文章编号：1673-3193(2024)03-0265-09

基于多特征融合嵌入与DCNN的临床命名实体识别模型研究

杨　旭，梁志剑

(中北大学 计算机科学与技术学院，山西 太原 030051)

摘　要：针对目前最先进的临床命名实体识别(Cinical Named Entity Recognition，CNER)模型未能充分挖掘文本的全局信息和语义特征，以及未能解决文本中的字符替换等问题，改进了传统的单词嵌入模型，并在此基础上提出了一种结合深度卷积神经网络和双向短时记忆条件随机场(DCNN-BiLSTM-CRF)的临床文本命名实体识别方法。改进的单词嵌入模型融合词根、拼音和字符本身意义，使用了来自Transformers的双向编码器表示，使单词嵌入向量具有汉字和临床文本的特点，该方法通过在临床命名实体识别任务中引入深度卷积神经网络(Deep Convolutional Neural Networks，DCNN)，解决了CNN预测时丢失部分信息无法找回的问题。通过使用DCNN，本文模型能够更有效地捕获全局信息、获取字符之间的权重关系和多层次语义特征信息，从而提高了临床命名实体识别的准确性。在数据集CCKS2017和CCKS2018上分别进行实验，实验结果表明，与基准模型相比，该模型$F1$值分别改善了0.48%，0.68%，0.6%，0.58%，0.04%和1.43%，2.36%，3.31%，1.11%，0.17%。为了进一步验证本文的模型，进行了两种消融实验。结果表明，在两个数据集CCKS2017和CCKS2018上本文模型对比变体模型M1，$F1$值分别改善了0.79%和0.84%；对比变体模型M2，$F1$值分别改善了0.53%和0.64%。这些实验结果证明了本文所提算法的可行性。

关键词：临床命名实体识别；多特征融合嵌入；深度卷积神经网络；BLSTM-CRF；BERT

中图分类号：TP391　　**文献标识码**：A　　**doi**：10.3969/j.issn.1673-3193.2024.03.003

引用格式：杨旭，梁志剑. 基于多特征融合嵌入与DCNN的临床命名实体识别模型研究[J]. 中北大学学报(自然科学版)，2024，45(3)：265-273.

YANG Xu，LIANG Zhijian. Research on clinical named entity recognition model based on multi-feature fusion embedding and DCNN[J]. Journal of North University of China(Natural Science Edition)，2024，45(3)：265-273.

Research on Clinical Named Entity Recognition Model Based on Multi-Feature Fusion Embedding and DCNN

YANG Xu，LIANG Zhijian

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

Abstract：In order to address the limitations of the current state-of-the-art clinical named entity recognition(CNER) models，which fail to fully exploit the global information and semantic features in text and address issues like character substitutions，we had improved the traditional word embedding model and proposed a novel approach that combines deep convolutional neural networks with bidirectional short-term

收稿日期：2023-04-24

作者简介：杨　旭(1999—)，男，硕士生，主要从事人工智能、自然语言处理的研究。

通信作者：梁志剑(1978—)，男，副教授，硕士生导师，主要从事人工智能、自然语言处理、信息获取等研究。E-mail：zhijianliang@163.com。

memory conditional random field (DCNN-BiLSTM-CRF) for clinical text named entity recognition. The enhanced word embedding model integrated the meanings of word roots, phonetics, and characters themselves. It utilized bidirectional encoder representations from Transformers, enabling the word embedding vectors to capture the characteristics of both Chinese characters and clinical text. By introducing DCNN in the task of clinical named entity recognition, we addressed the issue of losing information that cannot be retrieved during CNN prediction. Through the utilization of DCNN, our approach was capable of capturing global information more effectively, capturing weight relationships between characters, and extracting multi-level semantic feature information, thereby improving the accuracy of clinical named entity recognition. We conducted experiments on the CCKS2017 and CCKS2018 datasets. The experimental results show that $F1$ score of our model improves 0.48%, 0.68%, 0.6%, 0.58%, 0.04% and 1.43%, 2.36%, 3.31%, 1.11%, 0.17% respectively when compared to the baseline model. Furthermore, to further validate our model, we performed two ablation experiments. Compared to variant model M1, our model achieved $F1$ score improvements of 0.79% and 0.84% on the CCKS2017 and CCKS2018 datasets respectively. Compared to variant model M2, our model achieved $F1$ score improvements of 0.53% and 0.64% on the same datasets. These experimental results confirm the feasibility of the proposed algorithm in this study.

