计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
[doi: 10.15888/j.cnki.csa.009869] [CSTR: 32024.14.csa.009869]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

融合实体头尾关键特征的命名实体识别①

雷海卫, 宋朝帅

(中北大学 计算机科学与技术学院, 太原 030051)
通信作者: 宋朝帅, E-mail: 1733461873@qq.com

摘　要: 针对传统基于阅读理解框架的命名实体识别 (NER) 方法存在的单条样本实体数量稀释以及在预测实体头尾时缺乏对实体完整位置信息的利用这两方面问题, 本文基于阅读理解框架提出一种融合实体头尾关键特征的医学文本命名实体识别模型 (integrate key feature of entity start and end position, IKFSE). 首先, 设计一种实体头尾关键特征提取模块, 提取出针对医学实体起始位置和结束位置的关键特征, 减少冗余信息对模型的影响; 其次, 设计一种实体头尾特征交叉融合模块, 在对实体起始位置和结束位置进行预测时分别引入二者对彼此的影响, 从而引入实体完整的位置信息, 提高模型的语义表征能力. 在 cEHRNER 和 CCKS2017 两个公开数据集上将 IKFSE 与多个主流序列标注模型和阅读理解模型相比, 结果表明本文所提方法在中文医学 NER 任务中有着更好的性能.

关键词: 医学文本; 命名实体识别; 关键特征; 特征融合; 完整位置信息

引用格式: 雷海卫,宋朝帅.融合实体头尾关键特征的命名实体识别.计算机系统应用. http://www.c-s-a.org.cn/1003-3254/9869.html

Named Entity Recognition with Integrated Key Feature of Entity Start and End

LEI Hai-Wei, SONG Chao-Shuai

(School of Computer Science and Technology, North University of China, Taiyuan 030051, China)

Abstract: To address the challenges of entity dilution in single samples and the underutilization of complete entity location information in traditional named entity recognition (NER) methods based on reading comprehension frameworks, this study proposes a medical named entity recognition model that integrates key features of entity start and end positions (IKFSE). First, a key feature extraction module is designed to capture important features of the start and end positions of medical entities, thus reducing the impact of redundant information on the model. Second, an entity head-tail feature fusion module is introduced, which incorporates the mutual influence between the start and end positions during entity prediction. This integration enhances the model's ability to represent semantics by utilizing full positional information. Experimental results on two public datasets, cEHRNER and CCKS2017, demonstrate that the proposed method outperforms several mainstream sequence labeling models and reading comprehension models in Chinese medical NER tasks.

Key words: medical text; named entity recognition (NER); key feature; feature fusion; complete location information

近年来, 随着医疗信息化的不断推进, 各种医疗信息管理系统层出不穷, 随之也积累了庞杂的医疗资料, 这些资料包括但不限于电子病历记录、各类医学检测报告及丰富的医疗学术论文[1]. 其中电子病历中蕴含了大量的临床信息和医学专业知识, 如何挖掘电子病历中的知识, 提高医疗水平已成为国内外研究者共同关注的问题. 随着自然语言处理技术 (NLP) 和深度学习技术的出现和发展, 利用 NLP 从电子病历中提取有价值的知识以促进医疗数据的有效利用已成为医学和人工智能交叉领域的研究焦点[2].

① 收稿时间: 2024-11-06; 修改时间: 2024-12-09; 采用时间: 2024-12-25; csa 在线出版时间: 2025-04-28

命名实体识别任务 (NER) 是 NLP 技术挖掘电子病历的第 1 步, 其核心在于提取医学文本中的关键实体, 并正确分类它们的类别, 如“身体部位”“药品”“疾病与诊断”“治疗操作”及“检测项目”等. 相较于一般文本, 医学文本的特点是专业性强且结构更为复杂. 医学实体类型更加多样, 且其中蕴含着丰富的语义信息和大量的专业知识, 因此能否引入医学实体类型标签类型中的语义知识对于医学 NER 任务有着重要影响.

以往的研究大多将 NER 任务视为一个序列标注问题. 然而, 与基于阅读理解框架 (MRC) 的方法相比, 基于序列标注的 NER 方法无法充分利用医学实体类型标签所包含的语义知识. 基于 MRC 的方法通过将实体类型标签转化为自然语言问题 (Question), 并将该问题与目标文本一同输入模型进行训练, 从而有效地引入了标签中的语义知识. 但这种方法在实际应用中也存在一定的问题, 需要进一步优化和改进.

一方面, 基于 MRC 的方法将每个类型标签构建的 Question 与文本拼接, 增加了原有数据集样本的数量, 但实体总量没有增加, 因此导致单条数据中的实体数量被稀释, 冗余信息增加. Chen 等人[3]在情感识别领域提出一种基于情绪相关信息有效识别情绪的关键情绪提取模块, 本文借鉴其思想并应用于基于 MRC 的医学文本 NER 任务中, 设计了一种实体头尾关键特征提取模块, 提取出针对医学实体起始位置和结束位置的关键特征, 成功解决单条样本实体稀释的问题.

另一方面, 传统基于 MRC 的方法在预测实体头尾时缺乏对于实体完整位置信息的利用, 因此本文基于 Transformer[4]结构设计了一种实体头尾特征交叉融合模块, 在对实体起始位置和结束位置进行预测时分别引入二者对彼此的影响, 以提高模型的语义表征能力.

