DOI:10.3969/j.issn.1005-202X.2024.07.016

医学生物信息

基于特征融合的糖尿病命名实体识别

任建华,赵若涵
辽宁工程技术大学电子与信息工程学院,辽宁 葫芦岛 125105

【摘要】针对医学糖尿病领域命名实体识别中存在实体种类多样性、数据稀缺等问题,提出了基于特征融合的糖尿病命名实体识别方法。以BERT+BILSTM+CRF为基准模型,在3方面进行改进。首先,使用预训练模型RoBERTa-wwm-ext作为模型嵌入层,提供字符级嵌入,利用其在训练阶段进行全词掩码来获取含有先验知识的语义表示。其次,使用双向长短时记忆网络和迭代膨胀卷积神经网络并行提取特征,以获取不同粒度的特征。同时,结合注意力机制进行动态特征融合,从而更好地理解数据的关键信息,以获得更丰富的上下文特征。最后,采用条件随机场进行解码,获得最终的预测结果。该模型在包含18种实体类别的中文糖尿病数据集DiaKG上的F1值达到了79.58%,实验结果表明,与High-Order MKGragh模型相比,该模型的F1值提升了5.38%,充分说明了特征融合的方法能够有效识别糖尿病实体。

【关键词】糖尿病;命名实体识别;特征融合;注意力机制
【中图分类号】R318;R587.1 【文献标志码】A 【文章编号】1005-202X(2024)07-0890-07

Diabetes named entity recognition based on feature fusion

REN Jianhua, ZHAO Ruohan
School of Electronic and Information Engineering, Liaoning Technical University, Huludao 125105, China

Abstract: To overcome the challenges of entity diversity and data scarcity in diabetes named entity recognition, feature fusion based named entity recognition is proposed. With BERT+BILSTM+CRF as the benchmark model, improvements are made in 3 aspects. (1) The pre-trained model RoBERTa-wwm-ext is introduced as the model embedding layer to provide character-level embedding, and the whole word mask is used in the training stage to obtain semantic representation containing prior knowledge. (2) bidirectional long short-term memory network and iterated dilated convolutional neural network are used to extract features in parallel to obtain features of different granularities. At the same time, the dynamic feature fusion is combined with the attention mechanism to better understand the key information of the data, thus obtaining richer contextual features. (3) The conditional random field is decoded to obtain the final prediction results. The proposed model achieves an F1 value of 79.58% which is 5.38% higher than High-Order MKGraph on DiaKG, a Chinese diabetes data set containing 18 entity categories, fully demonstrating that the feature fusion based method can effectively identify diabetic entities.

Keywords: diabetes; named entity recognition; feature fusion; attention mechanism

前 言

命名实体识别是自然语言处理(Natural Language Processing, NLP)中的一项重要任务,其主要功能是从海量非结构化文本中识别和提取各种命名实体,如人名、地名、组织名称、领域特定词等[1]。命名实体识别任务是信息抽取[2]、智能问答[3]、机器翻译[4]、情感分析[5],知识图谱[6]等众多自然语言处理任务的基础。

在医学领域需要处理大量的文本信息来支持糖尿病的诊断和治疗;因此,糖尿病的命名实体识别应该引起广泛关注。糖尿病是21世纪发展最快的疾病之一[7],中国已成为世界上糖尿病患者人数最多的国家,成人患糖尿病的概率为11.7%,且这一概率还在上升[8]。然而,我国对糖尿病的相关知识关注度较低,所以如何迅速从众多文献中获取专业知识成为笔者的研究重点。

在早期的医学命名实体识别任务中,大多数采用基于规则和基于字典的方法[9],需要由专业的医学专家制定,导致劳动力成本高,便携性低。与基于规

【收稿日期】2024-02-20
【基金项目】国家自然科学基金(61772249)
【作者简介】任建华,硕士,副教授,研究方向:智能数据处理、数据库理论及应用,E-mail: renjh4665@163.com;赵若涵,硕士研究生,主要研究方向:自然语言处理,E-mail: 2689613378@qq.com

