]提出一种基于 Transformer 的双向编码表示(BERT)的语言预处理模型,显著增强了上下文语义的建模能力。LIU 等^[17]提出中文序列标注的词典增强 BERT 的方法,通过词典适配器将外部词汇直接集成到 BERT 层中,提升了中文序列标注任务的表现。

1.4 中文医疗命名实体识别方法

在中文医疗命名实体识别任务中,研究人员结合领域特性提出了多种改进方法。罗凌等^[18]利用字符的笔画特征进行医疗实体识别任务。宋雅文等^[19]使用多窗口大小的卷积神经网络提取字符级

别的特征。但是中文医疗命名实体识别基于字符级粒度特征嵌入的方法缺少词级别信息,LI 等^[20]使用 Lattice-LSTM 模型融入词汇信息,识别中文医疗实体。WANG 等^[21]针对电子病历中的命名实体识别任务,对 BERT 预训练模型进行了改进,以解决缺少上下文信息的问题,并添加了一种额外的机制来捕获字符之间的关系。

2 融合多级别粒度特征的医疗实体识别模型

本文设计的模型主要由多级别粒度特征嵌入层、特征融合层、特征解码层 3 个部分构成,总体结构如图 1 所示。

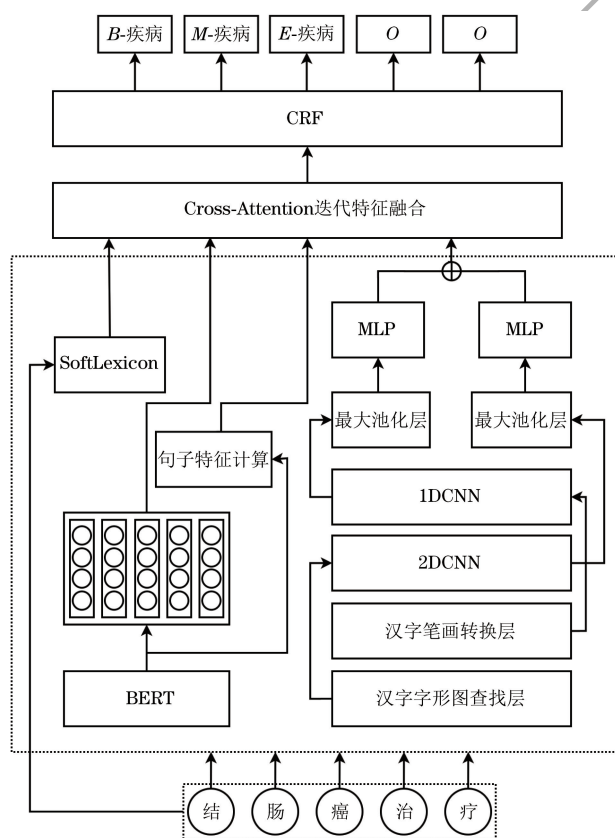


图 1 MGFCN 模型框架

Fig.1 MGFCN model framework

在多级粒度特征嵌入层中,采用基于大规模语料库训练语言模型 BERT 生成动态上下文字符嵌入,同时使用多层卷积模块提取字符的字形和笔画串特征嵌入。同时,通过医学词库对输入序列中的每个字符进行匹配,引入词汇的先验知识,以获得词特征嵌入。模型还利用 BERT 的开始标志位 [CLS] 位置的输出向量作为句子特征嵌入。在特征融合层中,通过迭代交叉注意力机制有效地融合上述多级粒度特征,从而生成最终的词嵌入表示。特征解码层通过 CRF 预测输入句子的全局最优标注序列。

2.1 多级别粒度特征嵌入层

2.1.1 字向量嵌入

BERT 作为一种预训练的双向 Transformer 模型,它通过对句子进行双向编码,区别于传统的词向量模型如 Word2Vec^[22]、GloVe^[23]。传统模型只能为每个词生成固定的向量表示,难以有效处理词义的上下文差异。然而 BERT 能捕捉到词汇本身的语义,并根据上下文动态调整每个字或词的向量表示,展现了 BERT 强大的字向量特征提取能力。在处理输入文本序列时,首先将输入的文本序列处理成单个字符形式,在每个句子的头和尾分别加入 [CLS] 和 [SEP],其中 [CLS] 为每个句子的开始标志位, [SEP] 为结束标志位,可用于分隔 2 个句子。给定长度为 n 的输入句子 $C=[c_1, c_2, \dots, c_n]$, 其中 $c_i \in C, i=1, 2, \dots, n$ 。第 i 个字符输入 BERT 模型,得到字向量 $x_{c_i} = \text{BERT}(C_i)$ 。字向量嵌入获取流程如图 2 所示。

