工程科学学报
Chinese Journal of Engineering

融合多特征嵌入与注意力机制的中文电子病历命名实体识别

巩敦卫 张永凯 郭一楠 王斌 樊宽鲁 火焱

Named entity recognition of Chinese electronic medical records based on multifeature embedding and attention mechanism

GONG Dun-wei, ZHANG Yong-kai, GUO Yi-nan, WANG Bin, FAN Kuan-lu, HUO Yan

引用本文:
巩敦卫, 张永凯, 郭一楠, 王斌, 樊宽鲁, 火焱. 融合多特征嵌入与注意力机制的中文电子病历命名实体识别[J]. 工程科学学报, 2021, 43(9): 1190-1196. doi: 10.13374/j.issn2095-9389.2021.01.12.006
GONG Dun-wei, ZHANG Yong-kai, GUO Yi-nan, WANG Bin, FAN Kuan-lu, HUO Yan. Named entity recognition of Chinese electronic medical records based on multifeature embedding and attention mechanism[J]. *Chinese Journal of Engineering*, 2021, 43(9): 1190-1196. doi: 10.13374/j.issn2095-9389.2021.01.12.006

在线阅读 View online: https://doi.org/10.13374/j.issn2095-9389.2021.01.12.006

工程科学学报, 第 43 卷, 第 9 期: 1190−1196, 2021 年 9 月
Chinese Journal of Engineering, Vol. 43, No. 9: 1190−1196, September 2021
https://doi.org/10.13374/j.issn2095-9389.2021.01.12.006; http://cje.ustb.edu.cn

融合多特征嵌入与注意力机制的中文电子病历命名实体识别

巩敦卫[1,2)]，张永凯[1,2)]，郭一楠[1,2)]✉，王　斌[1,2)]，樊宽鲁[3)]，火　焱[4)]

1) 中国矿业大学信息与控制工程学院, 徐州 221116　2) 中国矿业大学人工智能研究院智慧医疗研究中心, 徐州 221116　3) 徐州医科大学第二附属医院内分泌科, 徐州 221000　4) 中国矿业大学附属医院内分泌科, 徐州 221116

✉通信作者, E-mail: nanfly@126.com

摘　要　中文电子病历文本包含大量嵌套实体、句子语法结构复杂、句式偏短. 为有效识别其医疗实体, 提出一种融合多特征嵌入与注意力机制的命名实体识别算法, 在输入表示层融合字符、单词、字形三个粒度的特征, 并在双向长短期记忆网络的隐含层引入注意力机制, 使算法在捕获特征时更加关注于医疗实体相关的字符, 最终实现对中文电子病历中疾病、身体部位、症状、药物、操作五类实体的最优标注. 面向开源和自建糖尿病数据集的实验结果中所提算法的实体识别准确率、召回率和 F1 值都达到 97% 以上, 表明其可以更加有效地识别中文电子病历中各类实体.

关键词　中文; 电子病历; 命名实体识别; 多特征嵌入; 注意力机制

分类号　TP391.1

Named entity recognition of Chinese electronic medical records based on multifeature embedding and attention mechanism

GONG Dun-wei[1,2)], *ZHANG Yong-kai*[1,2)], *GUO Yi-nan*[1,2)]✉, *WANG Bin*[1,2)], *FAN Kuan-lu*[3)], *HUO Yan*[4)]

1) School of Information and Control Engineering, China University of Mining and Technology, Xuzhou 221116, China

2) Intelligent Medical Center, Institute of Artificial Intelligence, China University of Mining and Technology, Xuzhou 221116, China

3) Department of Endocrinology, the Second Affiliated Hospital of Xuzhou Medical University, Xuzhou 221000, China

4) Department of Endocrinology, Affiliated Hospital of China University of Mining and Technology, Xuzhou 221116, China

✉ Corresponding author, E-mail: nanfly@126.com

ABSTRACT　Medical records, as an essential part of the health care records of residents, save all the information about the clinical treatment of patients, which are traditionally written by doctors on paper. With the development of information technologies, electronic medical records that are more easily saved and managed gradually replace the traditional ones. Intelligent auxiliary diagnosis, patients' portrait construction, and disease prediction based on medical reports have become research hotspots in the field of intelligent medical care. To fully discover the hidden relationship between symptoms and diseases from the documents saved in electronic medical records, the development of an efficient named entity recognition algorithm is the key issue. Although several studies have been conducted on it, there is relatively little research on the information extraction of Chinese electronic medical records. To the best of our knowledge, the documents in Chinese electronic medical records contain a large number of nested named entities and short sentences. Moreover, there is weak logic among the sentences, causing a complex syntax structure. To effectively recognize the medical entities, a novel named entity recognition method based on multifeature embedding and attention mechanism was proposed. After embedding three types of features