则和基于词典的实体识别方法相比，基于统计学的机器学习算法利用手动注释语料库进行监督训练，其准确性显著提高[10]。随着深度学习的出现和计算能力的提高，越来越多的学者利用神经网络模型来处理医学命名实体识别任务[11]，并将其视为序列标记问题。Zhang等[12]提出了一种基于长短时记忆网络（Long Short-Term Memory, LSTM）的改进模型Lattice LSTM，将潜在的单词信息纳入传统的基于单词的LSTM模型中，避免了分词错误引起的错误传播，但是容易造成信息损失。随着深度学习方法的不断优化，识别结果精度也在逐步提升，但上述方法的效果依旧受训练语料规模和质量的影响。直到2018年Devlin等[13]引入了一种双向编码器表征法（Bidirectional Encoder Representations from Transformers, BERT）预训练模型，在自然语言处理领域取得了重大突破，该模型提高了嵌入词的质量，从而提高了识别性能，但BERT存在参数多模型大，少量数据训练时，容易出现过拟合等问题，所以又有对BERT进行改进的ALBERT、RoBERTa等预训练模型的研究出现，并已被应用到医学领域的命名实体识别任务中。Liu等[14]设计了一个Med-BERT预训练框架，该框架结合了医学语料库和与该领域相关的特定任务，以提高模型在医学命名实体识别中的性能。文献[15]将ALBERT与双向长短时记忆网络（Bi-directional Long Short-Term Memory, BILSTM）结合用于识别糖尿病命名实体，解决了BERT模型参数量大，训练速度过慢等问题。文献[16]使用RoBERTa预训练模型与对抗训练相结合用于医疗电子病历命名实体识别，解决了BERT只能获取字信息以及模型鲁棒性差的问题。除此之外，还有一些联合抽取的方法。Wei等[17]为解决重叠三元组问题，提出了一种新的用于关系三重抽取的级联二值标记框架，但这个框架需要将三元组抽取任务拆分为两个阶段，增加了模型的复杂性和计算成本。Yan等[18]提出的一种用于联合实体和关系提取的分区过滤网络，它将编码器分割为实体提取和关系提取部分，避免了手工设计特征和规则的繁琐过程，简化了模型的搭建和调整，但存在分割阈值难以确定的问题。Lai等[19]提出的基于知识增强集体推理的实体和关系提取，利用外部知识进行联合实体和关系提取新框架。但该方法的性能受限于外部知识库的规模和覆盖范围，如果外部知识库中的实体和关系数量有限或者与任务领域不匹配，可能会影响到抽取结果的准确性。Su等[20]提出了一种新的基于片段的NER框架-Global Pointer（GP），核心思想是能够从全局的视角考虑实体的起始和终止位置，并设计了两个模块来识别实体的头尾位置，保障训练和推理过程的一致性。由于该方法依赖于局部和全局上下文的信息来识别实体的边界，所以对于一些稀有实体，可能由于训练数据中缺乏足够的例子而导致识别准确性较低。Yang等[21]提出利用高阶医学知识图谱进行关节实体和关系提取，提出在医学知识图谱的基础上构建一个高阶异构图谱，这样，来自高阶异构图的邻居可以相互传递消息，以获得更好的全局上下文表示。但如果知识图谱中缺乏与任务相关的实体和关系，就可能会影响到该方法的抽取效果。特别是在使用训练数据较少或者不平衡的情况下，对缺失数据的敏感性可能更加明显。

然而，前面的研究和改进的模型都忽视了实际的糖尿病命名实体识别任务中存在实体种类众多，且有些实体数量稀缺导致模型识别效果差，准确率低，甚至预测不出数据稀缺的实体等问题。为此，本文在BERT+BILSTM+CRF上改进，提出了一种基于特征融合的DNER模型，首先，使用RoBERTa-wwm-ext代替BERT对字符进行编码以更好地适应和处理中文糖尿病字符；其次，目前大多数的模型都是直接将生成的字符向量输入到BILSTM和迭代膨胀卷积神经网络（IDCNN）中，而不综合考虑整体特征和局部最优特征，不能很好的理解命名实体的上下文背景和语义关系，使得实体识别的准确率没有得到很大提升，所以本文提出使用BILSTM和IDCNN并行提取特征，从不同的角度提取输入序列的特征，提高特征提取的效果、模型的鲁棒性和特征提取速度；传统的特征抽取层通常使用卷积神经网络或循环神经网络等方法来提取文本中的特征信息，然而，这些方法可能无法充分利用不同位置的特征之间的关系，为了解决这个问题，笔者引入了注意力机制对两种特征进行动态融合，从而提高模型的性能，更好地解释模型预测结果，也可以更好地适应不同的任务和数据集，提高模型的泛化能力；最后，使用条件随机场（Conditional Random Field, CRF）进行解码，得到最终的标签序列和预测结果。