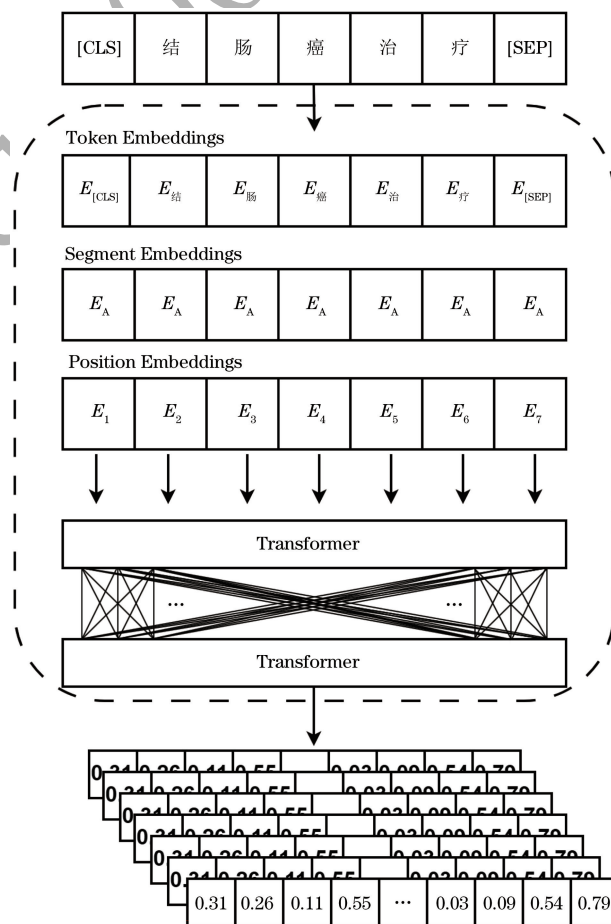


图 2 字向量嵌入提取

Fig.2 Character embedding extraction

2.1.2 字形笔画串特征嵌入

中文医疗文本不同于中文通用领域数据,前者具有较强的专业性,包含大量专业术语。为了全面融合中文医疗文本中的丰富特征信息,需要从多个

角度挖掘医疗数据集所包含的信息。汉字作为一种象形文字,所蕴含的信息主要是图像信息而非简单的编码信息。其中,字形和笔画有着丰富的语义特征。在中文医疗文本中,存在着大量相似字形的汉字,如“腹”“胸”“腿”“癌”“症”等。因此,从汉字的整体结构出发,利用卷积神经网络提取汉字的整体字形特征是至关重要的。在获得汉字的字形特征后,还可以考虑从汉字局部构成入手,提取汉字笔画串特征。将汉字的整体结构和局部构成特征相互补充,融合字形和笔画串特征,将有助于获取汉字的深层语义信息。

在字形特征提取过程中,先将字符统一转换为尺寸 32×32 的思源黑体灰度图。接着,利用多层卷积模块对灰度图进行处理,该模块包含 2 次二维卷积操作、ReLU 激活函数以及最终的池化层。提取到的特征图经过展平操作后,通过全连接层获得目标尺寸的嵌入表示,最后重新划分以建立句子界限。字形特征提取流程如图 3 所示。

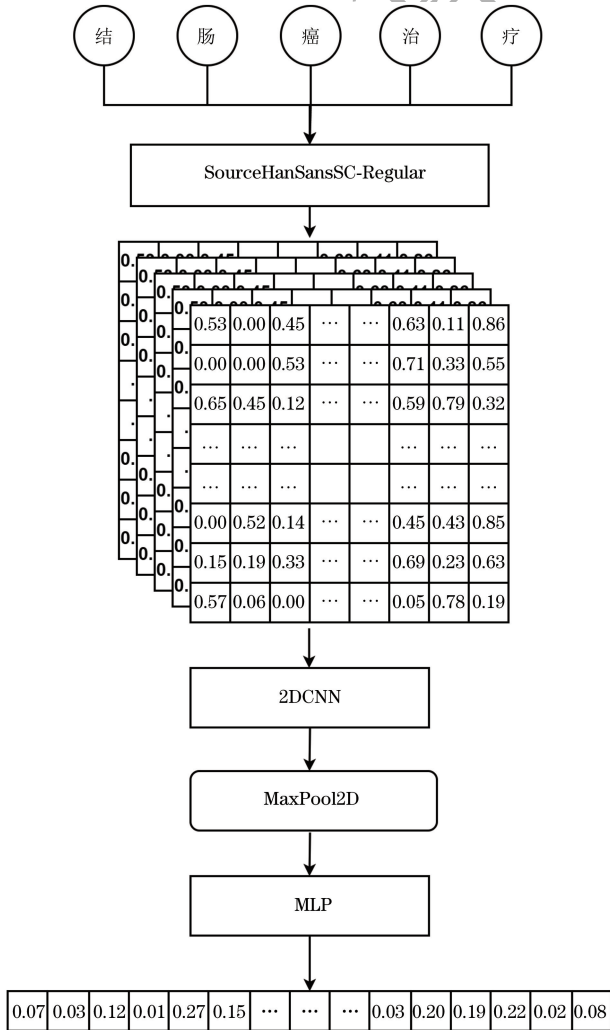


图 3 字形特征提取

Fig.3 Glyph feature extraction

笔画串特征提取与字形特征提取流程相似。首先,将字符转换为笔画串,并将序列截断或补充到固定的长度。接着,笔画串通过多层卷积模块、展平操作和全连接层,生成目标尺寸的嵌入表示,最后重新划分以建立句子界限。不同之处在于,该多层卷积模块采用一维卷积。一维卷积神经网络是一种专门用于处理一维序列数据的神经网络结构。与常用于图像数据的二维卷积神经网络相比,一维卷积在处理序列数据和局部模式识别任务时表现出更高的效率与准确性。笔画串特征提取流程如图 4 所示。

