计算机科学
COMPUTER SCIENCE

面向医学领域的文本特征增强多任务学习模型

郭瑞强, 贾晓文, 杨世龙, 魏谦强

引用本文

郭瑞强, 贾晓文, 杨世龙, 魏谦强. 面向医学领域的文本特征增强多任务学习模型[J]. 计算机科学, 2024, 51(11A): 240200041-7.

GUO Ruiqiang, JIA Xiaowen, YANG Shilong, WEI Qianqiang. Multi-task Learning Model for Text Feature Enhancement in Medical Field [J]. Computer Science, 2024, 51(11A): 240200041-7.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

基于多任务学习的复杂城市遥感图像道路提取

Road Extraction from Complex Urban Remote Sensing Images Based on Multi-task Learning

计算机科学, 2024, 51(11A): 240300095-8. https://doi.org/10.11896/jsjkx.240300095

基于改进Yolov8的敦煌壁画元素检测算法

Dunhuang Mural Element Detection Algorithm Based on Improved Yolov8

计算机科学, 2024, 51(11A): 231000034-6. https://doi.org/10.11896/jsjkx.231000034

基于位置交互感知网络的多任务情绪原因对抽取方法

Multi-task Emotion-Cause Pair Extraction Method Based on Position-aware Interaction Network

计算机科学, 2024, 51(11A): 231000086-9. https://doi.org/10.11896/jsjkx.231000086

对话场景下的情感引导问题生成模型

Emotion Elicited Question Generation Model in Dialogue Scenarios

计算机科学, 2024, 51(11): 265-272. https://doi.org/10.11896/jsjkx.231000002

基于AU的多任务学生情绪识别方法研究

Study on Multi-task Student Emotion Recognition Methods Based on Facial Action Units

计算机科学, 2024, 51(10): 105-111. https://doi.org/10.11896/jsjkx.240300059

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240200041

面向医学领域的文本特征增强多任务学习模型

郭瑞强[1,2,3]　贾晓文[1]　杨世龙[1]　魏谦强[1]
1 河北师范大学计算机与网络空间安全学院　石家庄 050024
2 河北师范大学河北省供应链大数据分析与数据安全河北省工程研究中心　石家庄 050024
3 河北省网络与信息安全重点实验室　石家庄 050024

摘　要　医学命名实体的识别和规范化是构建高质量医学知识图谱的基础。文中提出了一种基于文本特征增强的多任务学习模型,旨在解决现有模型中医学实体识别与规范化模型不能充分利用文本特征的问题。该模型添加词级、字符级特征和上下文语义信息来增强文本表示,再通过4个分级子任务,联合建模完成医学实体识别和规范化任务。实验表明,该模型能够学习实体识别和实体规范化这两个任务的共同特征,有效地提高学习的准确率。在NCBI和BC5CDR两个数据集上取得了较好的效果,在NER和NEN任务上的F1值分别为:91.09%,91.02%;92.05%,92%。

关键词:医疗命名实体识别;实体规范化;多任务;特征增强;联合建模
中图分类号　TP391

Multi-task Learning Model for Text Feature Enhancement in Medical Field

GUO Ruiqiang[1,2,3],JIA Xiaowen[1],YANG Shilong[1] and WEI Qianqiang[1]
1 School of Computer and Cyber Security,Hebei Normal University,Shijiazhuang 050024,China
2 Hebei Provincial Engineering Research Center for Supply Chain Big Data Analytics & Data Security,Hebei Normal University,Shijiazhuang 050024,China
3 Hebei Provincial Key Laboratory of Network and Information Security,Shijiazhuang 050024,China

Abstract　The recognition and standardization of medical named entities are the foundation for constructing high-quality medical knowledge graphs. This paper proposes a multi-task learning model based on text feature enhancement,aiming to address the issue of inadequate utilization of text features in existing models for medical entity recognition and standardization. The model incorporates word-level,character-level features,and contextual semantic information to enhance text representation. Through four hierarchical sub-tasks,it jointly models medical entity recognition and standardization tasks. Experiments indicate that the proposed model can learn common features for both entity recognition and entity standardization tasks,effectively improving the accuracy of learning. Satisfactory results are achieved on two datasets,NCBI and BC5CDR,with F1 scores for NER and NEN tasks 1.09%,91.02%;92.05%,92%,respectively.

Keywords　Medical named entity recognition,Entity normalization,Multitask,Feature enhancement,Joint modeling

1　引言

电子病历作为医生诊断后的记录,为人工智能与医疗方面的交叉研究提供了重要的数据基础。电子病例中的医学命名实体识别(Named Entity Recognition,NER)和命名规范化(Named Entity Normalization,NEN)研究在医学人工智能领域中是许多研究问题的工作基础,例如构建医学知识图谱。

NER是指识别文本中具有特定意义的实体,包括人名、地名、机构名等。NER任务可以识别出该实体的类别,对实体的边界进行划分。NEN是将文本中提取到的实体映射到标准标识符,如Mesh和OMIM。

现有的对于医学实体的研究多是面向实体识别,而没有进行规范化。而在少数进行医学实体识别和规范化的模型中[1],缺乏对文本特征的充分理解与运用,造成实体规范化的效果不佳。