1 基于特征融合的糖尿病命名实体识别方法

1.1 模型架构

本文提出的模型体系架构如图1所示，包括输入层、预训练模型的嵌入层、双通道特征提取层、注意力机制特征融合层、CRF解码层和输出层。首先采用RoBERTa-wwm-ext预训练模型从文本数据中提取词向量，将文本转换为语义表示，以便后续特征提取和融合；其次，采用BILSTM和IDCNN两种网络结构，并行提取输入序列的特征，这样可以从不同角度

获取丰富的特征表示，提高模型的性能和鲁棒性；双通道特征提取层之后，使用注意力机制对BILSTM和IDCNN提取的特征进行融合，从而更好地结合两种特征，这样可以进一步提高模型的性能，解释预测结果，并适应不同的任务和数据集；最后，输入到CRF层中进行解码，从而得到字符的标签序列。

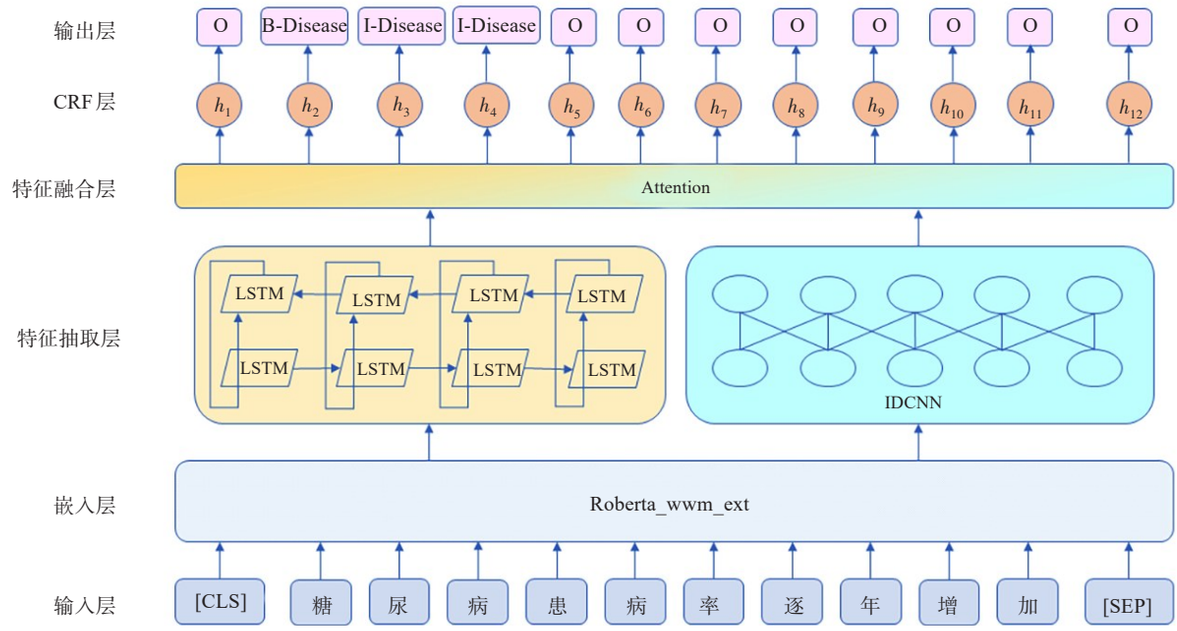


图1 模型架构图

Figure 1 Model architecture

1.2 嵌入层

BERT是一种以Transformer为主要架构的无监督深度双向语言表示模型，可以提高多种自然语言处理任务的性能。BERT已被广泛用于命名实体识别任务，用于预训练的语义表示。但BERT在中文中是以字为粒度切分的，没有考虑到中文分词，在处理特定任务时可能无法充分利用句子中的语义信息。因此，本文提出使用RoBERTa-wwm-ext预训练模型代替BERT模型对文本进行编码[22]。