收稿日期: 2021−01−12

基金项目: 国家自然科学基金资助项目(61973305, 61773384); 中国矿业大学中央高校基本科研业务费专项资金资助项目(2020ZDPY0302)

derived from characters, words, and glyphs in the input presentation layer, an attention machine was introduced to the hidden layer of the bidirectional long short-term memory network to make the model focus on the characters related to the medical entities. Finally, the optimal labels for the five types of entities in Chinese electronic medical records, including diseases, body parts, symptoms, drugs, and operations, were obtained. The experimental results for the open and self-built Chinese electronic medical records, recognition accuracy, recall rate, and F1 value of the proposed algorithm are all better than 97%, which shows that the proposed algorithm can effectively identify various entities in Chinese electronic medical records.

KEY WORDS Chinese; electronic medical records; named entity recognition; multifeature embedding; attention mechanism

电子病历主要用来记录患者过往病史、所患疾病及症状表现、体征检查数据、诊疗意见及治疗效果等一系列与患者健康状况相关的重要信息[1]. 随着医疗行业的信息化建设推进，用于记录患者临床治疗信息的电子病历也逐步完善. 基于电子病历的智能诊疗、患者画像构建及其病程追踪也逐渐成为智慧医疗领域的热点问题[2−3]. 为充分挖掘患者诊疗数据中的隐含特征和病症关联关系，高效准确的命名实体识别(Named entity recognition, NER)是电子病历文本信息抽取的关键. 虽然电子病历的命名实体识别已有较丰富的研究成果，但是面向中文电子病历的相关研究相对较少. 特别是，复杂的中文语言结构使中文电子病历文本存在专用词汇多、语言结构不规范、实体嵌套严重、中文词语边界模糊等特点，传统的命名实体识别模型难于获得满意的分类效果[4].

针对生物医学文本，研究人员先后提出词典与规则的统计学方法，基于 Transformer 编码的命名实体识别模型[5]、长短期记忆网络模型条件随机场[6](Bidirectional long short-term memory with conditional random field, BiLSTM-CRF)和用于解决疾病名称和实体标记不一致问题的 Dic-Att-BiLSTM-CRF 模型[7] 等. 然而，上述方法没有充分考虑中文电子病历的文本特性. 为进一步提高中文电子病历文本的命名实体识别准确性，本文提出一种融合多特征嵌入与注意力机制的中文电子病历命名实体识别模型(Multi-feature embedding-BiLSTM-Attention-CRF, MFBAC). 该模型针对中文电子病历的文本特点，首先将单词特征引入 NER 模型的输入表示层，以充分利用众多的专用医学词汇；其次，针对中文电子病历的稀疏标注数据和弱句子逻辑，采用 Glove 预训练与词典匹配，通过字词 Lattice 结构实现字词联合嵌入，从而解决传统字符向量表达中文电子病历文本时存在的局限性；再次，结合汉字字形的语义信息，通过卷积神经网络(Convolutional neural networks, CNN)提取字形局部特征，与上述字词特征充分融合，最终实现输入表示层的多特征嵌入；最后，在中间编码层，通过 BiLSTM 实现字符在句子中的语义特征提取，并采用注意力机制实现不同隐含层的权重分配，使语义特征提取聚焦在与实体词汇相关的特征上；进而，通过 CRF 解码，获得全局序列最优标签.

1 电子病历的命名实体识别方法概述

命名实体识别用于抽取序列文本中具有特定意义的实体，比如：人名、地名、时间等，并将其归类到预定义的类别中[8]，已被广泛应用于金融、互联网搜索、智慧医疗等领域[9−10]. 目前，命名实体识别方法主要有：

(1)基于词典匹配的方法：该类方法需要先构建领域词典，再通过匹配算法，完成命名实体识别，从而有效提升实体识别率[11]. 面向电子病历所适用的专病词库，通常从搜狗词库和维基百科词条中获得，再通过添加同义词、缩写词加以扩充. 为强化医学词典的个性化信息，可以对电子病历文本、医学文献先进行分词处理，再提取具有较大词频−逆向文件频率的若干词加入到领域词典中. 虽然基于词典匹配的方法对词识别率较高，但是由于中文医疗实体数量众多、个性化词汇丰富、难以完整的词典，所以容易导致对新词的错误识别[12].