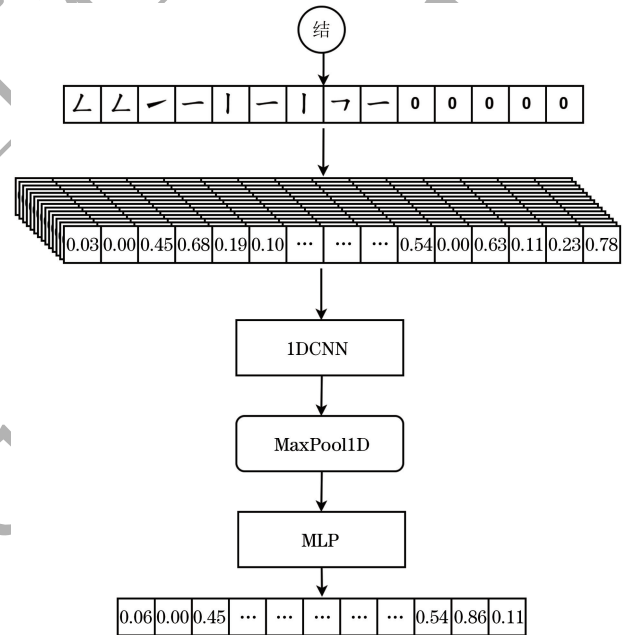


图 4 笔画串特征提取

Fig.4 Stroke string feature extraction

最后,将字形特征向量与笔画串特征向量拼接合并,得到字形笔画串特征嵌入。字形笔画串特征嵌入表示如式(1)所示:

$$\text{gsemb}_{c_i} = \text{gemb}(c_i) \oplus \text{semb}(c_i) \quad (1)$$

式中: c_i 为第 i 个字符; $\text{gemb}(\cdot)$ 为字形特征提取; $\text{semb}(\cdot)$ 为笔画串特征提取; $\oplus$ 为串联操作。

2.1.3 词特征嵌入

基于字符级嵌入的模型能够避免中文分词带来的误差,但是同样会缺乏词和实体的边界信息。然而,词级粒度特征相较于字符级粒度特征更能保留文本的词汇信息和实体边界信息。因此,本文模型在特征嵌入层引入词特征嵌入,以增强模型对医疗文本语义信息的利用能力,从而进一步提升模型对医疗文本语义的捕捉和表达。

词特征的引入将通过 SoftLexicon 方式进行。首先构建用于字符匹配的医学词典,以搜狗医学词库为基础,并收集大量医学领域的专业词汇用以扩充,构

成词典 L 。其中,搜狗医学词库共计 90 047 个词条,扩充词库具体的类别和词条数量如表 1 所示。

表 1 实体种类数量

Table 1 Number of entity types	单位:个
实体种类	实体数量
临床检查	2 239
医疗器械	1 403
手术操作	11 035
疾病诊断	30 608
症状体征	1 403
药物名称	438
解剖部位	1 206

将字符与医学词典匹配的词分为 B 、 M 、 E 、 S 共 4 个集合,集合 B 是包含以当前字符为起始的词汇,集合 M 是包含以当前字符在词汇中间的词汇,集合 E 是包含以当前字符为结尾的词汇,集合 S 是包含以当前字符独立构成的词汇。字符 c_i 对应集合的构造方式如式(2)所示:

$$\begin{aligned} B(c_i) &= \{w_{i,k}, \forall w_{i,k} \in L, i < k \leq n\} \\ M(c_i) &= \{w_{j,k}, \forall w_{j,k} \in L, 1 \leq j < i < k \leq n\} \\ E(c_i) &= \{w_{j,i}, \forall w_{j,i} \in L, 1 \leq j < i\} \\ S(c_i) &= \{c_i, \exists c_i \in L\} \end{aligned} \quad (2)$$

式中: $w_{i,k}$ 为字符 c_i 起始和字符 c_k 结尾的词; $w_{j,k}$ 为字符 c_j 起始和字符 c_k 结尾的词; $w_{j,i}$ 为字符 c_j 起始和字符 c_i 结尾的词。

如图 5 所示,以“治疗心脏病”中字符“心”的医学词典 L 匹配结果为例。如果某个分类集合没有匹配到词汇,将用 None 表示。

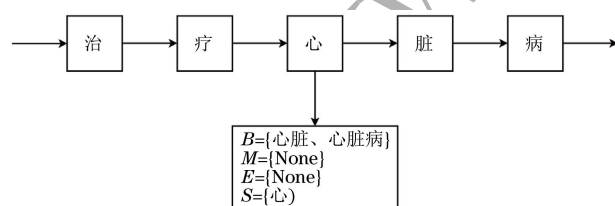


图 5 词典匹配实例

Fig.5 Lexicon matching instances

图 5 中,还存在分类集合中匹配到多个词汇的情况。对此,统计训练集和测试集中各个词汇的频率,并直接使用词频作为权重,将每类集合映射为固定长度的词嵌入。如词汇集合 B 的向量表示 $v(B)$ 如式(3)所示:

$$v(B) = \frac{1}{Z} \sum_{w \in B} z(w) e^w(w) \quad (3)$$

式中: $z(w)$ 表示词频; Z 表示 B 、 M 、 E 、 S 这 4 个集合的词频总数; $e^w(w)$ 表示词汇 w 的词向量。医学词典 L 使用 Word2Vec 训练词向量得到词向量

词典。

与式(3)相似,计算得到其他 3 个词汇集合的向量表示,将 4 个集合的向量表示进行串联拼接,得到对应字符的词特征嵌入,如式(4)所示:

$$v(B, M, E, S) = [v(B) \oplus v(M) \oplus v(E) \oplus v(S)] \quad (4)$$

式中: $v(M)$ 为词汇集合 M 的向量表示; $v(E)$ 为词汇集合 E 的向量表示; $v(S)$ 为词汇集合 S 的向量表示; $\oplus$ 为串联操作。