RoBERTa-wwm-ext预训练模型是哈工大与科大讯飞研究联合实验室在RoBERTa和中文全词掩码技术的基础上推出的中文预训练语言模型[23]。RoBERTa同BERT一样也是由堆叠的Transformer架构组成，并在海量文本数据上训练得到的。在模型架构层面，RoBERTa与BERT基本一致，如图2所示。RoBERTa使用动态掩码而不是BERT的静态掩码。动态掩码每次会随机选择不同的单词进行掩码，增加了模型输入的随机性，让模型学习到更多样化的语言表示。

RoBERTa-wwm-ext采用整个单词掩码，而不是BERT的单个字符掩码。整个单词掩码将遮罩整个单词而不是单个字符，这有助于提高模型对词汇的

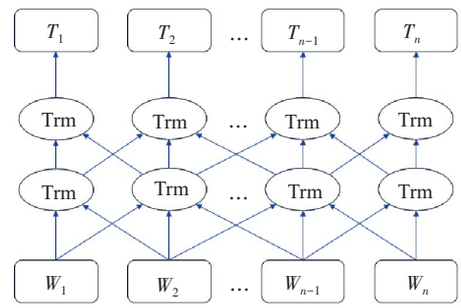


图2 BERT架构图

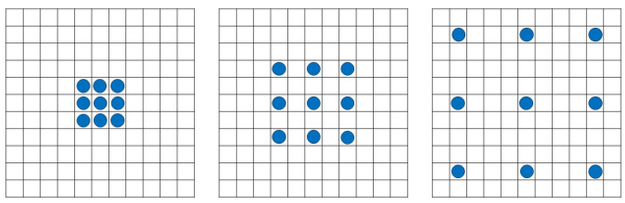
Figure 2 BERT architecture

理解。RoBERTa-wwm-ext先将一个句子分成多个单词，然后随机屏蔽一部分单词进行预测，通过这种训练方法，RoBERTa-wwm-ext可以在单词级别上学习语义表示，从而达到提高模型性能的整体效果。

1.3 特征抽取层

在嵌入层，本文使用RoBERTa-wwm-ext预训练模型从原始文本中提取出词向量。这些词向量具有丰富的语义信息。为了更好地利用这些向量，本文同时采用了BILSTM和IDCNN两种网络结构进行不同粒度的特征提取。BILSTM是一种循环神经网络，它能够捕捉输入序列的上下文信息，并通过前向和

后向传递来提取特征。它可以有效地处理序列数据，对于命名实体识别任务非常适用。IDCNN是一种卷积神经网络[24]，它通过多次迭代和膨胀卷积操作来提取局部最优特征。它能够自动学习到输入序列中的重要特征，并逐渐扩大感受野。IDCNN是在原卷积的基础上，增加了一个扩展步骤，卷积运算可以跳过步骤中间的数据，保持卷积核的大小不变，卷积核可以得到更大的输入矩阵，增加接收域。如图3所示，在图3a中，步长设置为1向外扩散，形成一个3×3区域的接收域；在图3b中，步长设置为2向外扩散，形成7×7区域的接收域；在图3c中，步长设置为4向外扩散，形成15×15接收域。因此，它可以扩展卷积核的感知视界，获取多尺度信息，有利于神经网络获取推文的上下文信息，提高模型的性能。总之，同时使用BILSTM和IDCNN，可以从不同的粒度上提取特征。这种多粒度的特征提取方法可以更好地理解文本的语义和上下文关系，从而提高了模型在糖尿病命名实体识别任务中的表现。