为了更好地解决医学NER和NEN任务,本文提出了基于文本特征增强的多任务学习模型:首先利用BioBERT[2]模型分别获取文本中单词和字母的表示,然后采用双向长短时记忆网络(Bi-directional Long Short-Term Memory,BiLSTM)对单词进行编码,利用卷积神经网络(Convolutional Neural Network,CNN)对单词中的字符进行编码。这一方法充分考虑句子中的长距离依赖关系和字符信息,同时获取词级和字符级特征,从而为医学命名实体识别与规范化任务提供更加全面和丰富的信息。为了充分利用文本和实体属性,本文设计了4个渐进难度的任务。一级任务是传统的NER任务,它负责提取文本中的实体,确定实体边界。二级任务是NEN任务,目的是确定文本中是否存在候选标准实体。三级任务结合了前两个任务,以获取文本中应该映射到候选标准实体

基金项目:2023年度河北省引才引智创新平台(606080123003)
This work was supported by the 2023 Hebei Province Talent Introduction and Intelligence Innovation Platform(606080123003).
通信作者:郭瑞强(rqguo@mail.hebtu.edu.cn)

的实体提及。二级任务和三级任务可以学习医学提及和候选标准实体的对应关系。四级任务结合了前两个任务，并利用自注意力机制实现一句话中多个实体的输出。一级任务可以从三、四级任务中获得监督信号，从而更准确地提取文本信息。

本文的主要贡献如下：

(1)提出了一种结合双向长短时记忆网络(BILSTM)和卷积神经网络(CNN)的方法，以增强文本语义表达能力，并提高模型的特征表达能力。

(2)利用卷积神经网络(CNN)学习字符级文本表示，以更好地利用文本内部的字符信息。这种方法在一定程度上解决了未见实体问题，并提升了对实体规范化的效果。

(3)通过设计四级任务，按照难度递增的方式对原始文本和标准实体中医学提及的细粒度特征进行建模，使模型对长实体的规范化效果得到提升，并通过这种方式降低了误差传播，提高了模型的鲁棒性和效率。

2 相关工作与研究

近年来，有不少研究在进行命名实体识别的任务。譬如 Zhao 等提出了一种融合词信息嵌入的注意力自适应模型的中文命名实体识别[3]，Yang 等提出了一种基于门控多特征提取器的中文命名实体识别[4]。这些方法的特点是能够有效地从文本中提取实体，但是没有考虑实体的规范化，即将实体映射到标准实体标识符。DNorm[5]是第一种使用机器学习来规范化疾病名称的技术。LeadMine[6]是一种基于字典/语法的实体识别器，用于识别化学物质和疾病并将其标准化为 MeSH 标识符。TaggerOne[7]可针对任意实体类型进行训练，其由半马尔可夫结构线性分类器组成，具有用于 NER 的丰富特征方法和用于规范化的监督语义索引。为了完成医学实体识别和规范化任务，之前传统方法是使用串行 NER 和 NEN 模型[8]。其中 NER 模块负责提取医学文本中的医学实体提及，NEN 模块负责将这些医学实体提及映射到标准实体标识符。但是，这种方法存在误差传播的问题。为了解决这个问题，一些学者提出将 NER 和 NEN 任务一起合并建模。近年来 Bert[9]等预训练模型在自然语言处理(NLP)领域取得了很好的效果。Chen 等以 Bert 为基础，设计了一个联合解决医学命名实体识别和规范化问题的机器阅读理解和多重序列标注任务框架[10]，用来解决医学命名实体识别和规范化的传统模型存在的误差传播问题。Zhou 等提出了一个结合实体知识和深度上下文化词表示的知识增强系统，用以解决蛋白质/基因命名实体识别与规范化[11]，该方法可以利用上下文信息，解决名称变化和实体歧义问题。Lou 等提出的 Transition-based joint mode[12]将输出构建过程转化为增量状态过渡过程，全局学习与联合结构输出项对应的过渡动作序列，使用非局部特征提高精度，来执行疾病命名实体识别和规范化。Emma 等为实现命名实体识别，提出 IDCNN 模型[13]，在大篇幅上下文和结构化预测方面比传统的 CNN 具有更好的能力，保持与 BILSTM-CRF 准确性相当的同时，速度更快。Zhao 等为实现疾病实体识别，提出 MCNN[14]，引入多标签策略来代替 CRF 层，将字符级嵌入、单词级嵌入和词典特征嵌入串联起来。然后，在级联嵌入上堆叠几个卷积层。最后，将 MLS 策略应用于输出层，以捕获相邻标签之间的相关性信息。Wonjin 等为实现生物医学命名实体识别，提出 CollaboNet[15]，利用了多个 NER 模型的组合，以便目标模型从其他合作者模型中获得信息，以减少误报。Zhao 等提出 MTL-MERN 模型[16]，来联合建模实体识别和规范化，该方法一方面受益于多任务学习提供的两个任务的一般表示，另一方面成功地将分层任务转换为并行多任务设置，同时保持任务之间的相互支持。这两个方面都提高了模型性能。Zhou 等提出的 E2EMERN[17]设计了 3 个难度递增的渐进式任务以有效地联合建模医学 NER 和 NEN，减少误差传播，提高了模型准确率。Chen 等提出 BiLSTM-Biaffine，基于词义增强进行生物医学命名实体识别，通过 BioBERT 获取语素表示信息，在单词层面利用 BiLSTM 分别获取语素的前向和后向序列信息[18]。Yu 等提出基于字符级特征自适应的生物医学命名实体识别，使用 CNN 和 BiLSTM 提取文本的字符向量，再动态计算两者向量权重进行拼接，让模型在字符粒度上的利用更加充分[19]。Zhou 等提出了一种用于医学命名实体识别和规范化的多任务对抗主动学习模型 MTAAL，利用样本间的异质性来满足多样性约束[20]。