基于以上两方面的考虑, 本文提出一种融合实体头尾关键特征的医学命名实体识别模型 (integrate key features of entities start and end positions, IKFSE), 本文的主要贡献有以下几点.

(1) 设计一种实体头尾关键特征提取模块, 借鉴情感识别领域的关键情绪提取思想, 将其应用于医学文本 NER 领域, 以解决 MRC 方法存在的单条文本实体数量稀释问题.

(2) 设计一种实体头尾特征交叉融合模块, 在对实体起始位置和结束位置进行预测时分别引入二者对彼此的影响, 引入了实体完整位置信息, 从而提高了模型的语义表征能力.

(3) 提出一种医学文本命名实体识别模型 IKFSE, 并在公开数据集 cEHRNER 和 CCKS2017 上取得了良好性能, 与基线模型 BERT-MRC 相比, $F1$ 分数分别提升了 1.83% 和 0.95%.

1 相关工作

传统的 NER 任务以基于规则的方法和基于机器学习的方法为主, 基于规则的方法需要针对具体的任务场景构造人工规则, 通过匹配的方式识别文本中的特定实体, 具有可解释性强的优点. 其中具有代表性的是 Friedman 等人[5]基于医学词典和语法规则提出的医学 NER 医疗语言提取与编码系统. 基于规则的方法在特定任务和语料上可以实现较高的准确率, 但这种方法往往伴随着高昂的规则制定成本和较差的泛化能力. 随着机器学习技术的不断发展, NER 领域开始采用基于机器学习的方法, 将 NER 任务视为序列标注任务, 通过训练模型来预测文本中每个词汇的实体标签, 从而找到最有可能的标签序列组合. 具有代表性的模型有隐马尔可夫模型 (HMM)[6]、支持向量机 (SVM)[7]和条件随机场 (CRF)[8]等, 在特定时期内取得了显著成效. 然而, 这些基于机器学习的方法往往对特征工程有较高的依赖, 而且在提取深层语义特征方面存在局限.

近年来, 随着深度学习在自然语言处理任务中的深入研究, 深度学习技术在当下的命名实体识别领域中占据了主导地位, 成为研究和应用的热点. 深度学习方法在构造词向量方面具有显著优势, 这些词向量不仅蕴含丰富的语义信息, 超越了传统人工特征的选择, 而且能将来自不同文本的数据映射到统一的向量空间, 因此深度学习在 NER 任务中表现出色[9]. 由于文本数据通常需要理解上下文才能准确地进行处理, RNN-CRF 方法利用了 RNN 对序列数据的强大处理能力, 以及 CRF 在标注任务中的优势, 因而在中文 NER 中取得了显著成效, 例如 Lample 等人[10]提出一种双向长短记忆网络 BiLSTM, 通过同时考虑文本序列的前向和后向信息, 有效地提取了文本中的双向特征, 并与 CRF 结合, 进一步提高了命名实体识别任务的精度.

为了提高中文字词表征的多义性, 表征更丰富的语义特征, 预训练模型应运而生, 其中最受欢迎的是 Google 团队 Devlin 等人[11]提出的预训练语言模型 BERT, 它基于 Transformer 编码器, 采用双向训练策略, BERT 的双

向特性使其在理解语境方面有着出色的性能. 近几年的中文 NER 研究中, BERT 及其变体在语义特征提取中发挥了巨大的作用. Cui 等人[12]提出一种基于语言信息增强的预训练模型 LERT, 利用 3 种语言学任务进行训练, 融入大量的语言学特征. Liu 等人[13]提出 LEBERT, 利用词典适配器, 将外部词典的丰富知识融入 BERT 模型层, 从而加强了 BERT 底层结构的深度知识整合.

上述工作中, NER 任务多被转为序列标注问题进行处理, 而此种方法忽略了实体类型标签中蕴含的语义信息. Li 等人[14]提出了一个统一的框架, 该框架能够应对平面和嵌套的命名实体识别任务, 将命名实体识别的任务转化为机器阅读理解的形式, 通过为每个实体类别设计独特的自然语言问题来实现这一过程. 提取某一类实体, 就形式化为提取该类实体对应自然语言问题的答案跨度. 通过自然语言问题的构造, 引入丰富的先验知识, 可以让模型学习到实体类型标签中的语义信息. 然而此种方法通过将每一条文本与类型描述进行拼接扩充了数据集, 导致单条数据中实体的数量被稀释, 文本中的冗余信息增加, 并且缺乏对实体完整位置信息的利用.

2 融合实体头尾关键特征的命名实体识别方法

2.1 问题描述

命名实体识别任务可以视为阅读理解过程, 通过为各类医学实体设计特定的问题或描述, 来指导模型在医学文本中寻找和识别相应的实体.