a：膨胀步长为1　　b：膨胀步长为2　　c：膨胀步长为4

图3 迭代膨胀卷积网络

Figure 3 Iterated dilated convolutional neural network

1.4 特征融合层

BILSTM通过在每个时间步上同时运行前向和后向的LSTM，能够有效地捕捉到序列数据中的上下文信息。这使得BILSTM能够捕捉到较长距离的依赖关系，更好地理解整个序列的上下文，可以捕捉全局特征。相比之下，IDCNN采用可扩展的卷积神经网络结构，通过堆叠多个卷积层来捕捉不同尺度的特征，使得它能够更加关注局部信息和细节。因此，本文通过在特征抽取层添加注意力机制，对得到的特征进行特征融合，实现优势互补。特征融合的实现方式如下：

$$\boldsymbol{t}=\sigma\left(\boldsymbol{W}_t^3\tanh\left(W_t^1\boldsymbol{x}_a+\boldsymbol{W}_t^2x_b\right)\right) \tag{1}$$

$$\tilde{x}=t\cdot\boldsymbol{x}_a+(1-t)\cdot\boldsymbol{x}_b \tag{2}$$

其中，σ为Sigmoid激活函数；$\boldsymbol{W}_t$为可学习的权重矩阵；$\boldsymbol{x}_a$为BILSTM输出的向量；$\boldsymbol{x}_b$为IDCNN输出的向量。向量$\boldsymbol{t}$与$\boldsymbol{x}_a$和$\boldsymbol{x}_b$的维数相同，是两个向量之间的权值，使BILSTM和IDCNN提取的不同维度特征得以融合。通过在特征抽取层引入注意力机制，对得到的特征进行特征融合，提高了模型在自然语言处理任务中的性能。这种方法能够更好地捕捉不同位置之间的语义关联，从而提高模型对文本信息的理解能力。

1.5 CRF层

模型的顶层是CRF，该层可以学习标签特征，根据标签邻居关系获得最优序列，并通过添加约束条件来检查标签的有效性。CRF层可以自动学习训练过程中的约束条件。给定的输入序列x与对应的输出序列y的概率得分如式(3)所示：

$$s(x,y)=\sum_{i=1}^{n}\boldsymbol{L}_{y(i-1),y_i}+\sum_{i=1}^{n}P_{i,y_i} \tag{3}$$

其中，$\boldsymbol{L}$是转移矩阵，$\boldsymbol{L}_{y(i-1),y_i}$表示从标签$y_{(i-1)}$到标签$y_i$的转移分数，$P_{i,y_i}$表示输入序列$x$的第$i$个字符标记为标签$y_i$的概率。标签序列$y$的条件概率分布如式(4)所示：

$$P(y|s)=\frac{e^{s(x,y)}}{\sum_{\tilde{y}\in y_x}e^{s(x,\tilde{y})}} \tag{4}$$

在CRF的训练过程中，使用最大似然法来最大化正确标签序列的概率y^*，如式(5)所示：

$$\log p(y^*|s)=s(x,y^*)-\log\left(\sum_{\tilde{y}\in y_x}e^{s(x,\tilde{y})}\right) \tag{5}$$

使用Viterbi算法获得的最高评分标签序列为CRF输出的全局最优结果，如公式(6)所示：

$$y^*=\arg\max_{\tilde{y}\in y_x}(x,\tilde{y}) \tag{6}$$

2 实验设计与分析

2.1 数据集

实验数据使用阿里云天池大数据平台提供的中文糖尿病科研文献实体数据集DiaKG[25]，本数据集由41篇中文糖尿病领域专家共识文献组成，数据包括基础研究、临床研究、药物使用、临床病例、诊治方法等多个方面，时间跨度达到7年，涵盖了近年来糖尿病领域最广泛的研究内容和热点，该数据集共标注了22 050个医学实体和6 890对实体关系，定义了18类实体类型，数据集实体的详细信息如表1所示。

本文使用BIO序列标注模式对字符进行标注。其中“B”表示实体的开头，“I”表示实体除开头外的其他部分，“O”表示非实体。最后将数据集按照8∶2的比例分为训练集和测试集进行实验。