(2)基于规则的方法：根据生成的大量规则，利用实体的上/下文信息，完成命名实体识别. 但是，规则依赖于领域专家经验，且不同领域之间的规则可移植性差. Kraus 等[12] 通过构建大量的正则表达式，用于识别临床记录中的药品、剂量等医疗实体.

(3)基于统计机器学习的方法：常见的统计机器学习方法有支持向量机、最大熵、隐马尔可夫模型、条件随机场 (Conditional random fields, CRF)等. 这些方法不需要过多的人工干预，但依赖于大规模的标注数据集[13] 和选择的特征.

(4)基于深度学习的方法：该方法采用端到端的模型训练与自动特征提取，不需要对数据进行人工处理. 针对电子病历文本，研究人员先后提出一类 Transformer 编码模型[5]、双向长短期记忆网

络卷积条件随机场模型[14]、基于字符与字典匹配实体联合编码的 Lattice-LSTM-CRF 模型[15]、谷歌公司开源的 BERT 模型[16] 等, 也有不少学者将迁移学习、半监督学习引入 NER 任务中.

(5)混合方法: Jiang 等[17] 将启发式规则与基于机器学习的实体识别模型相融合, 设计了一种临床实体的混合识别系统. Wei 等[18] 则针对单一疾病的医疗实体识别, 在条件随机场模型中引入规则. 龚乐君和张知菲[19] 提出基于领域词典与 CRF 双层标注的电子病历实体识别方法, 调和平均值(F1 Score, F1 值)达到 97.2%. Hu 等[20] 在 2017 年全国知识图谱与语义计算大会的临床命名实体识别竞赛中, 通过构建医疗实体规则, 获得了较好的医疗实体识别效果. 通过合理集成词典、规则、统计学习、深度学习等方法, 提升 NER 模型性能和实体识别效果.

在构建 NER 模型时, 通常采用词嵌入的方法, 将词表示为向量, 实现对中文电子病历文本的编码. 词嵌入应该能充分挖掘词在上下文的语义特征. 传统的独热表示方法, 不仅具有高稀疏性, 而且无法刻画词的语义信息. 基于此, 研究人员分别提出基于全局矩阵分解和局部上下文窗口的词嵌入方法[21]. 全局矩阵分解方法虽然利用了全局语料特征, 但是求解的计算规模较大. 相比而言, 基于局部上下文窗口的连续词袋模型 (Continuous bag of words, CBOW) 和跳读模型(Skip-Gram)等方法仅利用局部文本数据进行训练, 不能有效反映词汇的全局统计信息. 为了克服全局矩阵分解和局部上下文窗口方法的局限性, Pennington 等[22] 基于全局文本信息, 提出一种融合全局矩阵分解和 Word2Vec 的 Glove 方法, 显著提升了词嵌入效果.

2　融合多特征嵌入与注意力机制的中文电子病历命名实体识别

MFBAC 算法在输入表示层引入字符、字形、单词三个粒度的特征, 并在 BiLSTM-CRF 中融入注意力机制, 兼顾局部特征, 弥补了 BiLSTM 的不足, 提升了命名实体识别效果. 如图 1 所示, 输入文本序列经 Glove 预训练, 实现字符与单词嵌入, 并通过查表操作, 依次将序列文本转换为对应向量; 通过词典匹配, 基于字词 Lattice 结构, 实现字词联合嵌入; 采用 CNN 提取字形的部首局部特征向量; 通过双向长短期记忆网络(Long short term memory, LSTM), 对拼接后的特征向量实现特征提取; 基于注意力机制实现不同隐含层权重的重新分配; 经 CRF 解码, 获得全局序列最优标签. 由此可见, 多特征嵌入层、双向 LSTM、注意力机制层、条件随机场是 MFBAC 算法的关键技术.