给定一个医学文本 $X=\{x_1,x_2,\cdots,x_n\}$, 其中 x_i 代表文本中每一个字符, n 为给定文本长度. 我们需要找出文本中的每一个医疗实体并为其分配一个标签 $y\in Y$, 其中 Y 是所有可能的标签类型, 根据实体在文本 X 中的起始位置 X_{start} 和结束位置 X_{end}, 可以将每一个实体表示为 $X_{\text{start,end}}=\{x_{\text{start}},\cdots,x_{\text{end}}\}$, 而每一个标签 $y\in Y$ 一一对应一个自然语言问题 $Q_y=\{q_1,q_2,\cdots,q_m\}$, 其中 m 为问题 Q_y 的长度. 因此每一条文本可以表示为三元组 $\{Q_y,X_{\text{start,end}},X\}$, 则医学实体的识别过程为:

$$f:(Q_y,X)\rightarrow X_{\text{start,end}} \tag{1}$$

2.2 模型总体架构

模型的整体框架如图 1 所示, 首先使用预训练模型对实体类型标签描述和医学文本进行嵌入, 然后通过两个关键特征提取模块 (Key-Transformer) 分别提取出实体起始位置和结束位置的关键特征, 然后将二者一同输入实体头尾特征交叉融合模块 (Cross-SE-Transformer), 引入实体完整的位置信息, 最后通过两个二分类器分别对实体在文本中的起始位置和结束位置进行预测, 根据就近原则将二者匹配最终识别出完整的医学实体.

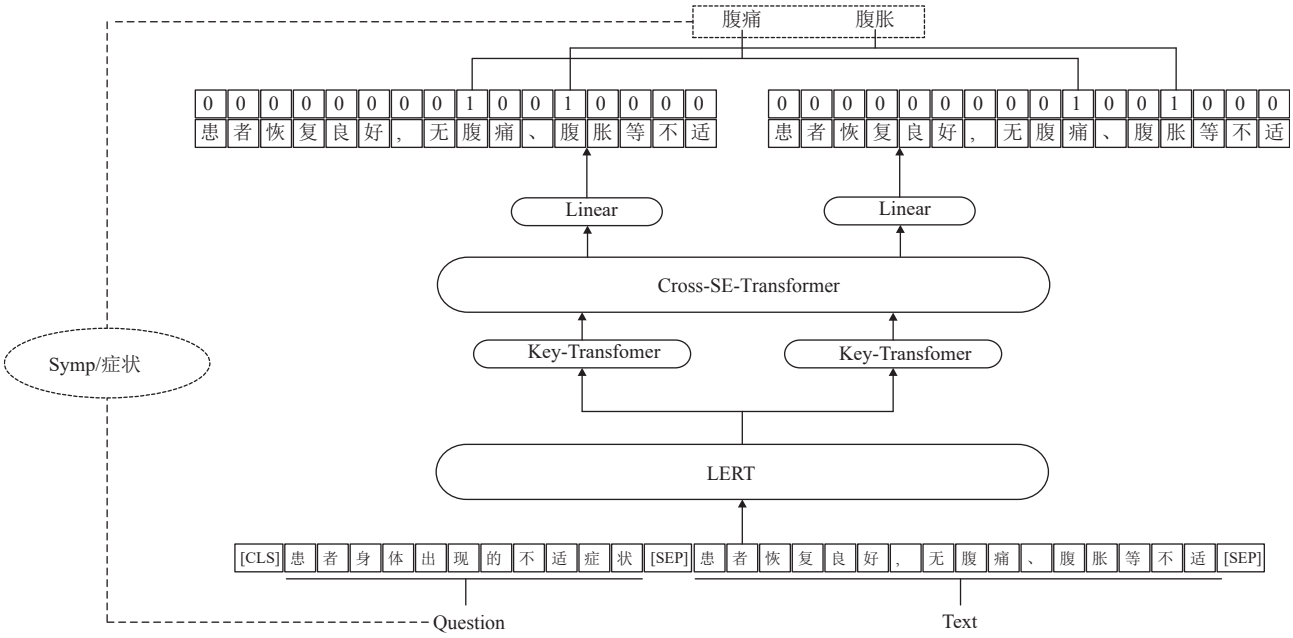

图 1 IKFSE 模型整体结构

2.3 预训练嵌入

2.3.1 LERT 预训练模型

LERT[12]预训练模型是哈工大讯飞联合实验室提出的一种基于语言信息增强的预训练模型, 与 BERT 预训练模型相比, LERT 融入了大量的语言学特征, 除了使用掩码进行训练之外, 还采用了 3 种语言学任务进行训练, 提出了一种语言学启发的预训练机制 (LIP), 使模型更好地学习到语言学特征.

2.3.2 问题和文本嵌入

将针对实体类型标签构造的自然语言问题Q_y与医学文本用[CLS]和[SEP]连接起来, 输入格式如下:

$$input = [\mathrm{CLS}], q_1, q_2, \cdots, q_m, [\mathrm{SEP}], x_1, x_2, \cdots, x_n, [\mathrm{SEP}] \tag{2}$$

使用 LERT 进行预训练编码得到问题和文本嵌入 $H_{input} \in R^{n \times d}$:

$$H_{input} = LERT\,(input) \tag{3}$$

2.4 实体头尾关键特征提取模块 (Key-Transformer)

为了帮助模型提取到针对于医学实体起始位置和结束位置更为相关的特征, 本文设计一种关键特征提取模块, 保留文本中注意力分数较高的部分权重, 其余部分视为无关的冗余信息, 将其权重置为 0.

2.4.1 Key-Attention

Key-Attention 机制能够自动判断文本序列中每个 token 对于实体起始或结束位置的重要性, 其机制如图 2 所示, 其中⊙和⊗分别表示位置乘法和矩阵乘法. 假设查询向量 Q 的长度为 i, 关键向量 K 的长度为 j, 将 Q 和 K 相乘并经 *Softmax* 函数进行归一化得到权重矩阵 W. V 中的值向量代表文本序列中的每一个 token, 而 W 中每一行的值均代表每个 token 所对应的权重, 将相同 token 对应的权重累加得到 S, 即 W 中的每一列之和.