这些方法都充分利用了深度学习模型的优势，在医学命名实体识别和规范化方面取得了较大的成果。但是当实体存在多义或歧义，即存在变型实体时，不能很好地进行规范化处理，并且对于长实体的规范化效果不佳。为此，本文提出了一种多任务模型，结合双向长短时记忆网络(BiLSTM)和卷积神经网络(CNN)，以增强文本语义表达和特征表达能力；同时利用 CNN 学习字符级文本表示，解决未见实体问题，提升实体的规范化效果；并通过四级任务设计，建模医学文本的细粒度特征，提高对长实体规范化效果，降低误差传播，以提高模型的鲁棒性和效率。

3 任务与模型

3.1 模型介绍

本文采用基于文本特征增强的多级任务模型来解决医学文本中的 NER 和 NEN 任务。一级任务是传统的 NER 任务，用于得到文本中出现的所有医学实体提及。二级任务是 NEN 任务，目标是确定该医学文本中是否包含某个特定医学标准实体。通过二级任务，模型建立起文本和查询实体的关系，获取到新的表示，但没有考虑到具体哪个医学实体提及应该映射到该标准实体。如果原文本中不包含查询的标准实体的内容，则该任务的输出为空。三级任务是确定映射到标准实体的实体提及。四级任务是用于得到文本中所有的规范化实体。

模型架构如图 1 所示，编码层 1 由 BioBERT 和 BiLSTM 组成；编码层 2 由 BioBERT 和 CNN 组成；编码层 3 由 BioBERT 构成。BioBERT 是一个在大规模生物医学语料库上预先训练的特定领域语言表示模型，采用 WordPiece 方法来解决 OOV(Out-Of-Vocabulary)问题，并且对英文大小写进行区分，在很大程度上优于 BERT 和以前的最先进的模型。所以本文采用 BioBERT 首先对原始英文医学文本和字母进行向量化处理，以得到较好的向量化结果。BiLSTM 可以通过考虑前后文信息来提高文本向量对上下文信息的表示

能力，而 CNN 则可以从字符级别上对文本进行特征提取。这样，结合 BILSTM 和 CNN 可以有效地捕捉文本中的上下文信息和字符级别特征，从而提高命名实体识别和实体规范化的准确性和鲁棒性。给定一个句子 $X=\{x_1,x_2,x_3,\cdots,x_n\}$，$n$ 代表句子中单词的个数，如图 1 所示，“HH is a common autosomal recessive genetic disorder of iron metabolism.”，这段文本的字符表示为：“H H i s a c o m m o n a u t o s o m a l r e c e s s i v e g e n e t i c d i s o r d e r o f i r o n m e t a b o l i s m.”，以变量表示为：$M=\{m_1,m_2,m_3,\cdots,m_k\}$，$k$ 代表句子中字母的个数。将利用 BioBERT 进行向量化的过程简化为式(1)和式(2)：

$$H=BioBert(X)=\{h_1,h_2,h_3,\cdots,h_n\} \tag{1}$$

$$J=BioBert(M)=\{j_1,j_2,j_3,\cdots,j_k\} \tag{2}$$

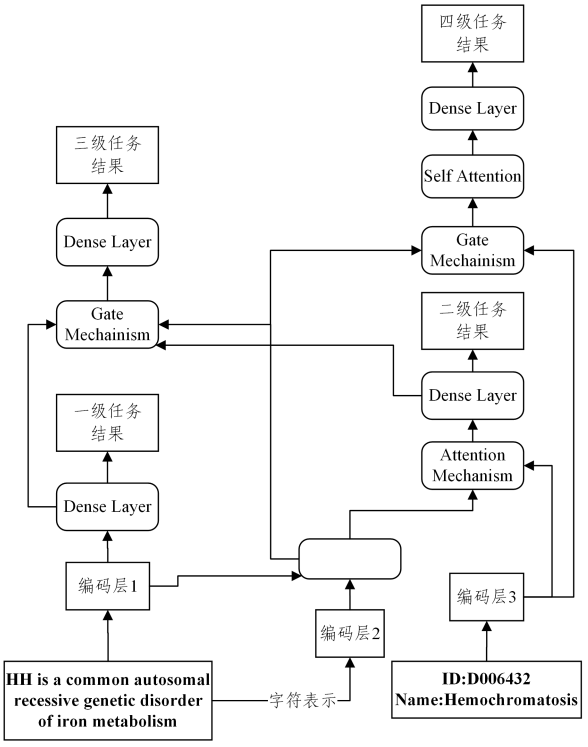

图 1　模型结构图

Fig. 1　Model structure diagram

3.2　任务介绍

3.2.1　一级任务

一级任务中使用 BiLSTM 处理文本获得时序特征 A，如式(3)所示：

$$A=BiLSTM(H)=\{a_1,a_2,a_3,\cdots,a_n\} \tag{3}$$

一级任务把 NER 标签作为目标，将经过 BiLSTM 得到的句子特征 A 送入 softmax 层，然后计算出一级任务的预测结果，计算过程如式(4)所示：

$$\hat{y}_i=\text{softmax}(W_k a_i+b_k) \tag{4}$$

其中，W_k 与 b_k 是可训练参数。在此例子中，一级任务的正确预测结果为：“B-Disease O O O B-Disease I-Disease I-Disease I-Disease O O O”。此阶段采用交叉熵损失函数来计算损失得分，一级任务的损失函数定义如式(5)所示：

