文章编号：2096-1472(2022)-12-30-07
DOI:10.19644/j.cnki.issn2096-1472.2022.012.007

结合部首特征和BERT-Transformer-CRF的中文电子病历实体识别方法研究

姚 蕾，蒋明峰，方 贤，魏 波，李 杨

(浙江理工大学计算机科学与技术学院，浙江 杭州 310018)
✉202030504181@mails.zstu.edu.cn; m.jiang@zstu.edu.cn; xianfang@zstu.edu.cn;
weibo@zstu.edu.cn; yangli@zstu.edu.cn

摘 要：在中文电子病历命名实体识别(CNER)中，中文文本缺乏划分单词边界的分隔符，一些现有的方法难以捕捉长距离相互依赖的特征。因此，文章提出一种利用预训练模型(BERT−Transformer−CRF，BTC)实现CNER的命名实体识别方法。首先，运用BERT(Bidirectional Encoder Representations from Transformers)提取文本特征。其次，使用Transformer捕捉字符之间的依赖关系，此过程不需要考虑字符间的距离；此外，由于汉字的术语字典信息和部首信息包含更深层次的语义信息，所以将术语字典和部首的特征纳入模型以提高模型的性能。最后，运用CRF解码预测标签。实验结果表明所提模型在CCKS2017和CCKS2021数据集上的*F*1值分别达到了96.22%和84.65%，优于当前主流的命名实体识别模型，具有更好的识别效果。

关键词：中文命名实体识别；部首特征；Transformer；BERT
中图分类号：TP391 **文献标识码**：A

Research on Chinese Clinical Named Entity Recognition Method based on Radical Feature and BERT-Transformer-CRF

YAO Lei, JIANG Mingfeng, FANG Xian, WEI Bo, LI Yang

(*School of Computer Science and Technology*, *Zhejiang Sci-Tech University*, *Hangzhou* 310018, *China*)
✉202030504181@mails.zstu.edu.cn; m.jiang@zstu.edu.cn; xianfang@zstu.edu.cn;
weibo@zstu.edu.cn; yangli@zstu.edu.cn

Abstract: In Chinese Clinical Named Entity Recognition (CNER), Chinese text lacks separators to delineate word boundaries, and some existing methods are difficult to capture the long-distance interdependent features. This paper proposes a pre-trained BERT-Transformer-CRF method to realize CNER. Firstly, BERT (Bidirectional Encoder Representation Transformer) is applied to extract text features. Then, Transformer is utilized to capture the dependencies between characters regardless of the distance between characters. In addition, as term dictionary and radical information of Chinese characters contain deeper semantic information, the features of term dictionary and radicals are incorporated into the model to improve its performance. Finally, CRF (Conditional Random Field) is applied to decode predicted labels. The experimental results show that *F*1 values of the proposed model on CCKS2017 and CCKS2021 datasets reach 96.22% and 84.65% respectively, which is superior to the current mainstream named entity recognition model and has better recognition effect.

Keywords: Chinese Clinical Named Entity Recognition; radical feature; transformer; BERT

1 引言(Introduction)

近年来，随着网络技术和信息系统的不断发展和完善，医疗系统产生的医疗数据将急剧增加。电子病历是指医务人员在开展医疗活动的过程中，使用信息系统生成的数字化资料，一般包括文字、图表、数据、符号、图形和影像[1]。电子病历中涉及大量的文字信息，中文电子病历命名实体识别作为重要的中文电子医疗数据信息抽取任务，普遍受到研究人员的重视。

命名实体识别是指从非结构化或半结构化的文本数据中提取实体，并将检测到的实体归类至预先定义好的一类中。其中，电子病历命名实体识别的主要目的是识别与分类医疗记录中的临床术语，包括实验室检验、手术和药物等。例如，某份电子病历中的“患者缘于2小时前无明显诱因出现左腹部疼痛……”，其中“左腹部”属于身体部位实体，“疼痛”属于症状和体征实体，“门诊经泌尿系超声”中的“泌尿系超声”为检查实体。命名实体识别研究成果能够为构建医学知识库、绘制信息抽取和知识图谱等后续的临床研究提供支撑。然而，手动抽取实体信息会消耗较大的时间和人力成本，因此很多研究者采用自然语言处理技术解决以上问题。