受到 Chen 等人[3]启发, 根据每个 token 所对应的权重之和来判断其对于文本中实体起始或结束位置的重要性. 将 S 中的 j 个权重之和从大到小排序并挑选前 k 个, 保持这 k 个 token 在权重矩阵 W 中的权值不变, 其余置为 0, 最后再和 V 相乘. 通过此操作便可减少不重要的冗余信息, 从而使模型提取出针对实体起始位置和结束位置更为关键的特征. 图 2 中 Top k mask 由式 (4) 得到:

$$M_x = \begin{cases} 0, & S_x < threshold \\ 1, & S_x \geqslant threshold \end{cases} \tag{4}$$

其中, *threshold* 为 j 个权重之和从大到小排序后的第 k 个值, $x \in [1, j]$.

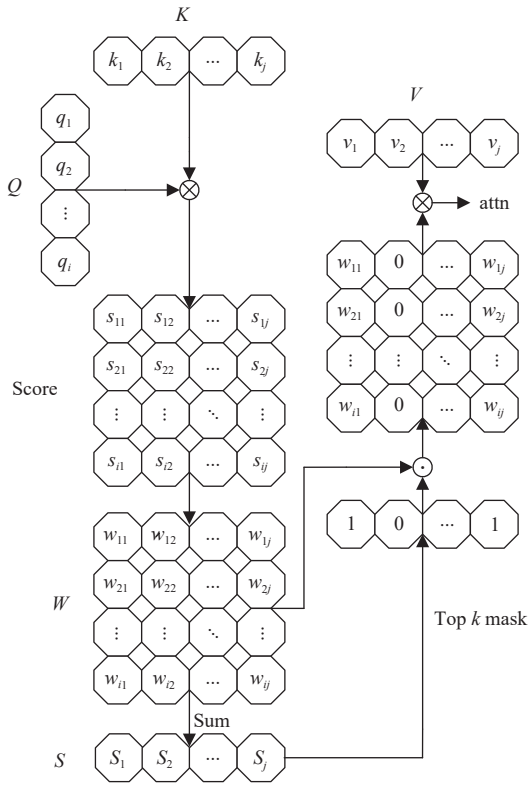

图 2 Key-Attention 机制图

2.4.2 Key-Transformer

Transformer[4]最初是由编码器和解码器组成, 本文仅使用其中的编码器, 由 Key-Attention 替换原本的多头注意力便得到 Key-Transformer. 将 LERT 预训练编码得到的问题和文本嵌入输入 Key-Transformer, 便可提取出实体起始位置和结束位置的关键特征:

$$KT_{S(E)} = KeyTransformer_{\mathrm{start(end)}}(H_{input}) \tag{5}$$

2.5 实体头尾特征交叉融合模块 (Cross-SE-Transformer)

通过两个 Key-Transformer 模块分别提取出实体起始位置和结束位置的关键特征, 然而只根据单一特征去预测实体的起始位置或者结束位置, 忽略了实体整体的位置信息. 实体的起始位置和结束位置共同表示了一个实体的完整信息, 因此本文基于 Transformer[4]结构设计了一个实体头尾特征融合, 通过两个 Transformer 模块对实体的起始位置特征和结束位置特征进行交叉注意力计算, 以得到二者对彼此的影响, 将注意力计算结果与实体起始和结束位置本身特征相结合从而引入实体完整的位置信息, 帮助模型更好地识别出

实体的边界. 该模块的具体结构如图 3 所示, 其中⊕为向量的拼接.

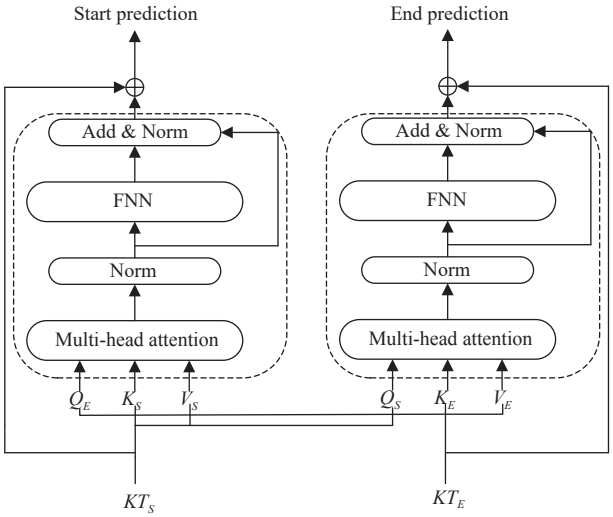

图 3 Cross-SE-Transformer 模块结构图

$Q_{S(E)}$, $K_{S(E)}$, $V_{S(E)}$分别由关键特征提取模块 Key-Transformer 的输出经线性变换得到:

$$\begin{bmatrix} Q_{S(E),i} \\ K_{S(E),i} \\ V_{S(E),i} \end{bmatrix}^{\mathrm{T}} = KT_{S(E),i} \begin{bmatrix} W_{S(E),Q} \\ W_{S(E),K} \\ W_{S(E),V} \end{bmatrix}^{\mathrm{T}} \tag{6}$$