2.2 评估指标

实验采用命名实体识别中常用的精确率P、召回率R和F1值作为评价指标，具体公式如下：

表1 DiaKG数据集

Table 1 DiaKG dataset

实体类型	示例	数量
疾病	对糖尿病微血管病变无改善作用	5 743
疾病分期分型	心功能III-IV级、终末期肾病	1 262
病因	若体重增加，可能加重胰岛素抵抗	175
发病机制	多数患者的β细胞完全破坏	202
临床表现	已发生明确的足趾、组长坏疽创面	479
检查方法	进行混合餐耐量试验（MMTT）	489
检查指标	血糖仪测量指血（毛细血管血）血糖	2 718
检查指标值	血糖＜3.3 mmol/L	1 356
药物名称	包括COX-2抑制剂	4 782
用药频率	12 h后按照0.5 mg，1~3次/d	156
用药剂量	根据0.3~0.5单位/千克体重	301
用药方法	胰岛素在餐前15~30 min皮下注射	399
非药治疗	认知-行为及心理干预是通过调整	756
手术	进行胰岛细胞移植手术	133
不良反应	贝特类可使胆结石的发生率升高	874
部位	糖尿病相关微血管和大血管并发	1 876
程度	中到重度肾功能不全的患者	280
持续时间	预防治疗维持3~6个月	69

$$P = \frac{T_P}{T_P + F_P} \tag{13}$$

$$R = \frac{T_P}{T_P + F_N} \tag{14}$$

$$F1 = \frac{2PR}{P + R} \tag{15}$$

其中，T_P为识别到正确实体的个数；F_P为识别到非实体的个数；F_N为未识别到正确实体的个数。

2.3 实验环境

实验是在Ubuntu20.04上进行的，Python版本为3.8，PyTorch版本为2.0.0，CUDA版本为11.8，GPU版本为RTX A5000，显存为24 GB，CPU版本为Xeon(R) Platinum 8350C，系统内存为42 GB。实验使用RoBERTa-wwm-ext预训练模型提供字符嵌入，学习率为8e-5，epoch为15，batch_size设置为32，最大序列长度max-length为156。

2.4 实验结果与分析

2.4.1 主流模型对比分析 为了验证所提模型在糖尿病领域中的性能，将该模型在中文糖尿病数据集DiaKG上与用于关系三重抽取的级联二值标记框架（Cascade-CRF）、联合实体和关系提取的分区过滤网络（PFN）、利用外部知识进行联合实体和关系提取的新框架（KECI）、基于片段的NER框架（Global Pointer）和基于高阶医学知识图谱（High-Order MKGraph）5个主流模型进行了对比实验，使用精确率，召回率和F1值作为评价指标。模型对比结果如表2所示。

表2 模型对比结果（%）

Table 2 Model comparison results (%)

对比模型	精确率	召回率	F1值
Cascade-CRF	59.52	65.97	62.58
PFN	-	-	63.00
KECI	-	-	72.60
Global Pointer	73.36	65.76	69.35
High-Order MKGraph	-	-	74.20
本文模型	78.13	81.09	79.58

由表2可知，本文模型在中文糖尿病DiaKG数据集上的F1值为79.58%，相比High-Order MKGraph和KECI分别提高了5.38%和6.98%；并且精确度为78.13%、召回率为81.09%，相较于其他模型均有所提升，充分展示出模型具有良好的性能。实验也充分表明所本文模型对中文糖尿病命名实体识别相比联合抽取和其他方法更有效。

本文还选取F1值最高的对比模型与本文模型在DiaKG数据集上进行了实体的细致评估，结果如表3所示。实验结果表明，本文模型在实体类型识别中获得了较高的分数，特别是针对实体数量较少的类别，也体现出了良好的识别效果，说明本文提出的模型能有效解决糖尿病命名实体任务中实体种类多样、数据稀疏的问题，充分证明了本文模型的有效性。

2.4.2 消融实验 为验证本文模型各部分的有效性，选择在中文糖尿病DiaKG数据集上进行消融实验。其中-RoBERTa-wwm-ext表示使用BERT预训练模型代替RoBERTa-wwm-ext，-IDCNN表示只使用BILSTM来进行特征提取，-Attention表示用直接拼接代替Attention进行特征融合。消融实验结果如表4所示。