目前，国外关于命名实体识别的绝大部分研究均基于英文[2]，而国内对于中文电子病历命名识别的研究尚处于初期阶段，没有建设全面的体系结构。主要原因是中文电子病历命名实体识别的规范和标准无法达成统一，中文电子病历文本中实体没有自然的分隔符，而且医疗实体的组成较为复杂。因此，本文将利用部首信息和术语字典开展命名实体识别研究任务。

针对中文电子病历命名实体识别任务，本文使用BERT获得输入的信息向量表示，结合部首级特征表示，以此作为Transformer模块的输入。Transformer模块对上下文长距离的位置依赖特征进行提取，并在CRF模块中对上下文标注进行约束，最终输出序列标注结果。本文提出的方法在CCKS2017和CCKS2021数据集上广泛评估了其可用性和实用性。

本文提出一种基于BERT−Transformer−CRF(BTC)的中文电子病历命名实体识别方法；其主要工作原理归纳如下：(1)鉴于句子中的实体间存在依赖关系，本文使用Transformer获得更好的上下文特征表示，从而捕捉字符之间的长距离依赖关系。其中，多头注意力机制可以直接捕捉角色之间的依赖关系，解决了一般深度模型随着实体间距离增加长期依赖能力下降的问题。(2)本文通过添加部首特征，并将部首信息和深度学习模型相结合，解决了移植一般深度学习模型导致的医疗实体识别性能差的问题。(3)本文在真实的电子病历语料上验证模型的效果，实验结果表明，BTC模型在CCKS2017和CCKS2021数据集上具有良好的性能优势。

2　命名实体识别相关工作(Related work of named entity recognition)

命名实体识别是从大规模非结构化文本数据中提取具有实际意义的实体[3]，主要分为基于词典和规则、基于统计和基于神经网络的命名实体识别方法。

基于规则的方法依赖于一个庞大且全面的领域字典，需要领域专家手动构建规则和模板[4]，无法在不同领域间复用。在统计的经典机器学习方法中，隐马尔科夫模型(Hidden Marko Model，HMM)[5]、支持向量机(Support Vector Machine，SVM)[6]、条件随机场(Conditional Random Field，CRF)[7]得到广泛应用。例如，扈应等[8]提出一种结合CRF的边界组合命名实体识别方法，有效地利用了生物医学实体特征。统计的机器学习方法需要设计特征模板提取特征，实体识别效果易受到构建的特征集合的影响。

近年来，基于神经网络的深度学习方法广泛运用于自然语言处理领域。WU等[9]在双向长短期记忆(Bidirectional Long Short−Term Memory，BiLSTM)网络后引入自注意力机制，并提出一种新的细粒度字符级表示方法用于获取更多的汉字语义信息。LI等[10]提出BERT−BiLSTM−CRF模型，在未标记的中文临床记录上预训练BERT模型增强语义信息，利用BiLSTM和CRF等不同层次提取文本特征和解码预测标签。YIN等[11]使用自注意力捕捉字符间的相关性，在CCKS2017和TP_CNER数据集中的$F1$评分分别达到93.00%和86.34%。KONG等[12]通过构建多层次卷积神经网络(CNN)融合长短期信息，设计一种注意力机制获取全局上下文信息。YA等[13]提出XLNET−BiLSTM−CRF模型，利用预训练的XLNET提取句子特征。QIU等[14]提出带有条件随机场的RD−CNN−CRF模型解决通过时间传播的隐形激活矢量导致训练时间过长的问题，残差卷积神经网络用来捕捉相邻标签间的依赖关系。然而，长短期记忆网络(Long Short−Term Memory，LSTM)和CNN捕捉字符长期依赖关系的能力将随着实体间距离的增加而下降。这些模型没有充分考虑医学领域数据信息的特点，在医学实体识别方面的效果不佳。

相较于传统的Word2vec、Glove、ELMO词向量方法[15]，BERT能获得更好的字符嵌入表示。除了单词嵌入，其他一些特征对于提升命名实体识别效果也有帮助。电子病历文本中有海量的医学专业术语，高质量的医学术语词典对于提取医疗领域知识的特征非常有用。因此，本文将部首特征和术语字典特征与深度学习模型相融合。