$$L_1=-\sum_{i=1}^{n} y_i \log \hat{y}_i \tag{5}$$

3.2.2　二级任务

二级任务的模型输入包含 3 个部分：(X,X_m,e)，其中 X 是原文本，X_m 是文本的字符表示，e 是标准实体。在二级任务中，初始阶段加入了 CNN 对文本进行字符级特征提取，如式(6)所示：

$$B=CNN(J)=\{b_1,b_2,b_3,\cdots,b_k\} \tag{6}$$

把 k 与 n 设置成相同大小，文本的上下文向量化结果为：$C=\{c_1,c_2,c_3,\cdots,c_n\}=\{a_i\}_{i=1}^{n}+\{b_i\}_{i=1}^{n}$。为了在标准实体与原文本之间建立起更强的联系，本文还对标准实体进行了特征的提取。标准实体库中有实体的编号、名称与描述。模型将标准实体库中的实体名称输入到 BioBERT 中，再进行平均池化操作，得到标准实体库中第 i 个实体的特征为 c_i^e。二级任务中，模型还使用注意力机制[21]，让模型在训练过程中更加关注句子中的局部词，使原始文本中的医学提及与标准实体相关性更强。注意力权重的特征计算式如式(7)所示：

$$c^{\beta}=\sum_{i=1}^{m}\beta_i x_i \tag{7}$$

注意力得分 β 的计算式如式(8)、式(9)所示：

$$\beta_i=\frac{\exp(s(x_i,c^e))}{\sum_{i=1}^{n}\exp(s(x_i,c^e))} \tag{8}$$

$$s(x_i,c^e)=W_{\beta}[x_i;c^e]+b_{\beta} \tag{9}$$

其中，W_{β} 和 b_{β} 是注意力模块中可训练的权重参数。在获取文本注意力特征 c^{β} 和 c^e 之后，可以得到标准实体是否存在于文本中的预测结果，计算过程如式(10)所示，其中 σ 表示激活函数 softmax。

$$\hat{y}_e=\sigma(W_e[c^{\beta};c^e]+b_e) \tag{10}$$

在此例子中，输入的标准实体名称为“Hemochromatosis”，二级任务的正确预测结果为“1”，表示原文本中存在标准实体“D006432”。二级任务的损失用交叉熵损失函数来计算，如式(11)所示：

$$L_2=-(y_e\log\hat{y}_e+(1-y_e)\log(1-\hat{y}_e)) \tag{11}$$

3.2.3　三级任务

三级任务的模型输入包含 3 个部分：(X',X_m,e)。X' 是原文本在经过二级任务注意力计算的文本特征，X_m 是这段文本的字符表示，标准实体 e 为“D006432”。为了充分利用一级与二级任务所获得的结果，并使模型不会过拟合且具有良好的泛化性，本文采用门机制来融合上述两个任务过程中所取得的实体特征和句子特征以完成三级任务。其中门机制如式(12)所示：

$$G(C,C^a)=\sigma(W_g[C;C^a]+b_g) \tag{12}$$

融合句子特征如式(13)所示：

$$C^f=C\odot(1-G(C,C^a))+C^e\odot G(C,C^a) \tag{13}$$

之后将得到的融合特征输入到 softmax 层以得到预测结果，计算过程如式(14)所示：

$$\hat{y}_i^c=\text{softmax}(W_h c_i^f+b_c) \tag{14}$$

在此例子中，三级任务的正确预测结果为：“B-Disease O O O O O O O O O O”。该任务只负责识别出给定实体对应的实体提及，在该例子中就是识别标准实体“D006432”对应的实体提及“HH”。三级任务的损失函数定义如式(15)所示：

$$L_3=-\sum_{i=1}^{n} y_i^c \log \hat{y}_i^c \tag{15}$$

3.2.4 四级任务

四级任务的模型输入包含两个部分：(X_m, e)。该任务同样采用门机制来融合一级与二级任务过程中所取得的实体特征和句子特征。同时，也采用自注意力机制使模型可以更好地理解上下文和语义关联。自注意力的公式定义如式(16)—式(19)所示：

$$Q = X * W_q \tag{16}$$

$$K = X * W_k \tag{17}$$

$$V = X * W_v \tag{18}$$

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}}\right)V \tag{19}$$

之后将特征输入 softmax 层以得到预测结果。在此例子中，四级任务的正确预测结果为："D006432 O O O D030342 D030342 D030342 D030342 O O O"。该任务负责得到文本中所有的规范化实体。四级任务的损失函数定义如式(20)所示：

$$L_4 = -\sum_{i=1}^{n} y_i^d \log \hat{y}_i^d \tag{20}$$

为了共同处理这 4 个分级任务，本文设计了求和方程(21)，以计算模型的总损失：

$$L = \theta L_1 + \varphi L_2 + \delta L_3 + \gamma L_4 \tag{21}$$

其中，θ、φ、δ 和 γ 是用来平衡不同任务损失的超参数，利用反向传播算法得到的损失更新网络的可训练参数。为了增强模型的泛化性，在每个训练阶段结束后，对样本进行重采样。