根据表4可知，使用RoBERTa-wwm-ext预训练模型的F1值要比BERT模型高1.26%，充分说明了RoBERTa-wwm-ext相比BERT有更好的语义表征能力，更适用于命名实体识别任务；使用BILSTM和IDCNN并行提取特征比只使用BILSTM的F1值提升了1.03%，说明使用双通道提取特征可以提取到不同粒度的特征，更好的理解上下文背景与语义关系；使用注意力机制进行特征融合比直接拼接F1值提升

表3 DiaKG数据集上的实体评估细节(%)

Table 3 Entity assessment details on DiaKG dataset (%)

实体类型	High-Order MKGraph			本文模型		
	精确率	召回率	F1	精确率	召回率	F1值
不良反应	77.55	52.78	62.81	70.85	83.93	76.84
用药剂量	26.67	23.53	23.53	63.24	70.49	66.67
部位	82.51	92.00	87.00	86.30	91.64	88.89
疾病分期分型	86.27	60.69	71.25	88.13	79.62	83.66
疾病	73.40	82.60	77.73	81.99	79.62	83.66
药物名称	66.80	77.62	71.80	92.45	93.76	93.10
用药方法	36.84	50.00	42.42	88.89	91.43	90.14
发病机制	22.22	25.00	23.53	44.64	60.98	51.55
病因	-	-	-	18.18	27.27	21.82
临床表现	52.17	37.50	43.63	29.51	54.55	38.30
检查方法	50.00	66.67	57.14	38.06	52.58	44.16
检查指标	53.99	68.75	60.48	77.13	73.60	75.32
非药治疗	38.24	50.00	43.34	47.92	43.40	45.54

表4 消融实验结果(%)

Table 4 Results of ablation study (%)

对比模型	DiaKG数据集		
	精确率	召回率	F1值
-RoBERTa-wwm-ext	77.20	79.51	78.32
-IDCNN	77.12	80.09	78.55
-Attention	77.74	80.42	79.05
本文模型	78.13	81.09	79.58

了0.53%,同时也说明了注意力机制可以有效融合全局信息和局部信息。综上所述,本文提出的模型对糖尿病命名实体识别有优秀的识别效果和强大的优越性。

3 结 语

本文提出了一种基于特征融合的糖尿病命名实体识别方法,旨在解决糖尿病文本中实体种类多样性、数据稀疏等问题。首先将数据使用预训练模型进行编码。其次,使用双通道的特征抽取方式进行字符级的特征抽取;同时,使用注意力机制将双通道提取的特征进行融合。最后,通过CRF得到最好的标签序列和最终结果,从而确定模型识别的准确性和有效性。在接下来的工作中,将在糖尿病命名实体识别的基础上,进行嵌套实体的识别工作。

【参考文献】

[1] Eligüzel N, Çetinkaya C, Dereli T. Application of named entity recognition on tweets during earthquake disaster: a deep learning-based approach[J]. Soft Comput, 2022, 26(1): 395-421.

[2] Pérez-Pérez M, Ferreira T, Igrejas G, et al. A deep learning relation extraction approach to support a biomedical semi-automatic curation task: The case of the gluten bibliome[J]. Expert Syst Appl, 2022, 195: 116616.

[3] Kung HY, Yu RW, Chen CH, et al. Intelligent pig-raising knowledge question-answering system based on neural network schemes[J]. Agron J, 2021, 113(2): 906-922.

[4] Xie SF, Xia YC, Wu LJ, et al. End-to-end entity-aware neural machine translation[J]. Mach Learn, 2022, 111(3): 1181-1203.

[5] Sharaf M, Hemdan EE, El-Sayed A, et al. An efficient hybrid stock trend prediction system during COVID-19 pandemic based on stacked-LSTM and news sentiment analysis[J]. Multimed Tools Appl, 2023, 82(16): 23945-23977.

[6] Khan N, Ma ZM, Ullah A, et al. Categorization of knowledge graph based recommendation methods and benchmark datasets from the perspectives of application scenarios: a comprehensive survey[J]. Expert Syst Appl, 2022, 206: 117737.