3　BERT-Transformer-CRF模型(The model of BERT-Transformer-CRF)

本文提出的BERT−Transformer−CRF模型架构主要由BERT、Transformer、CRF三个模块组成，模型架构如图1所示，其基本思想是通过微调的BERT模型进行预训练提取输入的信息特征，将字向量表示序列和部首特征嵌入连接起来，加入字典特征，通过Transformer层获得长距离的文本依赖，在CRF模块中对上下文标注约束进行解码，最终输出序列结果。

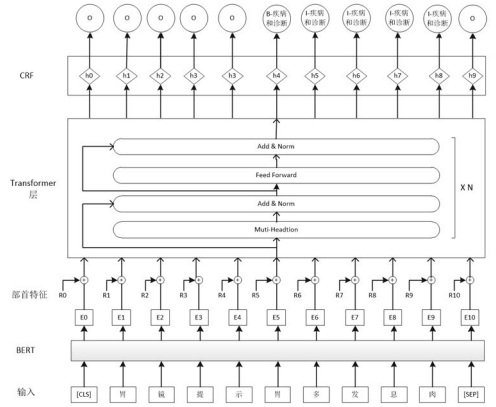

图1 BERT−Transformer−CRF+radical 模型架构图

Fig.1 Model architecture diagram of BERT−Transformer−CRF+radical

3.1 中文预训练模型BERT

BERT模型的主要创新点在于使用掩码语言模型(Mask Language Model，MLM)获取字符级特征表示和下一句预测进行预训练[16]，学习到的先验语义知识通过微调被应用到下游任务中。这样得到的向量不仅包含隐含的上下文信息，还能够更彻底地捕捉句子中的双向关系。谷歌官方发布的BERT-base及其中文版本并没有在中文临床领域进行预训练。北京大学国际数学研究中心发布了基于中文临床语料库的预训练模型[10]，本研究使用临床文本生成的PyTorch版本的预训练BERT模型，该模型从网络上爬取1.05 G临床文本，从现有的BERT检查点开始，在BERT的原始词汇表增加46 个字符并在特定域上进行预训练。本文模型从预训练好的BERT模型获得字符级的增强语义信息，结合部首信息输入Transformer层中。

3.2 部首特征

近些年，部首特征被广泛运用于命名实体识别任务中[17]。汉字是象形文字和方块字，它们有更深次的语义隐含在汉字内部。偏旁部首“月”通常与身体部位有关[18]，比如“肺”“肝”“脑”是用来代表人体器官的。“疒”通常与疾病和诊断有关，“口”通常出现在症状实体中。然而，目前主流的命名实体识别方法不能将预先训练好的模型与中文部首信息相结合。本研究从在线新华字典获取汉字的偏旁部首组成，它以“字符-偏旁部首”的形式生成一个键值对字典。部首信息编码与BERT获得字向量编码叠加融合到深度学习网络中。

本文引入字典信息提升命名实体识别的效果。针对药物、手术等术语字典，利用双向最大匹配算法[19]在文本中找到对应实体。具体来说，利用双向最大匹配算法分割文本和标注出现在字典中的实体，如果文本可以通过该算法被标注为第j 个标签，则通过向该元素添加常数修改线性层的输出，以此提高识别效果。

3.3 Transformer层

Transformer是一种全联接的多头自注意力神经网络模型。面向机器翻译等任务的Transformer[20]由编码组件、解码组件及它们中间的连接组成。本文提出的模型仅使用其中的编码器进行医疗文本序列的长距离位置依赖关系特征建模。

Transformer的编码器由多头注意力和前馈神经网络组成。如图2所示，BERT获得的字符嵌入与部首嵌入进行结合，输出信息进入Transformer层，与位置嵌入进行拼接得到$X_{\text{embedding}}$，作为多头注意力机制的输入。多头注意力由多个自注意力拼接组成。多头注意力结构由中心块的若干线性变换和点积注意力组成。Attention的工作原理如下：给定输入的$X_{\text{embedding}}$向量$\alpha^i \in R^{d_1 \times 1}$，然后读入输入向量通过矩阵$W^q \in R^{d_k \times d_l}$，$W^k \in R^{d_k \times d_l}$，$W^v \in R^{d_k \times d_l}$ 进行线性变换得到Query向量$(Q)\, q^i \in R^{d_k \times 1}$，Key向量$(K)k^i \in R^{d_k \times 1}$，以及Value向量$(V)v^i \in R^{d_k \times 1}$：