2.1.4 句子特征嵌入

虽然句子嵌入主要用于捕捉句子的整体语义信息,多应用于文本分类、情感分析等任务,但在命名实体识别任务中,句子嵌入能够提供全局上下文信息,有助于模型更好地理解句子中的实体关系,弥补细粒度嵌入在捕捉远距离上下文方面的不足。当句子结构较为复杂,尤其是在长句子和复杂句子中实体边界模糊的情况下,单靠细粒度嵌入难以精确识别实体边界。这时,通过融合句子的全局嵌入,模型能够更全面地理解句子整体语义,从而提供有效的辅助信息并做出更准确的判断。

句子嵌入的获取方法将使用 BERT 的开始标志位[CLS]位置的输出向量作为句子特征嵌入。BERT 在处理文本时,首先在输入序列的最前面插入[CLS],并通过双向 Transformer 对整个序列进行编码。在 BERT 的输出中,[CLS]位置对应的向量并不是单一字的向量,而是整个句子的全局向量表示。这个向量包含了上下文信息,并能够捕捉到输入句子的综合语义。例如,对于输入句子 $C = [c_1, c_2, \dots, c_n]$,其句子 C 的句子特征嵌入如式(5)所示:

$$\begin{aligned} H &= \text{BERT}(C_X) \\ h_{\text{cls}} &= H[0] \\ C_X &= ['[CLS]', c_1, c_2, \dots, c_n, '[SEP]'] \end{aligned} \quad (5)$$

式中: C_X 为加入开始标志位和结束标志位的句子 C ; h_{cls} 为句子 C 的句子特征嵌入向量。

2.2 特征融合层

在特征融合方面,现有研究多采用直接拼接或基于注意力机制的拼接方式。直接拼接的方式是假设不同粒度特征之间是相互独立的,信息没有交互,因此难以充分挖掘各粒度特征中的深层信息。而基于注意力机制的拼接方式通常依赖单向注意力机制,难以有效融合多种特征,限制了融合效果的提升。

在图 1 所示框架中,使用基于交叉注意力机制(Cross-Attention)进行多特征融合。交叉注意力可

在不同序列之间计算注意力,将不同的信息进行关联。优势在于建立不同特征之间的相关性,便于学习到各特征之间的深层关系。本文将多种粒度特征通过交叉注意力机制进行多次迭代融合,以充分利用不同粒度之间的信息。所采用交叉注意力机制结构如图 6 所示。

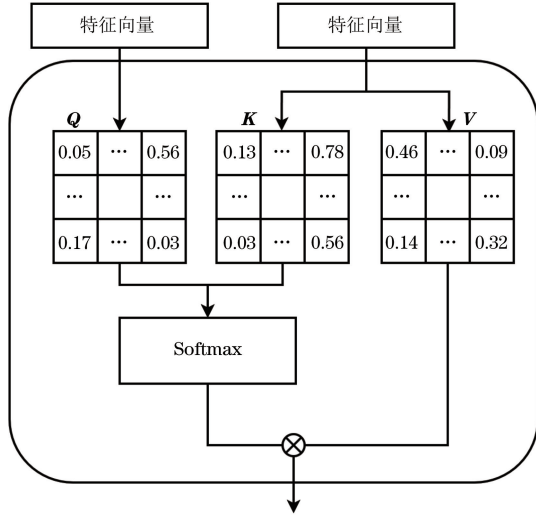


图 6 交叉注意力结构

Fig.6 Cross-attention structure

Q 为查询矩阵, K 为键向量矩阵, V 为值向量矩阵。通过计算 Q 和 K 的相似度,再通过 Softmax 函数得到 Q 对 K 的注意力分布,最后将注意力分布加权用于 V ,如式(6)所示:

$$\text{Cross-Attention}(Q, K, V) = \text{Softmax} \left(\frac{QK^T}{\sqrt{d_k}} \right) \quad (6)$$

式中: d_k 为 K 的维度,用于缩放内积结果。

具体来说,特征融合过程需要经过 3 次迭代交叉注意力机制完成。首先,将字向量特征与字形笔画串特征进行交互融合,此时字向量特征作为查询矩阵 Q ,字形笔画串特征作为键向量矩阵 K 和值向量矩阵 V ;接着,将词特征加入融合,使用第 1 次交互融合的输出向量作为查询矩阵 Q ,词特征作为键向量矩阵 K 和值向量矩阵 V ;最后,融合句子特征,将第 2 次融合的输出向量作为查询矩阵 Q ,句子特征作为键向量矩阵 K 和值向量矩阵 V ,通过逐步迭代融合的方式,将不同粒度的特征层层交互,实现深层次的特征融合。融合的具体流程如图 7 所示。