Key words: clinical named entity recognition; multi-feature fusion embedding; DCNN; BILSTM-CRF; BERT

0 引 言

临床命名实体识别(Clinical Named Entity Recognition, CNER)是指从临床文本中识别并分类出与医学临床相关的实体,如疾病、药物、手术、解剖部位等,并将它们归类到预定义的类别[1-2]。例如,给出一段临床文字描述,“患者脑部CT结果呈缺血性脑中风”,缺血性脑中风是病名,脑是人体名称。CNER是自然语言处理(NLP)和信息检索(IR)领域的一个基础和重要的任务,对于提高医疗质量,辅助临床决策,构建知识图谱等有着重要的意义[1-2]。

CNER的主要挑战是临床文本的特殊性和复杂性,如缩略语、同义词、多义词、专业术语、噪声数据等,导致传统的基于规则或语法的方法难以适应不同领域和场景[2-3]。早期的方法通常采用手动的方式提取上下文特征,如隐马尔可夫模型(HMM)、支持向量机(SVM)[4]、朴素贝叶斯模型[5]、结构支持向量机(SSVM)[6]等。这些方法通常需要大量标注的训练数据,可以利用临床文本中的上下文信息和领域知识来提高模型性能。近年来,深度学习网络被用于命名实体识别领域。Tang等[7]提出了一种基于卷积神经网络(Convolutional Neural Networks, CNN)和双向长短期记忆网络(Bi-directional Long Short-Term Memory, BiLSTM)的端到端序列标记模型,利用多尺度CNN和自我注意机制在不同专业的数据集之间有效提升了迁移休息能力。Hou和Lin等[8-9]通过整合CNN从根号序列中提取的根号嵌入,进一步丰富了文本表示中包含的语义信息,提高了临床命名实体识别模型的性能。虽然对普通实体的预测效果较好,但由于网络深度有限,对专业领域种类繁多的实体识别仍是一个具有挑战的任务。

临床命名实体识别在英语语言环境中取得了成功,这是由于英语CNER任务相比汉语CNER任务更加容易,因为英语中的名词比汉语的更容易识别。汉语CNER任务的难点有:1)汉语词汇多义性较强,同一词在不同语境下可能有不同的意义。2)英文文本包含空格、首字母大写等标识符来确定实体边界,而中文文本没有类似的实体边界标识符,这增加了实体边界识别的难度。3)中文CNER任务通常需要结合中文分词和浅层解析,这些方法的准确性直接影响实体识别模型的有效性和稳定性。4)中文临床文本中存在许多不规范的缩略语或首字母缩写、同一实体的多种变体和丰富的临床术语。5)与英语不同,汉字是象形文字,包含丰富的形态信息[10],也可以作为潜在的边界[11]。以“病”为例,含有偏旁“疒”,常出现在“糖尿病”、“结核病”等疾病中[12]。因此,使

用常用的词嵌入算法计算中文词之间的相似度时无法考虑中文的特殊性质。目前，在中文临床命名实体识别领域，需要一种改进的适合中文的词嵌入算法与模型来避免信息丢失。

本文结合深度学习提出了一种适用于中文临床文本的基于改进多重特征词嵌入算法并结合深度卷积神经网络和BILSTM-CRF的模型（RS-BERT-DCNN-BILSTM-CRF，RBDBC），并将其用于中文CNER任务。在本文模型中，新的词嵌入算法结合了双向编码器表征法（Bidirectional Encoder Representations from Transformers，BERT），并融合了汉语拼音和部分词根信息的多特征词嵌入算法，充分展示了汉语的文本特征。BERT是一种语言表示模型，可以从大规模文本语料库中预训练基于上下文的深度双向表示[13]，其在序列标注任务中表现良好，使本文的算法能更有效地表征词的歧义性，增强句子的语义表征。同时，本文利用深度卷积神经网络（Deep Convolutional Neural Networks，DCNN）的优势获得局部特征。然后，本文使用BiLSTM网络从输入序列中获取了时间特征和上下文依赖性，得到临床句子中每个字符在多个时间步上的加权表示，使模型能够更准确地关注句子中相关的字符或单词。

本文的主要工作：1）提出了一种RBDBC模型来识别中文临床命名实体，该模型利用深度学习和文本特征来提高知识提取能力，并结合基本的BiLSTM-CRF模型从多个不同方面学习医学文本中不同字符之间的相关权重。2）设计了一种改进的词嵌入算法，该算法融合词根、拼音和字符本身的意义，使用了来自Transformers（BERT）的双向编码器表示，使词嵌入向量具有汉字和临床文本的特点。3）在两个真实的数据集CCKS-2017、CCKS-2018上进行了实验，计算结果表明本文方法比目前最先进的方法具有更好的性能。