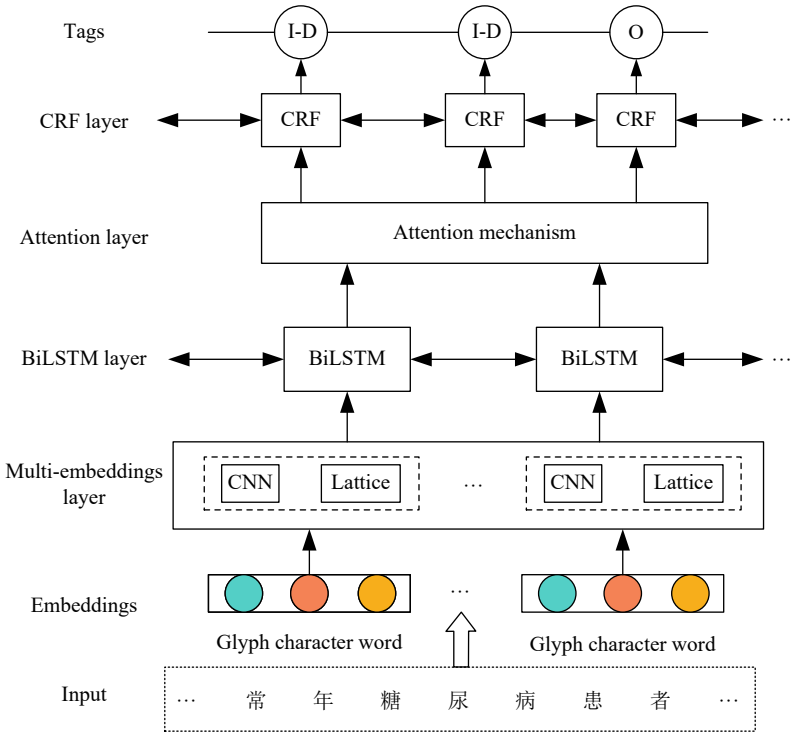

图 1　MFBAC 算法框架

Fig.1　MFBAC framework

2.1 多特征嵌入层

考虑到中文电子病历中存在的句间逻辑关系弱、实体嵌套严重、标签数据缺乏等特点，为有效提升 NER 模型性能，在输入表示层引入更多有效特征. 传统的字典匹配方法虽然可以有效解决中文电子病历文本中短句和弱逻辑关系问题，但是单一的词嵌入容易因词典中关键词的缺乏而产生识别误差. 以字符表征为主，采用 Lattice 结构的字词联合编码可以在字符特征中加入单词信息，从而避免分词误差. 基于此，本文将字符向量与领域词典匹配后的单词，通过 Lattice 结构实现字词混合编码，得到表征向量；再与字形特征向量拼接，得到序列文本的多特征嵌入.

为兼顾字符嵌入的计算代价和多特征嵌入的有效性，本文采用 Glove 实现字与词嵌入，为后续的字词联合嵌入提供基础. 基于维基百科、医疗论坛、医疗典籍上获取的大规模医疗文本，采用 Glove 模型，将包含 n 个字符的医疗文本序列转换为 100 维的字符向量. 同理，对分词处理的文本序列，通过 Glove 训练，得到 100 维的词向量. 进而，采用 Lattice 结构，将字词特征向量求均值后，得到 100 维的字词向量序列 $\boldsymbol{V}=\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_n\}$. 采用 CNN 提取汉字字形的局部特征[20]，为输入表示层引入更多有效的语义特征. 对于每个字符，采用 CNN 计算字形表示，再通过 CNN 对所有字符进行卷积与合并，获得字形级特征向量的序列 $\boldsymbol{W}=\{\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_n\}$. 进而，将 $\boldsymbol{W}$ 与上述字词特征向量 $\boldsymbol{V}$ 进行拼接，得到多特征嵌入表征 $\boldsymbol{U}$.

2.2 双向 LSTM

输入表示层的多特征向量 $\boldsymbol{U}$，经中间编码层，实现特征提取. 虽然传统的循环神经网络（Recurrent neural network, RNN）可以较好地建模词在句子中的语义且稳定性好，但是并行性弱、速度慢，在处理长序列文本时会发生梯度消失或梯度爆炸，不适合学习长距离的语义信息[23]. LSTM 在 RNN 结构中引入输入门、遗忘门和输出门，通过门控机制选择性地改变需要保留的内容，捕获长距离关联信息，有效克服了传统 RNN 模型在处理长序列文本时的梯度问题. 记 LSTM 层的输出为 $n\times k$ 维矩阵 $\boldsymbol{S}=[s_{ij}]_{n\times k}$，其中，$n$ 为输入文本序列长度，k 为标签个数.