$$q^i = W^q \cdot \alpha^i \tag{1}$$

$$k^i = W^k \cdot \alpha^i \tag{2}$$

$$v^i = W^v \cdot \alpha^i \tag{3}$$

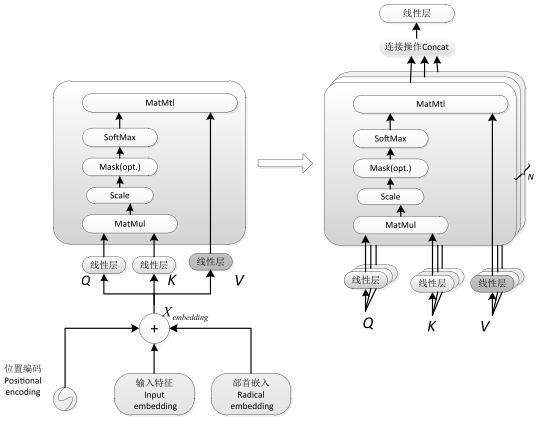

图2 多头注意力结构

Fig.2 Multi-head attention structure

接着利用得到的Q和K，使用点积法计算输入序列的相关性得分。对相关性得分进行归一化，使得训练时梯度稳定。经过Softmax函数将得分向量转化为[0,1]之间的概率分布并与V进行点积。如式(4)所示，令输出矩阵$B=(b^1,b^2,\cdots,b^n)\in R^{d_k\times n}$，则：

$$B^i = \text{Attention}(Q^i,K^i,V^i) = V^i \cdot \text{Softmax}\left(\frac{K^T \cdot Q}{\sqrt{d^k}}\right) \tag{4}$$

$$\text{MultiHead}(Q,K,V) = \left[B^1,B^2,\cdots,B^n\right]W_0 \tag{5}$$

$\text{MultiHead}(Q,K,V)$与$X_{\text{embedding}}$进行残差连接到L，对同一层的所有神经元进行归一化使其满足正态分布，从而可以更好地解决梯度消失和权重矩阵退化问题。如式(6)所示，编码器中的前馈神经网络层经过线性变换将L映射到一个更大维度的特征空间，使用ReLu激活函数引入非线性进行筛选，最后经过线性变换回到原始维度。

$$F(X) = \max\left(0,\ XW_1 + b_1W_2 + b_2\right) \tag{6}$$

$F(X)$与L再次进行残差连接和归一化，构造出一个编码器，叠加多个编码器，最终得到Transformer层的输出。

3.4 CRF层

CRF是自然语言处理的基础模型，广泛运用于序列标注模型，对上下文标注进行约束使得正确的输出标签最大化。在中文电子病例命名实体识别任务中，输出标注之间存在强相关性，相邻的标签之间有依赖关系，例如标签“B-疾病和诊断”不能跟在“I-疾病和诊断”之后。Transformer输出的向量只考虑了上下文之间的长距离依赖关系，没有考虑标签之间的顺序，而CRF层自动学习句子的约束条件，所以引入条件随机场解决这一问题。对于给定的输入$x=(x_0,x_1,\cdots,x_n)$，CRF通过Softmax函数运用随机条件概率预测输出向量$y=(y_0,y_1,\cdots,y_n)$标签序列Y的得分：

$$s(H,Y) = \sum_{i=1}^{N} P_{i,y_i} + \sum_{i=1}^{N} T_{y_{i-1},y_i} \tag{7}$$

Transformer模块生成输出序列$H=\{h_1,h_2,\cdots,h_3\}$，输入CRF模块中进行解码。T是CRF的转移分数矩阵，其中$T_{i,j}$是指标签i转移到标签j。使用维特比算法获得全局最佳预测的标签序列，如公式(8)所示：

$$y^* = \arg\max\left(S(x,y)\right) \tag{8}$$

维特比算法可以通过动态规划算法获得最优路径。

4　实验和结果(Experiment and result)

实验采用Python 3.8语言开发，软件模型基于PyTorch深度学习框架，采用Adam作为优化器，学习率为$2e^{-5}$，批处理大小为8，epochs迭代次数为15 次，*max_seq_length*=480，*hidden_size*为768，*dropout*为0.1，部首特征向量维度为20。硬件采用2 块NVIDIA GeForce RTX 3090显卡训练。