1　相关工作

在命名实体识别任务的方法中，传统的方法主要是利用机器学习来提取特征。近些年，神经网络已经被广泛应用于多个领域，如命名实体识别。Chen[14]提出了一种用于句子分类的CNN，该网络采用定长滑动窗口提取文本特征，但忽略了文本全局信息的相关性。Ma等[15]提出了一种多任务神经网络，将CNN和BiLSTM并行使用，能够从字词序列、语法信息和地名词典信息中学习到更高阶的特征，但该方法对于实体边界的处理效果仍然不太理想。Aguilar等[16]使用多准则融合方法构建了BERT-DNN-CRF模型以挖掘语料库间的共有信息，从而提高了中文命名实体识别的准确率和召回率。蔡庆[17]提出了一种融合BERT的多层次司法文书实体识别模型，使用掩码语言模型（Masked LM）在BERT层进行无监督预训练，在中国裁判文书网上公开的裁判文书训练中，其$F1$值达到了89.12%，明显优于对照模型，进一步证明了深度学习在命名实体识别中的优势。张芳丛等[18]提出的一种新的深度学习模型有效解决了中文电子病历命名实体识别中存在的一词多义和词识别不全的问题，在基于CCKS的数据集上的$F1$实验值达到了89%，但由于使用的数据集规模较小，部分类别实体识别效果一般。罗凌等[19]将语言嵌入模型（Embeddings from Language Model，ELMo）应用到了中文上，并提出了一种基于汉字笔画的中文ELMo模型，在CCKS2017和CCKS2018 CNER数据集上的$F1$实验值分别为91.75%和90.05%。

2　RBDBC的命名实体识别模型

由于汉语的特殊性，文本处理的第一步是分词，给定文本序列$S=[w_1, w_2, \cdots, w_n]$，其中$w_t(1\leqslant t\leqslant n)$是句子中的第$t$个字符，本文的目标是用BIOES（Begin，Inside，Other，End，Single）标签方案对句子S中的每个字符w_t进行标记，得到一个包含实体类型及其对应位置标签的标签序列$Y=[y_1, y_2, y_3, \cdots, y_n]$。本节将对提出的模型进行概述并介绍本研究涉及的NER模型方法。

2.1　模型概述

本文提出的模型框架如图1所示。该模型包括多层神经网络层：输入层（Input layer）、嵌入层（Embedding Layer）、上下文建模层（Context Modeling Layer）、CRF层（CRF Layer）。输入层的作用是将原始输入的临床文本序列转换为模型可接受的数字向量矩阵形式，以便于后续的计算。嵌入层将字义信息与词根、拼音等特征信息拼接成最终的词向量。在上下文建模层通过BERT模型和DCNN卷积层获取上下文信息和局部特征。然后，将融合向量输入到BILSTM层中，利用BILSTM在提取序列的上下文特征方面的优势学习单词在句子中的长期

依赖关系和上下文信息。最后，将输出输送到CRF层，得到最可能的标签序列。

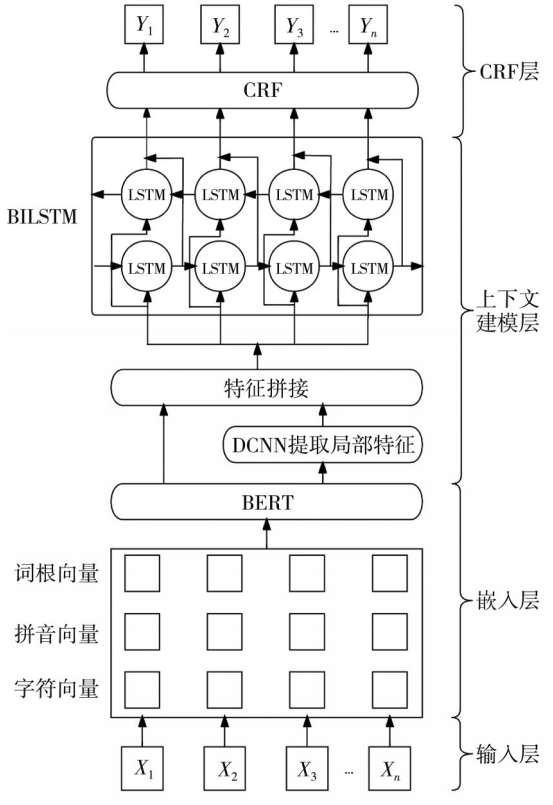

图1 模型框架示意图

Fig. 1 Schematic diagram of the model framework

2.2 输入层

现有的中文分词工具主要用于一般领域的文本处理，不能满足临床领域相关应用的需求。本文直接在字符层面对汉语临床句子进行切分，以避免分词错误的影响。给定汉语临床句子$S=[w_1,w_2,\cdots,w_n]$，输入层将每个字符w_n用独热编码(one-hot)表示，其维数等于训练数据集的词汇大小。经过处理后，句子表示变为$X=[x_1,x_2,\cdots,x_n]$。

2.3 嵌入层

本文嵌入层将尽可能识别字符的细粒度特征，提取字符的部首、拼音等特征，并将这三个特征融合起来，构建一个高质量的多特征单词嵌入模型，使单词向量能够充分应用于实体识别领域，从而识别更多的专有名词。