4 实验

4.1 数据集和实验设置

本文在两个医学数据集上比较了提出的模型和现有的方法，表 1 列出了两个数据集的信息。NCBI 数据集[22]中有 798 篇公开的医学摘要，正文中的每个医学提及都用 MeSH/OMIM 标识符进行注释。BC5CDR 数据集[23]包含 1500 篇公共医学摘要，这些摘要也用 MeSH 标识符进行了注释。

表 1 NCBI 和 BC5CDR 数据集的统计信息

Table 1 Statistical information of NCBI and BC5CDR datasets

	NCBI	BC5CDR
训练集	5424	4560
验证集	923	4581
测试集	940	4797
实体数	7025	28545
NER 类别	3	5
NEN 类别	743	2311

4.2 评价指标

本文使用精确率(Precision)和召回率(Recall)来计算 F_1 值，从而衡量模型性能。M 是本模型预测得到的正确医疗实体总数，N 是模型所预测的医疗实体总数，E 是数据集的医疗实体总数。计算公式如式(22)—式(24)所示：

$$Prescision = \frac{M}{N} \tag{22}$$

$$Recall = \frac{M}{E} \tag{23}$$

$$F_1 = \frac{2 \times Precisison \times Recall}{Precisison + Recall} \times 100\% \tag{24}$$

4.3 实验设置

实验采用 Tensorflow2 框架，GPU 选用 NVIDIA GeForce GTX V100，网络参数优化器选择 Adam[24]。超参数 $\theta=0.125$，$\varphi=0.5$，$\delta=0.5$，$\gamma=0.05$，批尺寸为 6，训练轮数为 25，学习率是 1×10^{-5}。

4.4 比较的方法

为了验证模型的有效性，将本文提出的模型与其他的机器学习和深度学习医学实体识别和规范化模型进行对比。

本节采用在相关工作中提到的 Transition-based joint mode[12]、IDCNN 模型[13]、MCNN[14]、CollaboNet[15]、MTL-MERN[16]、E2EMERN[17]、MTAAL[20]、BiLSTM-Biaffine[18]和基于字符级特征自适应的生物医学命名实体识别模型[19]进行对比。

4.5 实验结果分析

在医疗命名实体识别和规范化方面，将本文提出的模型和上节提到的各种方法进行了效果比较，结果如表 2 所列。与 IDCNN 相比，CollaboNet 利用多个数据集作为输入，进行多任务学习，提高了 NER 任务的性能。MTL-MERN 充分利用了多任务学习和深度语义表示的优点，其性能优于上述方法。E2EMERN 采用了渐进式任务并使用 BioBERT 做特征提取，准确性获得较大提升。BiLSTM-Biaffine 在 NCBI 和 BC5CDR 数据集上的 NER 效果表现良好，说明了词义增强的有效性。字符级特征自适应模型在 NCBI 数据集上效果一般，与其没有采用 BioBERT 这个预训练模型有关。

表 2 本文模型与近年工作对比(F1 分数)

Table 2 Comparison between the proposed model and recent works (F1 scores)

方法	NCBI		BC5CDR	
	NER	NEN	NER	NEN
Transition-based Model	0.8205	0.8262	0.8382	0.8562
IDCNN	0.7983	0.7425	0.8011	0.8107
MCNN	0.8517	—	0.8783	—
CollaboNet	0.8636	—	0.8818	—
MTL-MERN	0.8743	0.8823	0.8763	0.8645
E2EMERN	0.9151	0.8901	0.9175	0.8965
MTTAL	0.7744	0.9287	0.8600	0.9153
BiLSTM-Biaffine	0.8907	—	0.9214	—
字符级特征自适应模型	0.8714	—	—	—
本文模型	**0.9109**	**0.9102**	**0.9205**	**0.9200**
本文模型-无字符卷积	0.9086	0.9048	0.9141	0.9050
本文模型-无 BILSTM	0.8995	0.9067	0.9126	0.9198
本文模型-无第四级任务	0.9100	0.9048	0.9202	0.9161

与上述方法相比，本文提出的基于文本特征增强的多任务学习模型在 NCBI 数据集中的实体识别任务的效果比 E2EMERN 模型差 0.5%，在实体规范化任务中比 MTAAL 模型差 1.8%，但是 MTAAL 模型的实体识别效果只有 77.44%。在 BC5CDR 数据集中的实体规范化任务比现有较好的模型效果高 0.47%。实验表明，在进行医学命名实体识别和规范化时，将 BILSTM 和 CNN 分别加入 BioBERT 后，对原文本及其字符进行编码，然后采用四级任务的方式，能有效增强特征表示，使模型在实体规范化任务中表现更为优秀，从而提高了整体性能。

具体而言，多任务学习方法能够通过共享模型参数来学习多个任务，从而增强特征表示的能力。在进行 BioBERT 特征提取之后，将 BiLSTM 和 CNN 整合到模型中，有助于捕捉文本的上下文信息和字符级别特征，进一步提升特征表示的效果。这两种方法的结合使得模型能够更好地处理医学文