LSTM 本质上是一种前向传播学习算法. 为更加充分的利用上下文信息，将反向 LSTM 与之组合，构成双向 LSTM，将同一输入变量的两个反向隐含层状态向量进行拼接，更高效的表示字符在上下文中的含义. 记特征为 $\boldsymbol{H}=[\overrightarrow{h}_t;\overleftarrow{h}_t]$，$\overrightarrow{h}_t$ 与 $\overleftarrow{h}_t$ 分是正反向特征向量，[·;·]是拼接符号. 基于此，获得双向 LSTM 输出的状态转移矩阵，记为 $\boldsymbol{M}$.

2.3 注意力机制层

传统的嵌入表示方法没有考虑字符之间的相关性，导致输入序列中的信息不能充分利用. 为此，引入注意力机制，深层提取词汇特征和语义信息，从而对中文电子病历文本中与实体相关的字符加以自动关注，忽略无用信息，兼顾长文本序列的局部特征.

注意力机制源于对人类视觉的研究，已在机器翻译、机器视觉等领域得到应用广泛[24]. 考虑到中文电子病历的短句式和句间弱逻辑特性，在中间隐含层引入注意力机制，使不符合注意力模型的内容被弱化或者遗忘[25]. 针对每个字符，注意力机制使所提模型聚焦于输入序列的其他信息，通过获取更多线索，更好地对该字符进行编码. 注意力机制对不同字符的隐含层状态赋予不同的关注权重，从而使语义特征可以集中在与医疗实体相关的字符上[26].

记双向 LSTM 输出向量为 $\boldsymbol{H}_i=[\overrightarrow{h}_i\oplus\overleftarrow{h}_i]$，第 i 个字符在句子中的重要性量化为能量函数 e_i，式中 b 为偏置分量.

$$e_i=\tanh(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{H}_i+b) \tag{1}$$

对 e_i 进行归一化处理结果，记为

$$\alpha_i=\frac{\exp(e_i)}{\sum_i\exp(e_i)} \tag{2}$$

由此，基于动态尺度计算注意力权重为

$$\boldsymbol{H}_i'=\boldsymbol{H}_i\cdot\alpha_i \tag{3}$$

采用注意力权重分配方法来改变双向 LSTM 输出的概率矩阵，可以兼顾更多局部特征，也就能改善 CRF 层的序列标注结果.

2.4 条件随机场

CRF 解码过程中，将重新分配权重后的双向 LSTM 概率矩阵输入，获得序列标签. 记句子序列为 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\cdots,\boldsymbol{x}_n\}$，其预测的标签序列为 $\boldsymbol{Y}=\{\boldsymbol{y}_1,\boldsymbol{y}_2,\boldsymbol{y}_3,\cdots,\boldsymbol{y}_n\}$，则得分概率计算如下：

$$\boldsymbol{p}(\boldsymbol{X},\boldsymbol{Y})=\sum_{i=1}^{n}\boldsymbol{M}'_{y_i,y_{i+1}}+\sum_{i=1}^{n}\boldsymbol{N}_{i,y_i} \tag{4}$$

其中，$\boldsymbol{M}'_{y_i,y_{i+1}}$ 表示从 $\boldsymbol{y}_i$ 转移到 $\boldsymbol{y}_{i+1}$ 的概率，$\boldsymbol{N}_{i,y_i}$ 表示第 i 个词语被标记为 $\boldsymbol{y}_i$ 的概率，$\boldsymbol{p}(\boldsymbol{X},\boldsymbol{Y})$ 表示输入句子序列 $\boldsymbol{X}$ 被标记标签序列为 $\boldsymbol{Y}$ 的概率. 当前样本 $\boldsymbol{X}$ 的最佳标签序列具有最大 $\boldsymbol{p}(\boldsymbol{X},\boldsymbol{Y})$ 值.

3　实验结果与分析

为有效验证所提算法的有效性, 本文以 CCKS2017 开源数据集和自建糖尿病中文电子病历集作为实验数据来源, 通过与主流模型的多组对比实验, 深入分析所提命名实体识别方法性能. 所有实验是在 intel Xeon Silver 4210@2.20 GHz 处理器、256 GB 内存、Nvidia Quadro P5000 计算条件下进行, 模型搭建采用开源框架 TensorFlow 1.14 实现.

3.1　实验数据集和参数设置

自建糖尿病中文电子病历集共 500 份, 参考 CCKS2017 开源数据集, 采用 BIO 标注体系统一进行标注. BIO 标注体系中, B 代表实体的开始位置, I 代表实体的内部, O 代表非实体部分. 相应的命名实体包含疾病、症状、身体部位、药品、检查操作五种类别, 如表 1 所示.