其中, KT_S 和KT_E分别是由两个 Key-Transformer 输出的实体起始位置和结束位置的关键特征, 每个 W 都是一个可学习的参数. 接着, 为了引入实体完整的位置信息, 我们在计算注意力时将两者进行了融合. 注意力计算方式如式 (7)–式 (9):

$$Attn_{S(E)}(Q_{E(S)}, K_{S(E)}, V_{S(E)}) = Softmax(Q_{E(S)}K_{S(E)}^{\mathrm{T}})V_{S(E)} \tag{7}$$

$$head_{S(E),i} = Attn_{S(E)}(Q_{E(S),i}, K_{S(E),i}, V_{S(E),i}) \tag{8}$$

$$MultiHead_{S(E)} = Concat(head_{S(E),1}, \cdots, head_{S(E),i})W^{O} \tag{9}$$

其中, W^O 是最终输出的变换矩阵, Concat 是拼接操作. 以实体的起始位置特征为例, 使用多头注意力引入实体结束位置特征对其的影响, 经过层归一化、前馈层后, 再与 Key-Transformer 输出的实体起始位置关键特征 KT_S 拼接在一起, 得到最终的起始位置特征E_{start}. 最终的结束位置特征E_{end}的获得方式亦是如此.

2.6 答案预测

根据头尾信息融合模块 Cross-SE-Transformer 的输出, 我们可以分别得到实体起始位置和结束位置的特征, 采用两个独立的二分类器对文本序列中的每个 token 作为实体起始位置和结束位置的概率进行预测, 如式 (10) 和式 (11) 所示:

$$P_{\text{start}} = Softmax(E_{\text{start}} \cdot T_{\text{start}}) \in R^{n\times 2} \tag{10}$$

$$P_{\text{end}} = Softmax(E_{\text{end}} \cdot T_{\text{end}}) \in R^{n\times 2} \tag{11}$$

其中, T_{start}、$T_{\text{end}} \in R^{d\times 2}$ 是可以学习的参数矩阵, E_{start}、E_{end}是由 Cross-SE-Transformer 得到的交叉融合后的实体起始位置和结束位置特征, P_{start}、P_{end}分别表示该 token 在医学文本中作为实体起始位置或结束位置的概率. 在获取到每个 token 的概率后, 对P_{start}、P_{end}的每一行执行 argmax 操作, 即可得到两个长度为 n, 取值为 0 或 1 的序列:

$$I_{\text{start}} = \left\{ i \mid \text{argmax}(P_{\text{start}}^{(i)}) = 1,\ i = 1, \cdots, n \right\} \tag{12}$$

$$I_{\text{end}} = \left\{ i \mid \text{argmax}(P_{\text{end}}^{(i)}) = 1,\ i = 1, \cdots, n \right\} \tag{13}$$

获取到文本中实体起始位置和结尾位置的 0–1 序列后, 需将二者进行匹配才能得到实体完整的位置信息. 由于本文所用数据集均为平面数据集不存在实体嵌套情况, 因此根据就近原则, 将起始位置与最近的结束位置相匹配即可得到最终的实体边界.

2.7 训练和测试

我们对实体在医学文本中的起始位置和结束位置进行预测时涉及两个损失:

$$L_{\text{start}} = CE(P_{\text{start}}, Y_{\text{start}}) \tag{14}$$

$$L_{\text{end}} = CE(P_{\text{end}}, Y_{\text{end}}) \tag{15}$$

其中, Y_{start}和Y_{end}分别为医学文本中实体起始位置和结束位置的真实标签, CE 为交叉熵损失函数. 整体训练损失为:

$$L = \alpha L_{\text{start}} + \beta L_{\text{end}} \tag{16}$$

其中, α, $\beta \in [0, 1]$, 是控制L_{start}和L_{end}对整体训练损失 L 贡献的超参数. 测试时首先根据I_{start}和I_{end}分别确定实体在医学文本中的起始位置和结束位置, 然后根据就近原则将起始位置和结束位置进行匹配, 最终识别出医学实体.

3 实验结果及分析

3.1 数据集

本文选用中文医疗信息处理评测基准 CBLUE 中

的 cEHRNER 数据集和全国知识图谱与语义计算大会于 2017 年发布的中文临床数据集 CCKS2017 进行实验. cEHRNER 数据集中标注了解剖部位、药物、手术、疾病和诊断、症状、影像检查和实验室检验共 7 种不同的实体类型, 包括 914 条训练数据集、44 条验证数据集和 41 条测试数据集; CCKS2017 数据集中标注了症状和体征 (symp)、疾病和诊断 (dise)、检查和检验 (chec)、身体部位 (body) 以及治疗 (cure) 共 5 种不同的实体类型, 包括 2 006 条训练数据集和 223 条测试数据集. cEHRNER 和 CCKS2017 中不同类型的医疗命名实体统计数据如表 1、表 2 所示.

表 1 cEHRNER 数据集信息

类型	训练集	验证集	测试集
解剖部位	5 623	252	220
药物	1 646	84	72
手术	946	52	43
疾病和诊断	3 824	173	149
症状	2 095	78	88
影像检查	889	55	29
实验室检验	1 113	55	31

表 2 CCKS2017 数据集信息

类型	训练集	测试集
body	17 556	1 996
symp	12 821	1 473
dise	4 560	143
chec	17 655	1 750
cure	4 940	169

为将阅读理解框架引入医疗文本的命名实体识别任务中, 需将数据集格式转换为三元组格式: {问题, 答案, 上下文}.