在命名实体识别任务中，汉字与英文有很大的区别，汉字可以直接识别单个字符，因此可以获得每个字符的部首特征，同时，汉字是象形文字，具有丰富的深层语义信息。中国临床文本数据不同于一般的领域数据集，其包含专业词汇，数据中的实体具有一些专有的特征信息，例如，部首“月”在汉字中出现在与身体部位相关的单词中，如“膀”，“胱”，“腰”。在本文的模型中，从输入层获得的一个独热编码向量将被进一步嵌入到相应的低维密集语义向量中。将输入层输出的$X=[x_1,x_2,\cdots,x_n]$作为输入，使用Python提供的cnradial工具包提取每个单词的特征信息，如词根和拼音。如果单词不具有词根和拼音的特征，则将信息标记为PAD(Padding Additional Data)，将提取的信息输入基于BERT的模型进行建模训练，学习理解文本的前后文联系，得到词级词嵌入向量$\boldsymbol{c}_i\in R_{dc}$、偏旁部首词嵌入向量$\boldsymbol{m}_i\in R_{dm}$和拼音词嵌入向量$\boldsymbol{p}_i\in R_{dp}$，其中$dc$表示词级词嵌入向量的维度，$dm$表示部分部首词的嵌入向量的维度，$dp$表示拼音词的嵌入向量的维度。最后，并将$\boldsymbol{c}_i$，$\boldsymbol{m}_i$和$\boldsymbol{p}_i$拼接得到的多特征融合词嵌入向量$\boldsymbol{w}_i$作为最终输出的单词向量。

$$\boldsymbol{m}_i=[\boldsymbol{c}_i,\boldsymbol{m}_i,\boldsymbol{p}_i]。\tag{1}$$

本文使用BERT模型来进行词向量的处理，模型结构如图2所示。

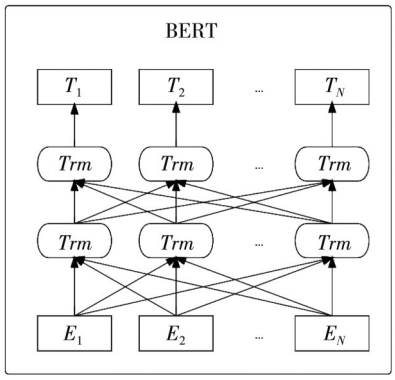

图2 BERT结构示意图

Fig. 2 Schematic diagram of BERT structure

本文BERT模型前馈大小均设置为4层，具体参数如表1所示。其中，T表示层数(Transformer blocks块)，C表示隐藏大小，H表示自注意力的头数，Total Para Meters表示模型的总参数数量。本文使用对汉语语料效果更好的BERT base版本。

表1 模型参数

Tab. 1 Model parameter

	T	C	H	Total Para Meters
BERTbase	12	768	12	110×10^6
BERTlarge	24	1 024	16	340×10^6

2.4　上下文建模层

为了更好地获取文本的时间特征、字符的上下文信息以及文本序列中字符之间的关联权重，在嵌入层之后部署了上下文建模层，该层由DCNN模块和BILSTM模块组成。将嵌入层输出的嵌入向量矩阵逐层馈送到上述模块进行进一步处理，然后将两个模块的最终输出向量作为上下文建模层的最终输出。

2.4.1　DCNN模块

在本文的模型中，DCNN模块由三层组成。在每一层中，将扩展卷积层的输出反馈给下一层扩展卷积层，每一层的卷积核和滤波器都设置为相同的数目。由于过多的层数会带来过多的参数，并导致过拟合问题。因此，本文将卷积层数设为3，迭代次数设为4。

DCNN模型第一层的膨胀宽度设置为1，并输出与输入大小相同的特征图，从而增大感受野以获取更多的上下文信息。膨胀卷积的输出为

$$i_n = D_1^0 w_n, \tag{2}$$

式中：D_ω^j表示宽度为ω的第j层膨胀卷积。

DCNN中第j层的输出为

$$C_n^j = r(D_{\omega_j}^{j-1} C_n^{L_c}), \tag{3}$$

式中：r表示RULE激活函数；L_c表示一个膨胀卷积层；j表示膨胀卷积层数。

将三层的卷积运算看作一个整体结构，在迭代后得到b_{kn}，k表示DCNN的迭代次数，将$(b_1, b_2, \cdots, b_n)$作为BILSTM的输入。