2.3 特征解码层

在经过特征提取和特征融合后,本文使用 CRF 对序列标注任务进行建模。CRF 能够建立相邻标签间的约束关系,可以避免出现标签偏置问题。例

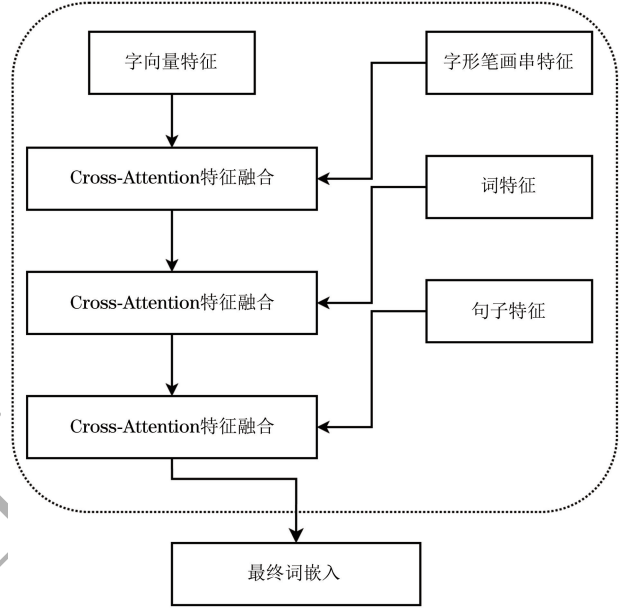


图 7 迭代特征融合层

Fig.7 Iterative feature fusion layer

如,标签“B-疾病”之后不可能直接出现“E-手术”的情况,这种关系通过转移矩阵得到有效约束。将特征融合后的输出 $X = [x_1, x_2, \dots, x_n]$ 作为 CRF 的输入,模型预测标签序列为 $y = [y_1, y_2, \dots, y_n]$ 。对应标签序列得分如式(7)所示:

$$S'(X, y) = \sum_{i=0}^n A_{y_i, y_{i+1}} + \sum_{i=1}^n P_{i, y_i} \quad (7)$$

式中: $A_{y_i, y_{i+1}}$ 为标签 y_i 转移到标签 y_{i+1} 的得分; P_{i, y_i} 为第 i 个序列被标注为 y_i 的得分。为了将得分函数转换为概率分布,使用 Softmax 函数对所有可能的标签序列进行归一化,得到标签序列的条件概率如式(8)所示:

$$P(y|X) = \frac{e^{S'(X, y)}}{\sum_{\bar{y} \in \mathcal{Y}} e^{S'(X, \bar{y})}} \quad (8)$$

式中: $\bar{y}$ 为真实标签。为了优化模型参数,本文采用负对数似然损失函数,损失函数如式(9)所示:

$$l = -\log_a P(y|X) = -\log_a \left(\sum_{\bar{y} \in \mathcal{Y}} e^{S'(X, \bar{y})} \right) - S'(X, y) \quad (9)$$

该损失函数的优化目标是最大化真实标签序列的条件概率,通过最小化 l ,从而提升模型对标签序列的预测能力。为了最小化损失函数 l ,本文采用梯度下降优化参数,其梯度计算如式(10)所示:

$$\frac{\partial l}{\partial \theta} = \frac{\partial \log_a \left(\sum_{\bar{y} \in \mathcal{Y}} e^{S'(X, \bar{y})} \right)}{\partial \theta} - \frac{\partial S'(X, y)}{\partial \theta} = \sum_{\bar{y} \in \mathcal{Y}} P(\bar{y}|X) \frac{\partial S'(X, \bar{y})}{\partial \theta} - \frac{\partial S'(X, y)}{\partial \theta} \quad (10)$$

式中: θ 为模型训练中需要优化所有参数的集合, 包括标签转移矩阵 $\mathbf{A}$ 以及特征得分的权重和偏置项等。通过前向-后向算法计算模型参数的更新梯度, 模型参数 θ 更新如式(11)所示:

$$\theta = \theta - \eta \cdot \frac{\partial l}{\partial \theta} \quad (11)$$

式中: η 为模型的学习率。在模型训练过程中, 通过上述优化, 逐步提升模型对输入序列的条件概率建模能力, 获得最大的条件概率, 最后使用式(12)的维特比算法计算出最优标签序列。

$$y^* = \operatorname{argmax}(S'(\mathbf{X}, \bar{y})) \quad (12)$$

3 实验及结果分析

3.1 实验数据

本文通过在 CCKS2017 和 CCKS2019 数据集上的实验验证所提方法的有效性。CCKS2017 和 CCKS2019 是全国知识图谱与语义计算大会在 2017 年和 2019 年的中文医疗命名实体识别评测任务的 2 个中文医疗实体数据集。其中 CCKS2017 包含了 2 229 条数据, CCKS2019 包含了 1 379 条数据, 将 2 个数据集以 8:2 的比例随机分配为训练集和测试集, 2 个数据集包括多类医疗实体类型, 详细的各类实体类型如表 2 和表 3 所示。

表 2 在 CCKS2017 数据集上实体类型统计

Table 2 Entity type statistics on CCKS2017 dataset 单位: 个

CCKS2017 数据集	身体部位	检查	症状	治疗	疾病
训练集	13 744	13 082	9 588	4 317	2 184
测试集	5 808	6 323	4 706	792	2 519

表 3 CCKS2019 数据集实体类型统计

Table 3 Entity type statistics on CCKS2019 dataset 单位: 个

CCKS2019 数据集	解剖部位	疾病和诊断	药物	实验室检验	影像检查	手术
训练集	9 118	4 217	1 795	1 337	906	905
测试集	2 402	1 318	512	448	411	286