关于自然语言问题 Q_y 的设置, 本文针对每一个实体类型进行详细的描述, 通过对实体类型标签进行扩充描述来帮助模型区分相近的实体类别. cEHRNER 和 CCKS2017 中每种实体及其对应的描述如表 3、表 4 所示.

最终转换后的数据集格式如图 4 所示.

3.2 实验设置

本研究在 NVIDIA GeForce RTX 3090 显卡上进行实验, 开发语言为 Python 3.7, 开发环境为 PyTorch 1.9.0. 选用“Chinese-LERT-base”作为预训练模型, 词嵌入维度 d 为 768 维, 选择 AdamW 算法作为优化器进行权重更新, Dropout 为 0.2, 训练轮数为 50 个 epoch. cEHRNER 数据集的学习率设为 5E–5, 批量大小为 8, CCKS2017 数据集的学习率设为 7E–5, 批量大小为 16.

表 3 cEHRNER 数据集实体类型描述

实体类型	实体描述
解剖部位	细胞、组织以及位于人体特定区域的由细小物质成分组合而成的结构、器官、系统、肢体
药物	用来预防、治疗及诊断疾病的物质
手术	人体局部开展去除病变组织、修复损伤、重建形态或功能、移植细胞组织或器官、植入医疗器械等医学操作的医疗技术
疾病和诊断	根据各种疾病的临床特点, 对病人做出相应的诊断, 确定病人所患何种疾病
症状	患者身体出现的不适症状
影像检查	利用超声波、X射线等介质, 将人体组织器官结构、密度以影像学的形式表现出来, 用于判断局部是否存在病变
实验室检验	对患者的血液、尿液、体液等进行检查, 以获得患者的生化指标、免疫学指标、血液学指标等信息

表 4 CCKS2017 数据集实体类型描述

实体类型	实体描述
body	细胞、组织以及位于人体特定区域的由细小物质成分组合而成的结构、器官、系统、肢体
symp	患者身体出现的不适症状
dise	根据各种疾病的临床特点, 对病人作出相应的诊断, 确定病人所患何种疾病
chec	通过一系列的实验室检验和医学影像检查等方法, 对患者的身体状况进行全面的评估和诊断的过程
cure	用药物、手术等方式消除疾病


```
{
 "context": "中年男性，48岁，主因：腹痛、腹胀7天，停止排气排便3天。于2016-2-25，22：40入院。",
 "end_position": [
  25
 ],
 "entity_label": "body",
 "impossible": false,
 "qas_id": "1.1",
 "query": "细胞、组织以及位于人体特定区域的由细小物质成分组合而成的结构、器官、系统、肢体",
 "span_position": [
  "25;25"
 ],
 "start_position": [
  25
 ]
},
```


图 4 转换为阅读理解格式的数据集样式

3.3 评价指标

本文中使用精确率 P、召回率 R 以及 $F1$ 分数 3 个指标来衡量模型性能, 定义如下:

$$R = \frac{T_P}{T_P + F_N} \times 100\% \tag{17}$$

$$P = \frac{T_P}{T_P + F_P} \times 100\% \tag{18}$$

$$F1 = \frac{2PR}{P + R} \times 100\% \tag{19}$$

其中, T_P 为真正例, F_P 为假正例, F_N 为假负例. 精确率 P 表示模型识别到的所有实体中正确实体的占比, 量化了模型识别实体的精确度; 召回率 R 表示所有实际存在实体中, 模型正确识别的比例, 衡量了模型发现实体的全面性; 二者是相互矛盾的, 无法同时达到最高水平, $F1$ 结合精确率 P 和召回率 R 对模型进行综合评估, 通常用于表征模型效果, 其值越高则表示模型综合性能越好.

3.4 实验结果

3.4.1 对比实验

本文选用 3 个优秀序列标注模型和 1 个机器阅读理解模型与本文模型对比.

BERT-BiLSTM-CRF[15]: 在双向 LSTM 和 CRF 的经典模型前面使用了预训练模型 BERT 进行文本预训练嵌入.

BERT-BiLSTM-contex-CRF[16]: 设计了一个全局上下文机制, 将全局句子信息集成到 BiLSTM 每个单元格的句子表示中, 增强了 BiLSTM 的句子表征能力.

LEBERT-CRF[11]: 通过词典适配器将外部词典知识集成到 BERT 层中, 促进了 BERT 底层的深层知识融合.

BERT-MRC[12]: 提出一种能够处理平面和嵌套 NER 的统一框架, 将命名实体识别任务表示为机器阅读理解任务, 也是本文的基线模型.

此外本文还选择了一些相同数据集上的先进模型来进行比较.

(1) cEHRNER 数据集

ERBEGP[17]: 将知识图谱与预训练模型相结合, 提出了一种基于知识增强的中文病历文本实体识别模型.

BBCPR[18]: 提出 POS 融合层整合外部语法知识, 并引入了一种新的正则化方法, 通过结合对抗性训练和 Dropout 来提高模型的鲁棒性.

(2) CCKS2017 数据集

FT-BERT+BiLSTM+CRF+Fea[19]: 在未标记的中文临床记录上训练 BERT 模型, 在 BiLSTM-CRF 的基础上融合外部词典和部首特征.