本数据，实现更精准的实体识别和规范化。

4.6 消融实验

如表 2 所列，本文通过消融实验来验证各个模块的有效性。当取消对医学文本字符进行的卷积操作时，仅在 BioBERT 的基础上引入 BiLSTM，以增强时序信息。在较小的数据集 NCBI 上，NER 的 F1 值下降了 0.23%，NEN 的 F1 值下降了 0.14%；而在较大的数据集 BC5CDR 上，NER 的 F1 值下降了 0.64%，NEN 的 F1 值下降了 1.5%。说明对于字符卷积来说，文本量越大，效果越好。

在去除对医学文本的 BiLSTM 操作后，我们仅在 BioBERT 的基础上引入了对文本字母的卷积操作。在较小的数据集 NCBI 上，NER 的 F1 值下降了 1.14%，NEN 的 F1 值下降了 0.35%；而在较大的数据集 BC5CDR 上，NER 的 F1 值下降了 0.79%，NEN 的 F1 值下降了 0.02%。说明 BiLSTM 模块对 NER 起到了作用，与上面模型所采用的对字母进行卷积的操作形成一个互补的作用。

当去掉第四级任务，只利用前三级任务时，在较小的数据集 NCBI 上，NER 的 F1 值下降了 0.09%，基本保持不变；NEN 的 F1 值下降了 0.54%，略微下降。而在较大的数据集 BC5CDR 上，NER 的 F1 值下降了 0.03%，NEN 的 F1 值下降了 0.39%。说明通过第四级任务，模型可以更好地利用实体提及和标准实体间的信息，从而更好地完成实体规范化任务。

4.7 超参数实验

如表 3 所列，本文进行超参数实验来比较选取方法的有效性，采用的是平均池化操作。当采用最大池化后，在 NCBI 上 NER 的值下降了 0.77%，NEN 的值下降了 1.54%，在较大数据集 BC5CDR 上 NER 的值下降了 0.16%，NEN 下降了 0.55%。本文原模型使用正态分布进行卷积核的初始化，当采用随机初始化方式后，在 NCBI 上 NER 的值上升了 0.13%，NEN 的值下降了 2%；在较大数据集 BC5CDR 上 NER 的值下降了 0.39%，NEN 下降了 1.06%。可以看出卷积核的初始化对 NEN 任务影响较大。本模型采用的字符级别和词级特征的连接方式为特征向量相加，为了证明其有效性，该节设计了多种特征连接方式包括加权连接、取最大、取最小和特征相减，可以看到从 NCBI 和 CDR 两个数据集及 NER 与 NEN 两个方面，效果均不如特征向量相加，证明了该方法的有效性。

表 3 模型在不同超参数下的 F1 分数

Table 3 F1 scores of the model with different hyperparameters

模型	NCBI		BC5CDR	
	NER	NEN	NER	NEN
本文模型	0.9109	0.9102	0.9205	0.9200
最大池化	0.9023	0.8948	0.9189	0.9145
核初始化为随机	0.9122	0.8902	0.9166	0.9094
特征连接方式为 0.3 * 字符特征+0.7 * 词特征	0.9022	0.8940	0.9190	0.8911
特征连接方式为取最大	0.8987	0.8955	0.9174	0.8946
特征连接方式为取最小	0.8995	0.8751	0.9123	0.8880
特征连接方式为字符特征-词特征	0.9034	0.8983	0.9203	0.9165

4.8 模型对于未见实体的实验结果

本节对 NCBI 和 BC5CDR 的测试集进行了统计分析，NCBI 测试集中有 53% 的实体未在训练集中出现，而 BC5CDR 测试集中有 39% 的实体未在训练集中出现。图 2 比较了 E2EMERN 模型和本文提出的模型在测试集中的不可见样本的实体规范化效果。本文提出的模型采用了标准实体信息和渐进式多任务学习，且对文本语义从时序和词特征方面进行了增强，具有更强的泛化能力，能够对未知样本进行有效的规范化处理。

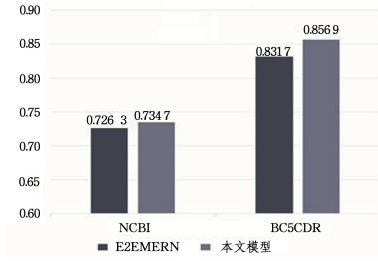

图 2 模型在未见实体数据集上的对比效果

Fig. 2 Comparison of model effects on unseen entity dataset

4.9 不同长度实体和特定实体类型实验结果

在本节中，首先将 NCBI 和 BC5CDR 中包含实体长度大于等于 6 的句子和包含实体长度小于 6 的句子分开处理，并在表 4 中呈现了本文模型在实体长度大于等于 6 的实体句子数据上的实验结果。本文模型在小数据集 NCBI 上的 NER 任务上的识别效果为 90.24%，在大数据集 BC5CDR 上的 NER 任务的识别效果为 86.35%。MTTAL 模型在 NCBI 上的 NER 任务上的识别效果为 83.20%，在 BC5CDR 上的 NER 任务的识别效果为 79.73%。这表明本模型在针对小数据集的 NER 任务上表现出更为显著的效果，并且本模型比 MTTAL 模型在两个数据集上的 NER 任务表现都好。此外，本模型在 NCBI 上的 NEN 任务的识别效果为 89.23%，而在 BC5CDR 上 NEN 任务的识别效果为 85.81%。MTTAL 模型在 NCBI 上的 NEN 任务上的识别效果为 93.23%，在 BC5CDR 上 NEN 任务的识别效果为 88.81%。这表明本模型在针对小数据集的 NEN 任务上表现出更为显著的效果。但是 MTTAL 模型在 NEN 任务上的表现更佳。