4.1　数据集

本研究的数据集源自2017 年全国知识图谱与语义计算大会(CCKS2017)和2021年全国知识图谱与语义计算大会(CCKS2021)。数据集包含实际的电子病历数据，由专业医学领域团队手工进行注释。首先对数据进行预处理，采用NER领域的标准标注策略BIO，“B”表示医疗实体的起始位置，“I”表示医疗实体的中间部分，“O”表示与医疗实体无关的部分。

在CCKS2017数据集中，有四种类型的电子病历，包括一般项目、病史特点、诊疗经过及出院记录，共有五类命名实体：DISEASE(疾病和诊断)、SIGNS(症状和体征)、CHECK(检查和检验)、BODY(身体部位)及TREATMENT(治疗)。由于本文的研究人员没有参加比赛，所以获得的数据集不完整。训练集有960 条临床记录，测试集包含120 条临床记录。表1列出了不同类别实体的统计数据；各实体分布比例如图3所示。

在CCKS2021数据集中，有1,150 条临床记录，按8∶2的比例划分训练集和测试集，共有实验室检验、药物、影像检查、解剖部位、疾病和诊断及手术6 类命名实体。表2列出了不同类别实体的统计数据；图4显示各实体分布比例。

表1 CCKS2017不同类别医疗实体统计

Tab.1 Statistics of different types of medical entities on CCKS2017 dataset

数据集	训练集	测试集
身体部位	9,116	402
检查和检验	8,483	389
疾病和诊断	691	30
症状和体征	7,753	167
治疗	808	381

表2 CCKS2021不同类别医疗实体统计

Tab.2 Statistics of different types of medical entities on CCKS2021 dataset

数据集	训练集	测试集
实验室检验	8,811	1,346
手术	923	124
药物	1,935	277
影像检查	1,002	138
疾病和诊断	4,345	607
解剖部位	1,002	138

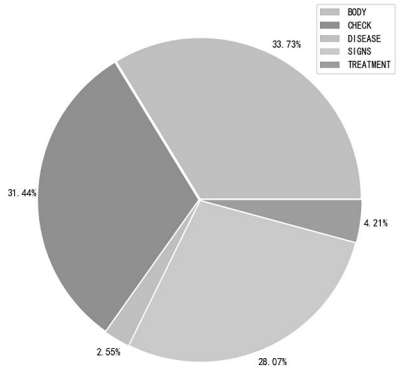

图3 CCKS2017各类别实体分布

Fig.3 Distribution of various entities on CCKS2017 dataset

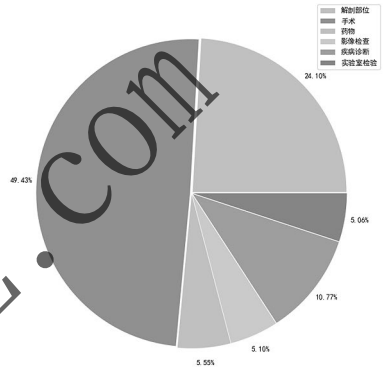

图4 CCKS2021各类别实体分布

Fig.4 Distribution of various entities on CCKS2021 dataset

4.2　评价指标

本实验采用命名实体识别通用的评价指标正确率*P*(Precision)、召回率*R*(Recall)、*F*1值(F−measure)对电子病例命名实体识别结果进行性能衡量，其计算公式分别如下：

$$P = \frac{TP}{TP + FP} \tag{9}$$

$$R = \frac{TP}{TP + FN} \tag{10}$$

$$F1 = \frac{2PR}{P + R} \tag{11}$$

式(9)—式(11)中，*TP*为识别正确的实体词数，*FP*为实体识别正确但类别或者边界判定出现错误，*FN*为应该被识别但实际没有被识别的医疗实体的数量。

4.3　结果和讨论

4.3.1　实验结果比较与分析

为了验证BTC模型在CCKS2017数据集的有效性，与其他的命名实体识别方法进行比较与分析，实验结果如表3所示。BiLSTM+CRF通过Word2vec训练单词嵌入。CRF保证了标签序列之间的顺序，BERT+CRF与BiLSTM+CRF两个基线模型进行对比，实验表明BERT能获得更好的特征表示。为了验证预训练BERT的有效性，进行BERT+BiLSTM+CRF和BiLSTM+CRF的对比实验，结果表明，在CCKS2017数据集中引入在临床预训练好的BERT模型后，*P*、*R*、*F*1值分别提高了4.98%、1.8%、3.41%。为了证明Transformer的有效性，进行BTC和BERT+BiLSTM+CRF的对比实验。在