2.4.2　BILSTM模块

本文利用BILSTM来获取输入序列的特征。将$(b_1, b_2, \cdots, b_n)$输入到前向和后向长短时记忆（LSTM），然后结合前向LSTM$(\overrightarrow{h_1}, \overrightarrow{h_2}, \cdots, \overrightarrow{h_n})$和后向LSTM的隐藏状态$(\overleftarrow{h_1}, \overleftarrow{h_2}, \cdots, \overleftarrow{h_n})$来获取隐藏状态序列$(h_1, h_2, \cdots, h_n)$。

$$h_t = [\overrightarrow{h_t}, \overleftarrow{h_n}] \in R^m。 \tag{4}$$

与传统的递归神经网络（RNN）相比，LSTM的优点是它可以决定哪些信息被遗忘，哪些信息被重新训练。这个过程是由遗忘门f_t中的s型函数决定的。

$$f_t = \sigma(W_f \cdot [h_{t-1}, w_t] + b_f), \tag{5}$$

式中：W_f和b_f是可训练的参数；σ是sigmoid函数，输出值在0和1之间，与σ相乘的信息在$\sigma=0$时被遗忘，在$\sigma=1$时被保留。

2.5　CRF层

CRF（Conditional Random Field）模块为解码器层。本文首先利用线性层将由堆叠神经网络生成的隐态序列（即$(h_1, h_2, \cdots, h_n) \in R_{n*m}$）的维数变换为$(p_1, p_2, \cdots, p_n) \in R_{n*k}$），其中，$k$是候选标签的数量。然后利用CRF层预测最可能的结果。CRF层的参数是一个矩阵，记为矩阵$\boldsymbol{A}$，其中$\boldsymbol{A} \in (k+2)*(k+2)$。$A_{ij}$表示从第$i$个标签到第$j$个标签的过渡分数。

给定输入序列X和候选标签序列$Y=[y_1, y_2, \cdots, y_n]$，得到的CRF综合评分为

$$score(X, Y) = \sum_{i=1}^{n} P_{i, y_i} + \sum_{i=1}^{n+1} A_{y_{i-1}, y_i}。 \tag{6}$$

根据式（6），标签序列的得分等于每个位置得分的总和。然后利用Softmax函数得到归一化概率

$$P(Y|X) = \frac{\exp(score(X, Y))}{\sum_{Y'} \exp(score(X, Y'))}。 \tag{7}$$

接下来利用最大化对数似然函数来训练模型

$$\log P(Y^X|X) = score(X, Y^X) - \log\left(\sum \exp(score(X, Y'))\right)。 \tag{8}$$

最后，使用维特比算法选择最优路径，得到最终的标签序列为

$$Y^* = \arg\max_{Y'} score(X, Y')。 \tag{9}$$

3　模型性能对比实验

通过实验与其他的方法进行比较，并评估本文所提出模型的性能。

3.1　数据集

本文的所有实验主要在CCKS2017-CNER数据集和CCKS2018-CNER数据集上进行，这两个数据集是中文临床命名实体识别方法性能评价中广泛使用的公共基准数据集。CCKS2017-CNER数据集包含来自不同临床科室的400份中文病历，共计1 596个注释实例（10 024句），包含5类临床命名实体，包括疾病（Disease）、症状（Symptom）、检查（Exam）、治疗（Treatment）和身体部位（Body）。这些句子被中文句点和感叹号进一步分割成从句。CCKS2018-CNER数据集共包含1 000条记录，分为600个训练数据集和400个测试数据集，包括解剖部位（Anatomical part）、症状描述（Symptom description）、独立症状（Independent symptom）、疾病（Disease）和手

术(Treatment)等5类临床命名实体。两个数据集中不同类型实体的详细统计如表2和表3所示。

表2 CCKS2018中不同类型实体数量的统计信息

Tab. 2 Number statistics of different types of entities in CCKS2018

数据集	临床命名实体				
	Anatomical part	Symptom description	Independent symptom	Disease	Treatment
Training set	7 838	7 831	2 066	1 005	1 116
Test set	6 339	918	1 327	813	735

表3 CCKS2017中不同类型实体数量的统计信息

Tab. 3 Number statistics of different types of entities in CCKS2017

数据集	临床命名实体				
	Body	Symptom	Exam	Disease	Treatment
Training set	10 719	7 831	9 546	722	1 048
Test set	3 021	2 311	3 143	533	465

3.2 比对模型

本文选取了几个典型的词级和字符级NER模型作为比对模型，以验证本文的模型在中文临床命名实体识别中的性能。

首先介绍在CCKS2017-CNER数据集上评估的一些基线模型：

1) HIT-CNER[20]：该模型提出了一种混合系统，它将Rule，CRF，RNN、带特征的RNN这4种方法融合在一起，并在最后添加一个投票机制。

2) BiLSTM-CRF-DIC[21]：该模型将字典融合到深度神经网络中，解决了一些稀有实体不能被识别的问题。

3) RD-CNN-CRF[22]：该模型为中文CNER提供了一种带有条件随机场的残差扩张卷积神经网络，它使模型在计算中异步，从而加快了训练周期。

4) LM-Att-BiGRU-CRF[23]：该模型首先从未标注的临床医疗数据中训练字符向量和语言模型，然后利用标注数据来训练标注模型。

5) MKRGCN[24]：该模型可以将来自中文词典和领域知识等多个来源的知识用于CNER。

此外，为了进一步验证模型在CCKS2018-CNER数据集上的识别能力，本文还选择了一些在相同数据集上经过测试的先进模型来进行比较，具体如下：

1) FT-BERT+BiLSTM+CRF+Fea[25]：该模型在未标记的中国临床记录上预训练BERT模型，使用不同层分别提取文本特征和解码预测标签。

2) DUTIR[26]：该模型提供了一种神经网络集成方法，该方法结合了5个单独的神经网络模型采用了各种特征。

3) BiLSTM-CRF[27]：将BiLSTM-CRF模型被应用于中文电子病历，以识别这些病历中的相关命名实体。

4) MSD_DT_NER[28]：该模型可以利用未标记的领域特定知识。使用不同的图层，例如长短期记忆(LSTM)和条件随机域(CRF)，分别提取文本特征和解码预测的标签。