ACNN[20]: 设计一个多级的卷积神经网络, 以同时捕捉局部短期和全局长期上下文信息.

CMNER[21]: 提出一种融合形态、字符、单词和句法层面语义信息的多层语义融合网络.

具体实验结果如表 5 和表 6 所示, 其中标有*为引用原论文数据, 其余为复现数据.

在两个公开医学命名实体识别数据集上, 本文模型的 $F1$ 分数分别达到了 92.87% 和 86.13%, 相比于 BERT-BiLSTM-CRF、BERT-BiLSTM-contex-CRF 和 LEBERT-CRF 这 3 个序列标注模型, 在 CCKS2017 数据集上分别提高了 0.52%、0.45% 和 0.28%, 在 cEHR-NER 数据集上分别提升了 2.48%、1.14% 和 1.66%. 可以看出本文通过阅读理解框架引入实体标签类型中的语义信息, 模型能够更好地识别出医学文本中的医学实体. 与基准模型 BERT-MRC 相比, 在两个公开数据集上分别提升了 0.95% 和 1.83%, 可以看出本文通过使用改进后的注意力计算方式提取实体起始位置和结束位置的关键特征, 解决了阅读理解范式带来的文本中实体稀释问题, 并在预测实体头尾时有效引入了实体完整的位置信息, 提高了模型实体识别精度, 使得阅读理解框架更好地应用于命名实体识别任务中. 与相同数据集上的先进模型对比, 本文模型同样有着明显的提升, 可知本文模型通过融合实体头部和尾部的关键特征能够更好地识别出医学实体从而提升医学文本的命名实体识别精度.

表 5 cEHRNER 数据集对比实验结果 (%)

模型	P	R	$F1$
ERBEGP	—	—	80.97*
BBCPR	83.28*	84.97*	84.11*
BERT-BiLSTM-CRF	82.69	84.64	83.65
BERT-BiLSTM-contex-CRF	83.61	86.41	84.99
LEBERT-CRF	82.79	86.22	84.47
BERT-MRC	**85.36**	83.28	84.30
IKFSE (本文)	85.18	**87.11**	**86.13**

表 6 CCKS2017 数据集对比实验结果 (%)

模型	P	R	$F1$
FT-BERT+BiLSTM+CRF+Fea	92.06*	91.15*	91.60*
ACNN	90.19*	90.78*	90.49*
CMNER	91.44*	92.30*	91.87*
BERT-BiLSTM-CRF	91.74	92.97	92.35
BERT-BiLSTM-contex-CRF	91.50	93.35	92.42
LEBERT-CRF	91.70	93.51	92.59
BERT-MRC	91.77	92.06	91.92
IKFSE (本文)	**92.14**	**93.61**	**92.87**

3.4.2 消融实验

为了深入探究 IKFSE 模型各个模块的有效性, 本文设计了 6 个模型进行消融实验研究, 实验结果详见表 7, 其中最优数据加粗表示.

(1) BERT-MRC: 使用 BERT 预训练模型, 去掉 Key-Transformer 和 Cross-SE-Transformer 模块.

(2) BERT-KT-MRC: 使用 BERT 预训练模型, 去掉 Cross-SE-Transformer 模块.

(3) BERT-KT-CT-MRC: 使用 BERT 预训练模型.

(4) LERT-MRC: 使用 LERT 预训练模型, 去掉 Key-Transformer 和 Cross-SE-Transformer 模块.

(5) LERT-KT-MRC: 使用 LERT 预训练模型, 去掉 Cross-SE-Transformer 模块.

(6) IKFSE: 本文模型.

表 7 消融实验结果 (%)

模型	cEHRNER			CCKS2017		
	P	*R*	*F*1	*P*	*R*	*F*1
(1)	**85.36**	83.28	84.30	91.77	92.06	91.92
(2)	83.39	85.71	84.54	91.95	92.37	92.16
(3)	84.35	86.41	85.37	92.19	92.72	92.45
(4)	85.06	84.32	84.69	**92.38**	92.22	92.30
(5)	85.11	84.67	84.89	91.96	92.97	92.46
(6)	85.18	**87.11**	**86.13**	92.14	**93.61**	**92.87**

从表 7 中数据可知, 在两个医学数据集 cEHRNER 和 CCKS2017 上, 本文完整的模型 *F*1 分数最高, 替换或者去掉本文模型中的任何一个模块都会导致精度下降.

模型 (1)–(3) 与 (4)–(6) 相比, 将预训练模型使用 BERT 替换 LERT, 模型在 cEHRNER 数据集和 CCKS-2017 数据集上 *F*1 分数均有所下降, 可知 LERT 通过引入语言信息增强能提升预训练模型的语义表征能力.

模型 (1) 和 (2) 相比, (4) 和 (5) 相比, 可知去掉模型中的 Key-Transformer 模块, 使模型不能识别出与实体起始和结束位置相关的关键特征, 导致模型在两个数据集上的识别效果都有一定程度下降.

模型 (2) 和 (3) 相比, (5) 和 (6) 相比, 可知去掉模型中的 Cross-SE-Transformer 模块, 单独使用实体的起始位置特征和结束位置特征去分别预测实体的起始位置和结束位置, 忽略了实体整体的语义信息, 导致模型的语义表征能力有所下降, 从而导致模型精度下降.