3.2 评价指标

本文采用命名实体识别中常用的评价指标, 分别为精确率(Precision, P)、召回率(Recall, R)和 F1 值(F_1, F_1), 对本文模型和对比模型进行评价, 计算如式(13)~式(15)所示:

$$P = \frac{N_{TP}}{N_{TP} + N_{FP}} \quad (13)$$

$$R = \frac{N_{TP}}{N_{TP} + N_{FN}} \quad (14)$$

$$F_1 = \frac{2P \times R}{P + R} \quad (15)$$

式中: N_{TP} 为模型识别正确的实体标签数量; N_{FP} 为模型识别错误的实体标签数量; N_{FN} 为模型未识别的实体标签数量。

3.3 对比实验

为验证所提方法的有效性, 在 2 个不同数据集上与多种当前主流模型进行比较, 其中表 4 为在 CCKS2017 数据集上不同模型的对比结果, 表 5 为在 CCKS2019 数据集上不同模型的对比结果, 其中“—”表示该方法的原论文未提供该项指标。

表 4 在 CCKS2017 数据集上不同模型的实验结果对比

Table 4 Comparison of experimental results among different models on the CCKS2017 dataset %

模型	Precision	Recall	F1
BERT-BiLSTM-CRF	90.86	87.11	88.95
MIN	91.95	87.93	89.90
CMNER	91.44	92.30	91.87
MBC	92.11	91.84	91.97
LIBERT-GAT	90.43	93.83	92.10
RoBERTa-wwm-MLDF-crf	92.60	91.90	92.24
MGFCN	93.47	92.30	92.88

表 5 在 CCKS2019 数据集上不同模型的实验结果对比

Table 5 Comparison of experimental results among different models on the CCKS2019 dataset %

模型	Precision	Recall	F1
FLAT	85.39	83.08	84.20
FBBC	—	—	85.62
MANM	—	—	86.12
MSD_DT_NER	86.09	87.29	86.69
MECT	87.43	86.21	86.82
文献[34]模型	86.70	87.80	87.25
MGFCN	89.04	86.71	87.86

BERT-BiLSTM-CRF: 将 BERT-BiLSTM 作为底层的文本特征编码器, 利用 CRF 进行实体标签预测, 是目前最流行的方法之一。

MIN^[24]: 利用句段级信息和词级依赖关系, 并结合交互机制来支持边界检测和类型预测之间的信息共享。

CMNER^[25]: 一种用于中文医学命名实体识别的多级语义融合网络, 该网络融合了形态、字符、单词和句法层面的语义信息。

MBC^[26]: 一种改进的字符级表示方法, 该方法融合字符嵌入和字符标签嵌入以增强特征表示的特殊性和多样性。

LIBERT-GAT^[27]: 利用词汇特征和词边界信

息,减少分词错误,同时通过图注意力网络提取文本的依存关系特征,优化语义依赖信息以区分实体边界。

RoBERTa-www-MLDF-crf^[28]:通过动态融合预训练模型各层隐向量,利用底层表层特征、中层句法特征和高层语义特征,提高实体识别的准确性和鲁棒性。

FLAT^[29]:凭借 Transformer 位置编码,结合词汇信息,以提升识别的性能和效率。

FBBC^[30]:在修正的数据集上使用 BERT 预训练模型提取字向量,并使用 BiLSTM-CRF 进行编解码,曾取得 CCKS2019 竞赛第一名的成绩。

MANM^[31]:一种结合字符词级特征和医疗领域专属注意力的模型,通过强化长距离依赖和专业术语识别能力。

MSD_DT_NER^[32]:通过多模态树提取语义词、边界信息并进行多粒度特征融合。

MECT^[33]:一种新颖的基于多元数据嵌入的跨 Transformer 模型,通过融合汉字的结构信息来提高中文 NER 的性能。

文献[34]模型:一种基于词性和汉语拼音特征的实体识别方法,采用缩放点积注意力模块将几种特征进行融合。

在 CCKS2017 数据集上,本文设计的 MGFCN 模型 F1 值达到了 92.88%,与 BERT-BiLSTM-CRF、MIN、CMNER、MBC、LIBERT-GAT、

RoBERTa-www-MLDF-crf 相比 F1 值分别提升了 3.93、2.98、1.01、0.91、0.78、0.64 百分点;在 CCKS2019 数据集上,MGFCN 模型 F1 值达到了 87.86%,与 FLAT、FBBC、MANM、MSD_DT_NER、MECT、文献[34]模型相比 F1 值分别提升了 3.66、2.24、1.74、1.17、1.04、0.61 百分点。

综上所述,通过与多种当前先进模型对比,本文提出的模型在中文医疗命名实体识别任务中表现出显著优势。该模型基于 BERT 提取字向量,结合医学文本的字形结构特性,引入字形特征以增强对复杂医学术语的辨识能力;同时加入词特征信息,以提高对实体边界的识别精度,使模型更准确地确定实体起止位置。模型还融入全局句子特征,从整体层面捕捉语义关系,避免局部信息孤立,并采用迭代交叉注意力机制实现字、词和全局特征的多级融合。实验结果显示,本文模型在多个指标上优于现有模型,尤其在识别较长实体或含多义词的医学术语方面表现突出,表明多种嵌入特征的交互融合能有效提升模型对文本深层语义的理解和命名实体识别效果。

3.4 消融实验

为了进一步验证所提模型各模块的有效性,本文设计了 5 种消融模型,对字形笔画串特征提取模块、词特征提取模块、句子特征提取模块和迭代特征融合模块进行了分析验证。在 CCKS2017 和 CCKS2019 数据集上分别进行消融实验,以评估各模块的贡献和效果。消融实验结果如表 6 所示。