5) Attention-BiLSTM-CRF+all[29]：该模型提供了一种基于历史实体信息的实体自动校正算法，并且构建了药物词典和后处理规则，对实体进行了校正，进一步提高了性能。

3.3 参数设置

本文使用基于Python 3.8的Transformers-4.19.2库来实现本文的模型，使用Adam算法进行参数优化。初始学习率设置为10^{-5}，遗忘率设置为0.1。如果最近3轮没有更好的精度，将对学习率进行适当调整。为了避免过拟合，所有方法都采用了早期停止和退出策略，最大早停次数设置为3。此外，采用梯度裁剪的方法解决了梯度爆炸问题。对于DCNN模块，扩张宽度设置为1×2×4。对于BiLSTM模块，输入特征维数设置为768×2，隐层状态的维数设置为384。批次大小设置为8，对于每个数据集，本文选择其中的90%作为训练数据，其余作为测试数据。

3.4 各模型的比较

本文模型在CCKS2017不同类型临床实体上的实验结果如表4所示，在CCKS2018上的实验结果如表5所示。在两个数据集上与比对模型比较的结果如表6所示。

由表6可以看出，本文提出的模型与比对模型相比，在数据集CCKS-2017上的$F1$分数分别提高了0.48%，0.68%，0.6%，0.58%，0.04%。在数据集CCKS-2018上分别提高了1.43%，2.36%，3.31%，1.11%，0.17%。这主要是因为深度卷积神经网络在并行提取特征方面具有较大

的优势，与传统的卷积神经网络相比，它能够以较快的速度获得扩展宽度距离内的特征，并获得文本的远距离特征。同时，添加拼音、部首等特征可以提高模型的识别性能和准确性。

表 4　本文模型在CCKS2017上的实验结果

Tab. 4　Experimental results on CCKS-2017 of this paper model

评价指标	临床命名实体					
	Body	Symptom	Exam	Disease	Treatment	Overall
Precision/%	89.50	93.58	92.40	88.51	87.64	**92.28**
Recall/%	87.92	93.67	92.01	89.65	88.48	**91.57**
*F*1/%	88.70	93.63	92.21	89.08	88.05	**91.92**

表 5　本文模型在CCKS2018上的实验结果

Tab. 5　Experimental results on CCKS2018 of this paper model

评价指标	临床命名实体					
	Anatomical part	Symptom description	Independent symptom	Disease	Treatment	Overall
Precision/%	86.48	90.20	90.02	88.53	94.00	**88.31**
Recall/%	89.56	93.17	90.95	89.89	88.68	**93.82**
*F*1/%	88.42	91.66	90.48	88.92	91.26	**90.99**

表 6　各模型在CCKS2017和CCKS2018上的实验结果对比

Tab. 6　The experimental comparison results of different models on CCKS2017 and CCKS2018

数据集	评价指标			
	模型	Precision/%	Recall/%	*F*1/%
CCKS2017	HIT-CNER	91.99	90.30	91.44
	BiLSTM-CRF-DIC	90.83	91.64	91.24
	RD-CNN-CRF	90.63	92.02	91.32
	LM-Att-BiGRU-CRF	88.60	94.25	91.34
	MKRGCN	93.33	92.19	91.88
	Our model	92.28	91.57	**91.92**
CCKS2018	FT-BERT＋BiLSTM＋CRF＋Fea	92.06	91.15	89.56
	DUTIR	88.89	88.37	88.63
	BILSTM-CRF	88.52	86.86	87.68
	MSD_DT_NER	89.84	89.93	89.88
	Attention-BiLSTM-CRF＋all	91.26	90.38	90.82
	Our model	88.31	93.82	**90.99**

3.5　DCNN和BERT对实验的影响

为了研究DCNN和BERT对中文CNER任务执行的影响，本文设置了两组消融实验，使用相同的表示方法和数据集，将本文模型（RS-BERT-DCNN- BILSTM-CRF）与其两个变体进行了比较，变体模型分别为M1和M2。模型M1是将本文模型中的DCNN模块替换为CNN模块而获得的，模型M2是将本文模型中的BERT嵌入方法替换为word2vec获得的。表7和表8给出了这两个模型在两个CNER数据集上的性能结果。