表 1　命名实体类别

Table 1　Types of named entities

The entity class	Identifier	Definition of categories
Diseases	B-diseases I-diseases	Terms of various diseases
Symptom	B-symptom I-symptom	Abnormal physical manifestations
Body	B-body I-body	Various parts of the human body
Drug	B-drug I-drug	The names of various medicines
Test	B-test I-test	Various physical examinations

从 CCKS2017 开源数据集和自建数据集中, 随机选取 800 份糖尿病中文电子病历构成数据集, 其中, 80% 作为训练集, 20% 作为测试集. 训练集和测试集中各类实体分布, 如表 2 所示.

表 2　训练集与测试集医疗实体分布

Table 2　Distribution of training and test datasets for medical entities

Dataset	Training data	Test data
Diseases	856	382
Symptom	3845	1526
Body	563	214
Drug	657	289
Test	3426	1647
Total	9347	4058

为实现多特征嵌入, 从维基百科与医疗论坛上爬取 1000 万条句子作为训练语料, 词向量和字符维度设置为 100, 窗口尺寸设为 8; 用于字形局部特征提取的 CNN 采用 13 层, 选用 TrueType 字体将每个汉字渲染为 48×48 的 8 位灰度位图; 卷积核设为 3×3, 通道数选取 64、128、256 和 512; BiLSTM-Att-CRF 的隐含层节点数选取 300, dropout 层参数设置为 0.5, Adam 优化算法的学习率为 0.001, batch size 设为 64, epoch 选取 80. 采用准确率 P、召回率 R 和 F1 作为 NER 模型评价指标[25].

3.2　实验结果对比分析

实验一: 基于相同的 BiLSTM-CRF 结构, 在输入表示层分别采用字符嵌入 (Char embedding)、字形嵌入 (Font embedding)、字词联合嵌入 (CW Embedding)、字词形嵌入 (CWF embedding), 对比分析所提算法中多特征嵌入的合理性与必要性. 不同特征嵌入下的命名实体识别性能均值如表 3 所示. 可见, 单独使用字符嵌入优于单独字形嵌入; 将字符与单词特征通过 Lattice 结构联合嵌入, 其效果优于单纯使用字符嵌入或词嵌入, 表明单词特征可以显著表达中文电子病历文本中的实体关系. 相比而言, 所提字词形多特征嵌入比字词联合嵌入具有更优识别性能, 并且随着引入特征的增加, 命名实体识别性能显著改善, 表明引入的多类特征符合中文电子病历特点, 可以有效增强模型性能.

表 3　不同特征嵌入下的命名实体识别性能

Table 3　Performance of NER embedding different features

Model	P/%	R/%	F1/%
Font embedding-BiLSTM-CRF	79.51	80.35	79.72
Char embedding-BiLSTM-CRF	88.61	87.43	87.96
Word embedding-BiLSTM-CRF	85.82	86.87	86.32
CW embedding-BiLSTM-CRF	86.58	87.23	87.62
CWF embedding-BiLSTM-CRF	96.24	97.25	96.94

实验二: 基于实验一的五种输入表示层嵌入方式, 在 BiLSTM-CRF 结构中引入注意力机制, 构成 BiLSTM-Att-CRF, 其识别性能如表 4 所示. 通过与实验一的命名实体识别性能对比, 表明引入注意力机制可以显著提升 NER 模型性能. 这是因为, 注意力机制通过重新计算各个隐含层的权重分布, 弥补了双向 LSTM 对多特征嵌入信息的提取不足, 通过更加合理的编码层特征提取, 有效改善模型的实体识别性能.

实验三: 为充分验证所提模型的合理性, 将其与其他主流算法进行性能对比. 其他主流算法的输入表示层均采用字符嵌入. 由表 5 所示的不同算法识别效果可知, 本文所提 MFBAC 方法的识别性能均优于其他主流算法, 表明引入多特征嵌入和注意力机制对识别语言结构特殊的中文电子病

表 4 注意力机制对不同特征嵌入的影响

Table 4 Performance of NER with attention

Model	*P*/%	*R*/%	F1/%
Font embedding-BiLSTM-Att-CRF	92.46	93.12	92.68
Char embedding-BiLSTM-Att-CRF	93.41	93.56	93.49
Word embedding-BiLSTM-Att-CRF	96.36	96.18	96.21
CW embedding -BiLSTM-Att-CRF	96.52	96.18	96.45
CWF embedding -BiLSTM-Att-CRF	97.21	97.83	97.54