[7] Sun H, Saeedi P, Karuranga S, et al. IDF diabetes atlas: global, regional and country-level diabetes prevalence estimates for 2021 and projections for 2045[J]. Diabetes Res Clin Pract, 2022, 183: 109119.

[8] Luo Z, Fabre G, Rodwin VG. Meeting the challenge of diabetes in China[J]. Int J Health Policy Manag, 2020, 9(2): 47-52.

[9] Hettne KM, Stierum RH, Schuemie MJ, et al. A dictionary to identify small molecules and drugs in free text[J]. Bioinformatics, 2009, 25 (22): 2983-2991.

[10] Gao WC, Zheng XH, Zhao SS. Named entity recognition method of Chinese EMR based on BERT-BiLSTM-CRF[J]. J Phys Conf Ser, 2021, 1848(1): 012083.

[11] An Y, Xia XY, Chen XL, et al. Chinese clinical named entity recognition *via* multi-head self-attention based BiLSTM-CRF[J]. Artif Intell Med, 2022, 127: 102282.

[12] Zhang Y, Yang J. Chinese NER using lattice LSTM[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2018: 1554-1564.

[13] Devlin J, Chang MW, Lee K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg, PA, USA: ACL, 2019: 4171-4186.
[14] Liu N, Hu Q, Xu HY, et al. Med-BERT: a pretraining framework for medical records named entity recognition[J]. IEEE Trans Industr Inform, 2022, 18(8): 5600-5608.
[15] 马诗语, 黄润才. 基于ALBERT与BILSTM的糖尿病命名实体识别[J]. 中国医学物理学杂志, 2021, 38(11): 1438-1443.
Ma SY, Huang RC. Named entity recognition of diabetes based on ALBERT and BILSTM[J]. Chinese Journal of Medical Physics, 2021, 38(11): 1438-1443.
[16] 郭瑞, 张欢欢. 基于RoBERTa和对抗训练的中文医疗命名实体识别[J]. 华东理工大学学报(自然科学版), 2023, 49(1): 144-152.
Guo R, Zhang HH. Chinese medical named entity recognition based on RoBERTa and adversarial training[J]. Journal of East China University of Science and Technology, 2023, 49(1): 144-152.
[17] Wei ZP, Su JL, Wang Y, et al. A novel cascade binary tagging framework for relational triple extraction[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: ACL, 2020: 1476-1488.
[18] Yan ZH, Zhang C, Fu JL, et al. A partition filter network for joint entity and relation extraction[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: ACL, 2021: 185-197.
[19] Lai T, Ji H, Zhai CX, et al. Joint biomedical entity and relation extraction with knowledge-enhanced collective inference [C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. Stroudsburg, PA, USA: ACL, 2021: 6248-6260.
[20] Su JL, Murtadha A, Pan SF, et al. Global pointer: novel efficient span-based approach for named entity recognition[EB/OL]. (2022-08-05). https://arxiv.org/abs/2208.03054.
[21] Yang Z, Huang Y, Feng JL. Learning to leverage high-order medical knowledge graph for joint entity and relation extraction[C]//Findings of the Association for Computational Linguistics: ACL 2023. Stroudsburg, PA, USA: ACL, 2023: 9023-9035.
[22] Xu Z. RoBERTa-wwm-ext fine-tuning for chinese text classification [EB/OL]. (2021-02-24). https://arxiv.org/abs/2103.00492.
[23] Cui YM, Che WX, Liu T, et al. Revisiting pre-trained models for chinese natural language processing[C]//Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg, PA, USA: ACL, 2020: 657-668.
[24] Fang Y, Gao J, Liu ZL, et al. Detecting cyber threat event from twitter using IDCNN and BiLSTM[J]. Appl Sci, 2020, 10(17): 5922.
[25] 阿里云天池实验室. 中文糖尿病科研文献实体关系数据集DiaKG [EB/OL]. (2021-01-21) [2023-12-22]. https://tianchi.aliyun.com/dataset/88836.
Alibaba Yuntianchi Laboratory. Chinese diabetes research literature entity relationship dataset DiaKG [EB/OL]. (2021-01-21)[2023-12-22]. https://tianchi.aliyun.com/dataset/88836.

（编辑：薛泽玲）