由表7和表8可以看出，模型M1在CCKS2017和CCKS2018数据集上的*F*1分数分别为91.13%和90.15%，比本文所提模型分别低0.79%和0.84%。模型M2在CCKS2017和CCKS2018数据集上的*F*1分数分别为91.39%和90.35%，比本文所提模型分别低0.53%和0.64%。以上实验结果表明，与传统算法相比，本文基于多特征的词嵌入算法能够在训练单词向量中提取出更多的细粒度汉字特征，并且在上下文建模层和CRF层相同的情况下，将模型M1，M2与本文提出的模型相比较，本文的模型在实体识别方面比传统的word2vec和卷积神经网络有更好的效果，与传统的卷积神经网络相比，它能解决CNN预测时丢失的部分信息无法找回的问题，而且还能获得文本的远距离特征。综上所述，引入DCNN和BERT并结合多特征词嵌入的方法是有效的。

表7 模型M1在CCKS2017和CCKS2018上的实验结果

Tab. 7 Experimental results of model M1 on CCKS2017 and CCKS2018

数据集	模型	Precision/%	Recall/%	*F*1/%
CCKS2017	M1	90.81	91.45	91.13
	Our model	92.28	91.57	91.92
CCKS2018	M1	88.79	91.56	**90.15**
	Our model	88.31	93.82	**90.99**

表8 模型M2在CCKS2017和CCKS2018上的实验结果

Tab. 8 Experimental results of model M2 on CCKS2017 and CCKS2018

数据集	模型	Precision/%	Recall/%	*F*1/%
CCKS2017	M2	91.61	91.17	91.39
	Our model	92.28	91.57	**91.92**
CCKS2018	M2	90.30	90.41	90.35
	Our model	88.31	93.82	**90.99**

4 结 论

本文针对临床命名实体识别任务，提出了一种基于多特征融合嵌入和DCNN的模型，该模型提供了一种融合词义、部首、拼音特征的单词嵌入算法，利用Transformers的双向编码器表示，生成了具有中文和临床特点的单词向量。将DCNN与BiLSTM相结合，用于提取文本的全局信息和语义特征，解决了CNN信息丢失的问题，提高了文本的表达能力。在两个公开的临床命名实体识别数据集上进行了实验，结果表明，本文模型的准确率均优于两个数据集共10个基准模型，本文模型的*F*1值与比对模型相比分别提升了0.48%，0.68%，0.6%，0.58%，0.04%和1.43%，2.36%，3.31%，1.11%，0.17%。另外，通过消融实验验证了DCNN和多特征融合嵌入的有效性。本文模型在临床命名实体识别任务上取得了较好的效果，但仍有改进空间，未来的工作将在更多的数据集和领域上进行验证和应用，以提高模型的鲁棒性和泛化性。

参考文献：

[1] AL-NABKI M W, FIDALGO E, ALEGRE E, et al. Improving named entity recognition in noisy user-generated text with local distance neighbor feature[J]. Neurocomputing, 2020, 382: 1-11.

[2] 杨飞洪，张宇，覃露，等. 中文电子病历的命名实体识别研究进展[J]. 中国数字医学，2020，15(2)：9-12.
YANG Feihong, ZHANG Yu, TAN Lu, et al. A research progress of clinical named entity recognition from Chinese electronic medical records[J]. China Digital Medicine, 2020, 15(2): 9-12. (in Chinese)

[3] GAJENDRAN S, MANJULA D, SUGUMARAN V. Character level and word level embedding with bidirectional LSTM-dynamic recurrent neural network for biomedical named entity recognition from literature[J]. Journal of Biomedical Informatics, 2020, 112(445): 103609.

[4] GAIZAUSKAS R, DEMETRIOU G, HUMPHREYS K. Term recognition and classification in biological science journal articles[C]//Proceddings of the Computional Terminology for Medical and Biological Applications Workshop of the 2nd International Conference on NLP, 2000: 37-44.

[5] KRISHNA A, AKHILESH V, AICH A, et al. Sentiment analysis of restaurant reviews using machine learning techniques[C]//Emerging Research in Electronics, Computer Science and Technology. Singapore: Springer, 2019: 687-696.

[6] AWWALU J, BAKAR A, YAAKUB M R. Hybrid n-gram model using naive bayes for classification of political sentiments on Twitter[J]. Neural Computing and Applications, 2019, 31(12): 9207-9220.

[7] TANG B, CAO H, WU Y, et al. Recognizing clinical entities in hospital discharge summaries using structural support vector machines with word representation features[J]. BMC Medical Informatics & Decision Making, 2013, 13(S1): 1-10.

[8] HOU L L, ZHANG J, WU O, et al. Method and dataset entity mining in scientific literature: a CNN+BiLSTM model with self-attention[J]. Knowledge-Based Systems, 2022, 235: 107621.