历实体是有效且必要的. 此外, 通过对比图 2 所示的变换器条件随机场(Transformer-CRF)、BiLSTM-CRF、MFBAC 算法的 F1 值可知, 本文所提算法对 5 类医疗实体的识别率均取得最好性能. 字符和单词的联合编码可以为与身体部位相关的实体提供丰富信息, 确保该类实体的识别性能; 对于存在较多嵌套的检查和操作类实体, 其他算法不能提取足够有效的特征, 而 MFBAC 算法通过引入注意力机制解决了该问题. 为进一步分析所提算法的识别效率, 统计不同识别算法的平均加载时间和平均测试时间, 如表 5 所示. 可见, 本文所提算法的模型加载速度和测试时间虽稍劣于基于注意力机制的双向记忆神经网络与条件随机场(Attention-BiLSTM-CRF)、双向门控循环神经网络与条件随机场(BiGRU-CRF)和 BiLSTM-CRF, 但是其识别性能却显著优于后者.

表 5 不同算法的性能对比

Table 5 Comparison of the performance of different NER models

Model	*P*/%	*R*/%	F1/%	Loading time/s	Testing time/s
Transformer	85.46	86.32	85.68	4.33	12.6
BiGRU-CRF	85.87	86.23	86.14	2.95	9.4
BiLSTM-CRF	88.61	87.43	95.16	3.21	9.81
Attention-BiLSTM-CRF	94.52	96.18	96.45	3.56	10.56
Transformer-CRF	95.32	94.62	94.14	5.32	13.57
MFBAC	97.21	97.83	97.54	4.34	11.68

4 结论

针对中文电子病历文本, 提出一种融合多特征嵌入与注意力机制的命名实体识别算法. 该算法根据中文电子病历的特点, 在输入表示层集成了字词和字符形状等多种特征嵌入, 并通过引入注意力机制, 对双向 LSTM 各个隐含层的编码信息进行权值分配. 针对中文电子病历文本的 5 类实体, 基于开源和自建糖尿病数据集, 通过三组实验的对比分析, 表明输入表示层引入多种特征是必要的, 在双向 LSTM 与 CRF 层中间引入注意力机制, 能有效兼顾序列文本的局部特征, 显著提升所提 NER 算法性能, 同时不显著增加计算代价. 基于该思想, 在未来工作中将深入挖掘不同特征的重要性, 并引入半监督学习和迁移学习方法, 解决中文电子病历中标签样本少等问题.

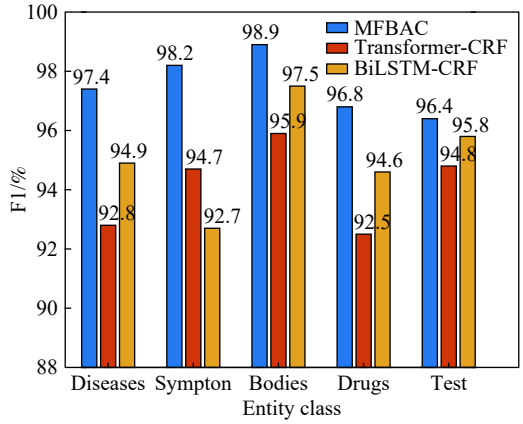

图 2 不同算法的 F1 值

Fig.2 Comparison on the F1 values of different NER models

参 考 文 献

[1] Tang G Q, Gao D Q, Ruan T, et al. Clinical electronic medical record named entity recognition incorporating language model. *Comput Sci*, 2020, 47(3): 211
(唐国强, 高大启, 阮彤, 等. 融入语言模型和注意力机制的临床电子病历命名实体识别. 计算机科学, 2020, 47(3): 211)

[2] Topol E J. High-performance medicine: The convergence of human and artificial intelligence. *Nat Med*, 2019, 25(1): 44

[3] He J, Baxter S L, Xu J, et al. The practical implementation of artificial intelligence technologies in medicine. *Nat Med*, 2019, 25(1): 30

[4] Li B, Kang X D, Zhang H L, et al. Named entity recognition in Chinese electronic medical records using transformer-CRF. *Comput Eng Appl*, 2020, 56(5): 153
(李博, 康晓东, 张华丽, 等. 采用Transformer-CRF的中文电子病历命名实体识别. 计算机工程与应用, 2020, 56(5): 153)

[5] Luo L, Yang Z H, Yang P, et al. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics*, 2018, 34(8): 1381