表 6 消融实验结果

Table 6 Ablation experimental results

模型	框架	CCKS2017			CCKS2019			%
		Precision	Recall	F1	Precision	Recall	F1	
Model1	BERT+CRF	90.77	87.03	88.86	84.22	83.72	83.97	
Model2	Model1+字形笔画串特征	91.38	88.45	89.89	84.97	84.73	84.85	
Model3	Model2+词典特征	92.21	90.75	91.47	87.10	85.59	86.34	
Model4	Model3+句子特征	92.90	91.19	92.04	87.85	85.81	86.82	
Model5	Model4+迭代交叉注意力特征机制	93.47	92.30	92.88	89.04	86.71	87.86	

从表 6 的消融实验结果可以看出,Model1 仅使用 BERT 预训练语言模型的方法,在数据量较小、语境复杂、专业术语较多的医疗数据集上获取文本信息不足,识别效果较差。在 Model1 模型基础上加入字形笔画串特征后,模型分别在 CCKS2017 和 CCKS2019 数据集上 F1 值分别提高了 1.03 和 0.88 百分点,说明加入汉字的字形笔画串特征可以丰富特征嵌入的语义信息。在 Model2 模型基础上加上词典特征后,模型分别在 CCKS2017 和 CCKS2019 数据集上 F1 值分别提高了 1.58 和

1.49 百分点,表明加入词典特征有助于保留文本的词汇信息并增强实体边界信息。在 Model3 模型基础上加上句子特征后,模型分别在 CCKS2017 和 CCKS2019 数据集上 F1 值分别提高了 0.57 和 0.48 百分点,可以表明加入句子特征使得模型获得句子的整体语义信息,缓解细粒度嵌入难以捕捉远距离上下文的问题。最后,在 Model4 模型基础上加上迭代交叉注意力特征融合机制后,模型分别在 CCKS2017 和 CCKS2019 数据集上 F1 值分别提高了 0.84 和 1.04 百分点,说明迭代交叉注意力特征

机制的引入使得多级别粒度特征之间进行交互,通过学习各特征之间的深层关系进一步提升模型的识别性能。

4 结束语

本文提出一种融合多级别粒度的中文医疗命名实体识别模型,在特征提取阶段引入字符的字形特征和笔画串特征、词典信息、句子的整体语义多个不同粒度级的特征,以挖掘文本深层的语义信息;然后将多个粒度级特征通过迭代交叉注意力特征机制融合,最后通过 CRF 解码得到最优的标签序列。实验结果表明,本文模型在 CCKS2017 和 CCKS2019 数据集上分别获得了 92.88% 和 87.86% 的 F1 值,通过消融实验验证了所提模型的有效性。下一步将尝试挖掘中文医疗文本中其他有价值的特征,并探索更优的特征融合方法。

参考文献

- [1] KE J, WANG W J, CHEN X J, et al. Medical entity recognition and knowledge map relationship analysis of Chinese EMRs based on improved BiLSTM-CRF [J]. Computers and Electrical Engineering, 2023, 108: 108709.
- [2] YI F, LIU H, WANG Y, et al. Medical named entity recognition fusing part-of-speech and stroke features [J]. Applied Sciences, 2023, 13(15): 8913.
- [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.
- [4] ZHAO J G, QIAN Y R, WANG K, et al. Survey of Chinese named entity recognition research [J]. Computer Engineering and Applications, 2024, 60(1): 15-27. (in Chinese)
- [5] 崔博文, 金涛, 王建民. 自由文本电子病历信息抽取综述 [J]. 计算机应用, 2021, 41(4): 1055-1063.
- [6] CUI B W, JIN T, WANG J M. Overview of information extraction of free-text electronic medical records [J]. Journal of Computer Applications, 2021, 41(4): 1055-1063. (in Chinese)
- [7] LAFFERTY J D, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data [C] // Proceedings of the 18th International Conference on Machine Learning. San Francisco, USA: Morgan Kaufmann Publishers Inc, 2001: 282-289.
- [8] NOBLE W S. What is a support vector machine? [J]. Nature Biotechnology, 2006, 24(12): 1565-1567.
- [9] SAMPATHKUMAR H, CHEN X W, LUO B. Mining adverse drug reactions from online healthcare forums using hidden Markov model [J]. BMC Medical Informatics and Decision Making, 2014, 14(1): 91.
- [10] TRAN Q, MACKINLAY A, YEPES A J. Named entity recognition with stack residual LSTM and trainable bias decoding [EB/OL]. [2025-01-23]. <https://arxiv.org/pdf/1706.07598>.
- [11] HE H, SUN X. A unified model for cross-domain and semi-supervised named entity recognition in chinese social media [C] // Proceedings of the AAAI conference on artificial intelligence. San Francisco, California, USA: AAAI Press, 2017: 3216-3222.
- [12] SCHUSTER M, PALIWAL K K. Bidirectional recurrent neural networks [J]. IEEE Transactions on Signal Processing, 1997, 45(11): 2673-2681.
- [13] XU K, ZHOU Z F, HAO T Y, et al. A bidirectional LSTM and conditional random fields approach to medical named entity recognition [C] // Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2017. Berlin, Germany: Springer, 2017: 355-365.
- [14] 张红民, 颜鼎鼎, 田钱前. 改进时空图卷积网络的视频异常检测方法 [J]. 光电工程, 2024, 51(5): 240034.
- [15] ZHANG H M, YAN D D, TIAN Q Q. Improved spatio-temporal graph convolutional networks for video anomaly detection [J]. Opto-Electronic Engineering, 2024, 51(5): 240034. (in Chinese)
- [16] 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.
- [17] ZHANG Y, YANG J. Chinese NER using lattice LSTM [EB/OL]. [2025-01-23]. <https://arxiv.org/pdf/1805.02023>.
- [18] MA R T, PENG M L, ZHANG Q, et al. Simplify the usage of lexicon in Chinese NER [EB/OL]. [2025-01-23]. <http://arxiv.org/pdf/1908.05969>.
- [19] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2025-01-23]. <https://arxiv.org/pdf/1810.04805>.
- [20] LIU W, FU X Y, ZHANG Y, et al. Lexicon enhanced Chinese sequence labeling using BERT adapter [EB/OL]. [2025-01-23]. <https://arxiv.org/pdf/2105.07148>.
- [21] 罗凌, 杨志豪, 宋雅文, 等. 基于笔画 ELMo 和多任务学习的中文电子病历命名实体识别研究 [J]. 计算机学报, 2020, 43(10): 1943-1957.
- [22] LUO L, YANG Z H, SONG Y W, et al. Chinese clinical named entity recognition based on stroke ELMo and multi-task learning [J]. Chinese Journal of Computers, 2020, 43(10): 1943-1957. (in Chinese)
- [23] 宋雅文, 杨志豪, 罗凌, 等. 基于字符卷积神经网络的生物医学变异实体识别方法 [J]. 中文信息学报, 2021, 35(5): 63-69.
- [24] SONG Y W, YANG Z H, LUO L, et al. Biomedical mutation entity recognition method based on character convolution neural network [J]. Journal of Chinese Information Processing, 2021, 35(5): 63-69. (in Chinese)
- [25] LI Y B, WANG X H, HUI L H, et al. Chinese clinical named entity recognition in electronic medical records: development of a lattice long short-term memory model with contextualized character representations [J]. JMIR Medical Informatics, 2020, 8(9): e19848.
- [26] WANG Q C, HAIHONG E. A BERT-based named entity recognition in Chinese electronic medical record [C] // Proceedings of the 2020 9th International Conference on Computing and Pattern Recognition. New York, USA: ACM Press, 2020: 13-17.
- [27] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [EB/OL]. [2025-01-23]. <https://export.arxiv.org/pdf/1301.3781>.
- [28] PENNINGTON J, SOCHER R, MANNING C. GloVe: global vectors for word representation [C] // Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Stroudsburg, USA: ACL, 2014: 1532-1543.
- [29] LI F, WANG Z, HUI S C, et al. Modularized interaction network for named entity recognition [C] // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg, USA: ACL, 2021: 200-209.
- [30] SHI J T, SUN M X, SUN Z Y, et al. Multi-level semantic