模型 (1) 和 (3) 相比, (4) 和 (6) 相比, 可知去掉模型中的 Key-Transformer 和 Cross-SE-Transformer 模块, 模型在两个数据集上的精度均有明显下降, 可知二者组合之后对模型的性能仍有积极影响.

综上可知, 本文模型中各个模块对模型的整体性能均有积极影响.

4 结论与展望

本文介绍了融合实体头尾关键特征的医学命名实体识别模型 IKFSE, 针对基于阅读理解的命名实体识别方法存在的两方面问题进行了改进. 考虑到阅读理解框架存在的单条语句实体数量稀释问题, 为减少冗余信息对模型的影响, 设计一种实体头尾关键特征提取模块, 提取出实体起始位置和结束位置的关键特征; 考虑到传统阅读理解模型预测实体头尾时缺乏对实体完整位置信息的利用, 设计一种实体头尾特征交叉融合模块, 以提高模型语义表征能力. 通过在 cEHRNER 和 CCKS2017 两个公开数据集上进行对比实验表明, 在中文医学命名实体识别任务中, 本文模型有着更优秀的性能, 并且通过消融实验对模型各组成部分的有效性进行了验证. 未来的工作可以针对阅读理解框架中自然语言问题的构造, 即优化实体类型标签的语义表达, 从而增强实体类型标签和文本的交互作用, 进而提高模型的实体识别精度.

参考文献

1 徐国海. 面向中文医疗文本的命名实体识别研究 [硕士学位论文]. 上海: 华东师范大学, 2019.

2 王红, 王彩雨. 中文医疗命名实体识别方法研究综述. 山东师范大学学报 (自然科学版), 2021, 36(2): 109–117.

3 Chen WD, Xing XF, Xu XM, *et al*. Key-sparse Transformer for multimodal speech emotion recognition. Proceedings of the 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). Singapore: IEEE, 2022. 6897–6901.

4 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.

5 Friedman C, Alderson PO, Austin JHM, *et al*. A general natural-language text processor for clinical radiology. Journal of the American Medical Informatics Association, 1994, 1(2): 161–174. [doi: 10.1136/jamia.1994.95236146]

6 Morwal S, Jahan N, Chopra D, *et al*. Named entity recognition using hidden Markov model (HMM). International Journal on Natural Language Computing (IJNLC), 2012, 1(4): 15–23. [doi: 10.5121/ijnlc.2012.1402]

7 Ju ZF, Wang J, Zhu F. Named entity recognition from biomedical text using SVM. Proceedings of the 5th International Conference on Bioinformatics and Biomedical Engineering. Wuhan: IEEE, 2011. 1–4.

8 Song SL, Zhang N, Huang HT. Named entity recognition based on conditional random fields. Cluster Computing,

2019, 22(3): 5195–5206.

9 Chang Y, Kong L, Jia KJ, *et al*. Chinese named entity recognition method based on BERT. Proceedings of the 2021 IEEE International Conference on Data Science and Computer Application (ICDSCA). Dalian: IEEE, 2021. 294–299.

10 Lample G, Ballesteros M, Subramanian S, *et al*. Neural architectures for named entity recognition. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego: ACL, 2016. 260–270.

11 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional Transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: ACL, 2019. 4171–4186.

12 Cui YM, Che WX, Wang SJ, *et al*. LERT: A linguistically-motivated pre-trained language model. arXiv:2211.05344, 2022.

13 Liu W, Fu XY, Zhang Y, *et al*. Lexicon enhanced Chinese sequence labeling using BERT adapter. Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). ACL, 2021. 5847–5858.

14 Li XY, Feng JR, Meng YX, *et al*. A unified MRC framework for named entity recognition. Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. ACL, 2020. 5849–5859.

15 Dai ZJ, Wang XT, Ni P, *et al*. Named entity recognition using BERT BiLSTM CRF for Chinese electronic health records. Proceedings of the 12th International Congress on Image and Signal Processing, Biomedical Engineering and Informatics (CISP-BMEI). Suzhou: IEEE, 2019. 1–5.

16 Xu CL, Shen K, Sun HG. Supplementary features of BiLSTM for enhanced sequence labeling. arXiv:2305.19928, 2023.

17 李宛泽, 宋波, 齐岳山. 基于知识增强的中文电子病历命名实体识别. 计算机系统应用, 2023, 32(12): 112–119. [doi: 10.15888/j.cnki.csa.009322]

18 Jiang M, Zhang X, Chen CH, *et al*. Leveraging part-of-speech tagging features and a novel regularization strategy for Chinese medical named entity recognition. Mathematics, 2022, 10(9): 1386. [doi: 10.3390/math10091386]

19 Li XY, Zhang H, Zhou XH. Chinese clinical named entity recognition with variant neural structures based on BERT methods. Journal of Biomedical Informatics, 2020, 107: 103422. [doi: 10.1016/j.jbi.2020.103422]

20 Kong J, Zhang LX, Jiang M, *et al*. Incorporating multi-level CNN and attention mechanism for Chinese clinical named entity recognition. Journal of Biomedical Informatics, 2021, 116: 103737. [doi: 10.1016/j.jbi.2021.103737]

21 Shi JT, Sun MX, Sun ZY, *et al*. Multi-level semantic fusion network for Chinese medical named entity recognition. Journal of Biomedical Informatics, 2022, 133: 104144. [doi: 10.1016/j.jbi.2022.104144]

(校对责编: 王欣欣)