表 4 所提模型和 MTTAL 在长实体数据下的 F1 分数

Table 4 F1 scores of the proposed model and MTTAL model on long entity data

(%)

datasets	本文模型		MTTAL	
	NER	NEN	NER	NEN
NCBI	90.24	89.23	83.20	93.23
BC5CDR	86.35	85.81	79.73	88.81

在实体长度小于 6 的实体句子数据上评估了模型的性能，结果如表 5 所列。在小规模数据集 NCBI 上，本文模型对 NER 任务的识别效果为 91.32%，对 NEN 任务的识别效果为 91.04%；MTTAL 模型对 NER 任务的识别效果为 81.13%，对 NEN 的识别效果为 93.31%。本文模型在大规模数据集 BC5CDR 上，对 NER 任务的识别效果为 91.83%，对 NEN 任务的识别效果为 91.39%；MTTAL 模型对 NER 任务的识别效果为 82.67%，对 NEN 任务的识别效果为 89.38%。这说明本模型在中短实体数据集中，对 NER 任务的识别效果依然优于 MTTAL 模型，NEN 任务的识别与 MTTAL 的差距有所缩小，在 BC5CDR 数据集中效果比 MTTAL 要高 2%。

表 5 所提模型和 MTTAL 在中短实体数据下的 F1 分数

Table 5 F1 scores of the proposed model and MTTAL model on medium to short entity data

(%)

datasets	本文模型		MTTAL	
	NER	NEN	NER	NEN
NCBI	91.32	91.04	81.13	93.31
BC5CDR	91.83	91.39	82.67	89.39

本项工作中，针对包含 Disease 与 Chemical 两种实体的 BC5CDR 数据集，表 6 对不同类别的 NER 和 NEN 进行了效果比较。从横向比较来看，在 Disease 类别下，本文模型在 NER 方面识别效果为 88.60%，而在 NEN 方面的识别效果为 88.46%；在 Chemical 类别下，本文模型在 NER 方面的识别效果为 93.61%，在 NEN 方面的识别效果为 93.41%。从纵向比较来看，本文模型在 Chemical 类别下的 NER 和 NEN 效果比在 Disease 类别下分别高 5.01%和 2.95%。这是由于 BC5CDR 数据集中 Chemical 类别下的实体长度较短，模型对于较短实体的边界识别效果比较长实体要好，因此会有这样的差别。对比 MTTAL 模型，本文模型在 NER 任务上的表现较为突出，NEN 任务上两者表现较为接近。

表 6 所提模型和 MTTAL 在不同实体类别下的 F1 分数

Table 6 F1 scores of the proposed model and MTTAL model in different entity categories

(%)

实体	本文模型		MTTAL	
	NER	NEN	NER	NEN
Disease	88.60	88.46	79.95	88.76
Chemical	93.61	93.41	81.87	88.90

4.10 变型实体实验结果

医学文本中存在大量的同义词和一个实体的交替拼写。例如实体编号 D011125，它在文本存在的变型写法有：FAP；APC；familial adenomatous polyposis；adenomatous polyposis coli；familial adenomatous polyposis syndrome；Inherited colorectal polyposis。表 7 显示了模型对于部分存在特定变型实体句子的识别效果。其中 D011125 来自小数据集 NCBI，D007674 来自大数据集 BC5CDR。可以看出，对于变型实体，在大小数据集上，本文模型的 NER 的识别效果比 NEN 要高 2%～6%，说明对于变型实体的规范化，还存在一定的困难。本文提出的模型在小数据集中的变型实体 D011125 上，结果为 89.62%，在大数据集的变型实体 D007674 上，实现了 90.58%的效果。MTTAL 模型在 NER 任务中的表现都比本文模型要差，在 NEN 任务中表现与本文模型相近。

表 7 模型对于变型实体的 F1 分数

Table 7 F1 scores of the proposed model and MTTAL model on variant entities

(%)

编号	本文模型		MTTAL	
	NER	NEN	NER	NEN
D011125	95.65	89.62	79.43	92.97
D007674	92.57	90.58	79.46	88.99

结束语 本文提出了一个新的模型，该模型结合深度神经网络和多任务学习方法，完成命名实体识别和规范化两个任务，并通过联合训练来提高两个任务的准确性。与现有方法相比，该模型从单词和字母两个方面进行了文本语义增强，并且由 4 个难度递增的任务组成，有助于对原始文本中医学实体提及和标准实体的细粒度特征进行建模，并减少误差传播。通过对该模型的实验效果详细分析，证明了它的有效性。

参 考 文 献

[1] ZHOU B Z, CAI X R, ZHANG Y, et al. MTAAL: Multi-Task Adversarial Active Learning for Medical Named Entity Recognition and Normalization[C]// Proceedings of the 35th AAAI, California. Palo Alto, AAAI Press, 2021: 14586-14593.