[6] Xu K, Yang Z G, Kang P P, et al. Document-level attention-based BiLSTM-CRF incorporating disease dictionary for disease named entity recognition. *Comput Biol Med*, 2019, 108: 122

[7] Yang J F, Yu Q B, Guan Y, et al. An overview of research on electronic medical record oriented named entity recognition and entity relation extraction. *Acta Autom Sin*, 2014, 40(8): 1537
(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. 自动化学报, 2014, 40(8): 1537)

[8] Lei J, Tang B, Lu X, et al. A comprehensive study of named entity recognition in Chinese clinical text. *J Am Med Inform Assoc*, 2014, 21(5): 808

[9] Hirschberg J, Manning C D. Advances in natural language processing. *Science*, 2015, 349(6245): 261

[10] Wang Q, Zhou Y M, Ruan T, et al. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition. *J Biomed Informatics*, 2019, 92: 103133

[11] Shang J B, Liu L Y, Gu X T, et al. Learning named entity tagger using domain-specific dictionary//*Proceedings of the* 2018 *Conference on Empirical Methods in Natural Language Processing*. Brussels, 2018: 2054

[12] Kraus S, Blake C, West S L. Information extraction from medical notes [J/OL]. *arXiv preprint* (2007-07-24) [2020-12-26]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.120.3671&rep=rep1&type=pdf

[13] Gorinski P J, Wu H H, Grover C, et al. Named entity recognition for electronic health records: A comparison of rule-based and machine learning approaches [J/OL]. *arXiv preprint* (2019-04-25) [2020-12-26]. https://arxiv.org/pdf/1903.03985.pdf

[14] Ma X Z, Hovy E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF [J/OL]. *arXiv preprint* (2016-05-29) [2020-12-26]. https://arxiv.org/pdf/1603.01354.pdf

[15] Zhang Y, Yang J. Chinese NER Using Lattice LSTM [J/OL]. *arXiv preprint* (2018-07-05) [2020-12-26]. https://arxiv.org/pdf/1805.02023.pdf

[16] Alsentzer E, Murphy J R, Boag W, et al. Publicly available clinical BERT embeddings [J/OL]. *arXiv preprint* (2019-6-20) [2020-12-26]. https://arxiv.org/pdf/1904.03323.pdf

[17] Jiang M, Chen Y K, Liu M, et al. A study of machine-learning-based approaches to extract clinical entities and their assertions from discharge summaries. *J Am Med Inform Assoc*, 2011, 18(5): 601

[18] Wei Q K, Chen T, Xu R F, et al. Disease named entity recognition by combining conditional random fields and bidirectional recurrent neural networks. *Database* (*Oxford*), 2016, 140: 1

[19] Gong L J, Zhang Z F. Clinical named entity recognition from Chinese electronic medical records using a double-layer annotation model combining a domain dictionary with CRF. *Chin J Eng*, 2020, 42(4): 469
(龚乐君, 张知菲. 基于领域词典与CRF双层标注的中文电子病历实体识别. 工程科学学报, 2020, 42(4): 469)

[20] Hu J L, Shi X, Liu Z J, et al.HITSZ_CNER: a hybrid system for entity recognition from Chinese clinical text//*Proceedings of the Evaluation Tasks at the China Conference on Knowledge Graph and Semantic Computing* (*CCKS* 2017). Chengdu, 2017: 1

[21] Mikolov T, Grave E, Bojanowski P, et al. Advances in pre-training distributed word representations [J/OL]. *arXiv preprint* (2017-12-26) [2020-12-26]. https://arxiv.org/pdf/1712.09405.pdf

[22] Pennington J, Socher R, Manning C. GloVe: global vectors for word representation//*Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*). Doha, 2014: 1532

[23] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need [J/OL]. *arXiv preprint* (2017-12-06) [2020-12-26]. https://arxiv.org/pdf/1706.03762.pdf

[24] Choi E, Bahadori M T, Kulas J A, et al. RETAIN: interpretable predictive model in healthcare using reverse time attention mechanism [J/OL]. *arXiv preprint* (2016-08-19) [2020-12-26]. https://arxiv.org/pdf/1608.05745.pdf

[25] Zhu Q L, Li X L, Conesa A, et al. GRAM-CNN: a deep learning approach with local context for named entity recognition in biomedical text. *Bioinformatics*, 2018, 34(9): 1547

[26] Wu G H, Tang G G, Wang Z R, et al. An attention-based BiLSTM-CRF model for Chinese clinic named entity recognition. *IEEE Access*, 2019, 7: 113942