- fusion network for Chinese medical named entity recognition [J]. Journal of Biomedical Informatics, 2022, 133: 104144.
- [26] AN Y, XIA X Y, CHEN X L, et al. Chinese clinical named entity recognition via multi-head self-attention based BiLSTM-CRF[J]. Artificial Intelligence in Medicine, 2022, 127: 102282.
- [27] 唐卓然, 柳毅. 基于词汇融合和依存关系的中文命名实体识别[J]. 计算机工程, 2024, 50(10): 145-153.
TANG Z R, LIU Y. Chinese named entity recognition based on lexicon fusion and dependency relation [J]. Computer Engineering, 2024, 50(10): 145-153. (in Chinese)
- [28] 林令德, 刘纳, 徐贞顺, 等. 基于多层动态融合的中文医疗命名实体识别[J]. 计算机工程与应用, 2024, 60(15): 161-169.
LIN L D, LIU N, XU Z S, et al. Chinese medical named entity recognition based on multi-layer dynamic fusion [J]. Computer Engineering and Applications, 2024, 60(15): 161-169. (in Chinese)
- [29] LI X N, YAN H, QIU X P, et al. FLAT: Chinese NER using flat-lattice transformer [C] // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: ACL, 2020: 6836-6842.
- [30] 乔锐, 杨笑然, 黄文亢. 基于 BERT 与模型融合的医疗命名实体识别 [C] // 全国知识图谱与语义计算会议评估任务论文集. 杭州, 中国: [s. n], 2019: 1-6.
QIAO R, YANG X R, HUANG W K. Medical named entity recognition based on BERT and model fusion [C] // Proceedings of Evaluation Tasks at the China Conference on Knowledge Graph and Semantic Computing. Hangzhou, China: [s. n], 2019: 1-6. (in Chinese)
- [31] 陈明, 刘蓉, 张晔. 基于多重注意力机制的中文医疗实体识别 [J]. 计算机工程, 2023, 49(6): 314-320.
CHEN M, LIU R, ZHANG Y. Chinese medical entity recognition based on multiple attention mechanism [J]. Computer Engineering, 2023, 49(6): 314-320. (in Chinese)
- [32] WANG C Y, WANG H, ZHUANG H, et al. Chinese medical named entity recognition based on multi-granularity semantic dictionary and multimodal tree [J]. Journal of Biomedical Informatics, 2020, 111: 103583.
- [33] WU S, SONG X N, FENG Z H. MECT: multi-metadata embedding based cross-transformer for Chinese named entity recognition [C] // Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg, USA: ACL, 2021: 1529-1539.
- [34] 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]. 小型微型计算机系统, 2025, 46(2): 330-338.
LU X T, SUN L P, LING C, et al. Named entity recognition of Chinese electronic health records incorporating phonetic and part-of-speech features [J]. Journal of Chinese Computer Systems, 2025, 46(2): 330-338. (in Chinese)

文字编辑 薛晋栋
栏目编辑 赖玉玲