[2] LEE J, YOON W, KIM S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining [J]. Bioinformatics, 2020, 36(4): 1234-1240.

[3] ZHAO P, DOU Q S, TANG H L, et al. Attention Adaptive Model with Word Information Embeding for Named Entity Recognition[J]. Computer Engineering and Applications, 2023, 59(8): 167-174.

[4] YANG R Y, HE Q, DU N S. Chinese Named Entity Recognition Based on Gated Multi-Feature Extractors[J]. Computer Engineering and Applications, 2022, 58(8): 117-124.

[5] ROBERT L, REZARTA L, ZHIYONG L. Dnorm: disease name normalization with pairwise learning to ran. [J]. Bioinform, 2013, 29(22): 2909-2917.

[6] LOWE D M, O'BOYLE N M, ASAYLE R. Leadmine: Disease identification and concept mapping using wikipedia[C]// Proceedings of the Fifth BioCreative Challenge Evaluation Workshop. Seville, CEUR Workshop Proceedings, 2015: 240-246.

[7] ROBERT L, ZHIYONG L. Taggerone: joint named entity recognition and normalization with semi-markov models. [J]. Bioinform, 2016, 32(18): 2839-2846.

[8] SAHU S, ANAND A. Recurrent neural network models for disease name recognition using domain invariant features[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Stroudsburg, Association for Computational Linguistics, 2016: 2216-2225.

[9] DEVLIN J, MINGWEI C, LEE K, TOUTANOVA K. 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. Florence, Minneapolis, Association for Computational Linguistics, 2019: 4171-4186.

[10] XIONG Y, HUANG Y H, CHEN Q C, et al. TANG B Z. A joint model for medical named entity recognition and normalization [C]// Proceedings of the Iberian Languages Evaluation Forum colocated with 36th Conference of the Spanish Society for Natural Language Processing. Málaga, IberLEF@SEPLN, 2020: 499-504.

[11] ZHOU H, NING S, LIU Z, et al. Knowledge-enhanced biomedical named entity recognition and normalization: application to proteins and genes. [J]. Bioinform, 2020, 21(1): 35-50.

[12] LOU Y X, ZHANG Y, QIAN T, et al. A transition-based joint model for disease named entity recognition and normalization. [J]. Bioinform, 2017, 33(15): 2363-2371.

[13] EMMA S, PATRICK V, DAVID B, ANDREW M. Fast and accurate entity recognition with iterated dilated convolutions[C]// Proceedings of the 25th Conference on Empirical Methods in Natural Language Processing. Copenhagen, Copenhagen, Asso-

ciation for Computational Linguistics,2017:2670-2680.

[14] ZHAO Z H,YANG Z H,LUO L,et al. Disease named entity recognition from biomedical literature using a novel convolutional neural network. [J]. BMC Medical Genomics, 2017, 10(5):73-82.

[15] WONJIN Y,CHAN H S,JINHYUK L,et al. Collabonet:collaboration of deep neural networks for biomedical named entity recognition[J]. BMC Bioinform,2019,20(10):55-65.

[16] ZHANG S D,LIU T,ZHAO S C,et al. A neural multi-task learning framework to jointly model medical named entity recognition and normalization[C]// Proceedings of the 33th AAAI. Hawaii,Honolulu,AAAI Press,2019:817-824.

[17] ZHOU B H,CAI X R,ZHANG Y,et al. An End-to-End Progressive Multi-Task Learning Framework for Medical Named Entity Recognition and Normalization[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing,Online,Aug 1-8,Florence. Association for Computational Linguistics,2021:6214-6224.

[18] CHEN M X,CHEN Y P,HU Y,et al. Biomedical Named Entity Recognition Method Based on Word Meaning Enhancement[J]. Computer Engineering,2023,49(10):305-312.

[19] YU X Q,WANG X,LI Z Q,et al. Biomedical Named Entity Recognition Based on Character Level Feature Adaptation[J]. Journal of Chinese Computer Systems,2023,44(9):1876-1883.

[20] ZHOU B,CAI X,ZHANG Y,et al. MTAAL:multi-task adversarial active learning for medical named entity recognition and normalization[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2021:14586-14593.

[21] ZHOU P,SHI W,TIAN J,et al. Attention-based bidirectional long short-term memory networks for relation classification [C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, Berlin, Aug 7-12, Stroudsburg. Association for Computational Linguistics,2016:207-212.

[22] REZARTA I D,ROBERT L,LU Z Y. NCBI disease corpus:A resource for disease name recognition and concept normalization. [J]. Journal of Biomedical Informatics,2014,47(1):1-10.

[23] JIAO L,SUN Y P,ROBIN J J,et al. Biocreative V CDR task corpus:a resource for chemical disease relation extraction. [J]. Database(Oxford),2016,2016(2016):68-78.

[24] DIEDERIK P. KINGMA,JIMMY L B. Adam:A method for stochastic optimization[C]// Proceedings of the 3rd International Conference on Learning Representations, San Diego, May 7-9, Ithaca. Conference Track Proceedings,2015:602-616.

GUO Ruiqiang, born in 1974, Ph.D, professor, master supervisor, is a member of CCF (No. 17546M). His main research interests include database system design, data mining, big data processing.