[9] YIN M, MOU CH, XIONG K, et al. Chinese clinical named entity recognition with radical-level feature and self-attention mechanism[J]. Journal of Biomedical Informatics, 2019, 98: 103289.

[10] 陈剑，何涛，闻英友，等. 基于BERT模型的司法文书实体识别方法[J]. 东北大学学报(自然科学版)，2020，41(10)：1382-1387.
CHEN Jian, HE Tao, WEN Yingyou, et al. Entity recognition method for judicial documents based on BERT model[J]. Journal of Northeastern University (Natural Science), 2020, 41(10): 1382-1387. (in Chinese)

[11] MENG Y, WU W, WANG F, et al. Glyce: glyph-vectors for chinese character representations[DB/OL]. (2019-01-29)[2023-04-24]. https://arxiv.org/abs/

1901. 10125.

[12] SONG C H, SEHANOBISH A. Using chinese glyphs for named entity recognition (student abstract)[C]// The 34th AAAI Conference on Artificial Intelligence, 2020: 13921-13922.

[13] XUAN Z, BAO R, JIANG S. FGN: Fusion glyph network for Chinese named entity recognition [M]. Springer: Singapore, 2021.

[14] CHEN Y. Convolutional neural network for sentence classification [D]. Waterloo: University of Waterloo, 2015.

[15] MA X Z, HOVY E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF[DB/OL]. (2016-03-04) [2023-04-24]. https://arxiv. org/abs/1603. 01354v5.

[16] AGUILAR G, MAHARJAN S, LÓPEZ-MONROY P, et al. A multi-task approach for named entity recognition in social media data[C]//Proceedings of the 3rd Workshop on Noisy User-generated Text, 2017: 148-153.

[17] 蔡庆. 多准则融合的中文命名实体识别方法[J]. 东南大学学报(自然科学版), 2020, 50(5): 929-934.
CAl Qing. Chinese named entity recognition based on multicriteria fusion[J]. Journal of Southeast University (Natural Science Edition), 2020, 50(5): 929-934. (in Chinese)

[18] 张芳丛, 秦秋莉, 姜勇, 等. 基于 RoBERTa-WWM-BiLSTM-CRF 的中文电子病历命名实体识别研究[J]. 数据分析与知识发现, 2022, 6(2/3): 251-262.
ZHANG Fangcong, QIN Qiuli, JIANG Yong, et al. Named entity recognition for Chinese EMR with RoBERTa-WWM-BiLSTM-CRF [J]. Data Analysis and Knowledge Discovery, 2022, 6(2/3): 251-262. (in Chinese)

[19] 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.
LUO Ling, YANG Zhihao, SONG Yawen, 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)

[20] HU J L, SHI X, LIU Z J, et al. HITSZ_CNER: a hybrid system for entity recognition from Chinese clinical text [C]//CEUR Workshop Proceedings. Chengdu: The Technical Committee on Language and Knowledge Computing of the Chinese Information Processing Society of China, 2017: 25-30.

[21] WANG Q, ZHOU Y, RUAN T, et al. Incorporating dictionaries into deep neural networks for the Chinese clinical named entity recognition [J]. Journal of Biomedical Informatics, 2019, 92: 103133.

[22] QIU J, ZHOU Y, WANG Q, et al. Chinese clinical named entity recognition using residual dilated convolutional neural network with conditional random field[J]. IEEE Transactions on Nanobioscience, 2019, 18(3): 306-315.

[23] 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[J]. 计算机科学, 2020, 47(3): 211-216.
TANG Guoqiang, GAO Daqi, RUAN Tong, et al. Clinical electronic medical record named entity recognition incorporating language model and attention mechanism[J]. Computer Science, 2020, 47(3): 211-216. (in Chinese)

[24] XIONG Y , PENG H , XIANG Y , et al. Leveraging multi-source knowledge for Chinese clinical named entity recognition via relational graph convolutional network [J]. Journal of Biomedical Informatics, 2022, 128: 104035.

[25] LI X, ZHANG H, ZHOU X H. Chinese clinical named entity recognition with variant neural structures based on bert methods[J]. Journal of Biomedical Informatics, 2020, 107(5): 103422.

[26] LUO L, LI N, LI S C. DUTIR at the CCKS-2018 Task1: A neural network ensemble approach for chinese clinical named entity recognition [C]//Proceedings of the Evaluation Tasks at the China Conference on Knowledge Graph and Semantic Computing, 2018.

[27] JI B, LIU R, LI S, et al. A BILSTM-CRF method to Chinese electronic medical record named entity recognition [C]//International Conference on Algorithms, Computing and Artificial Intelligence, 2018: 1-6.

[28] WANG C , 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 (1): 103583.

[29] JI B, LIU R, LI S, et al. A hybrid approach for named entity recognition in Chinese electronic medical record [J]. BMC Medical Informatics and Decision Making, 2019, 19(S2): 64.