CCKS2017数据集上，*P*、*R*、*F*1值分别提高了4.4%、4.99%、4.69%，这优于其他现有技术方法。充分表明Transformer中多头注意力机制使其具有较强的长距离依赖关系表征能力，能获得更好的上下文表示。除此以外，本文将术语字典和部首特征等应用于微调的BERT模型，Transformer能获得丰富的语义信息，*F*1提高了0.22%。

除了观察整个测试数据集的评估指标，本研究仔细观察了预测的结果。在BTC模型的基础上，添加部首特征信息并应用字典信息后处理，模型在不同类型和临床实体上的性能如表4所示。在CCKS2017数据集中，在检查和检验、疾病和诊断、症状和体征及治疗方面取得了较好的效果，但未能有效识别身体部位特征。检查结果发现：预测实体遗漏了一些部位，例如正确实体为“头皮”“咽部”“右侧顶骨”“头颅”“右侧上下肢”，模型则预测为“头”“咽”“右侧顶”“头”“右侧上下”，部分部位遗漏；其次，类别标注错误，如检查类实体“查体双肺呼吸音粗”，模型提取“双肺部位”，识别类别错误。

表3 CCKS2017模型对比实验

Tab.3 Model comparison experiment on CCKS2017 dataset

模型	*P*/%	*R*/%	*F*1/%
BiLSTM+CRF（基线一）	86.66	89.18	87.90
BERT+CRF（基线二）	91.50	89.94	90.71
BERT+BiLSTM+CRF	91.64	90.98	91.31
BTC	**96.04**	**95.97**	**96.00**
BTC+radical+dictionary	**96.42**	**96.22**	**96.22**

表4 CCKS2017各类型实体的识别比较

Tab.4 Recognition and comparison of different types of entities on CCKS2017

实体类型	*P*/%	*R*/%	*F*1/%
身体部位	84.67	84.04	84.36
检查和检验	98.97	98.97	98.97
疾病和诊断	98.97	99.48	99.48
症状和体征	99.81	99.40	99.10
治疗	99.21	99.21	99.21
总体	96.42	96.22	96.22

本文还在CCKS2021数据集上做了实验，实验结果如表5所示。同样，选取BiLSTM+CRF与BERT+CRF对比，*F*1提高了3.52%，表明引入临床语料上预训练的BERT模型优于BiLSTM模型。BTC相较于BERT+BiLSTM+CRF在*P*、*R*、*F*1上分别提升1.01%、0.59%、0.81%；结果表明，Transformer能够获得长距离依赖关系的能力优于BiLSTM。相较于基线BiLSTM−CRF，本文的*F*1提高了4.39%。本文引入药物和手术等词典，利用字典信息后处理的方法修改线性层的输出。引入外部部首特征和术语词典，总体*P*、*R*、*F*1值提升了0.03%、0.6%、0.3%，方法受限于领域词典的质量，结果不显著。各实体的*P*、*R*、*F*1值如表6所示。在CCKS2021数据集中，疾病和诊断及实验室检验识别效果较差：第一，预测实体的位置是缺失或者冗余的。比如，预测的实体是“口腔”“子宫”，而正确的实体识别是“口腔溃疡”“子宫内膜分段诊刮”。第二，标注的命名实体识描述过长，预测的命名实体通常较短，比如“宫内孕不全流产”，模型预测的实体为“宫内孕不全”。第三，预测的实体存在错误的标注，比如“结肠癌病变”正确类别为疾病和诊断，而模型预测的结果提取“结肠”，该实体识别的类别为解剖部位，识别结果错误。

表5 CCKS2021模型对比实验

Tab.5 Model comparison experiment on CCKS2021 dataset

模型	*P*/%	*R*/%	*F*1/%
BiLSTM+CRF（基线一）	80.17	80.35	80.26
BERT+CRF（基线二）	83.31	84.29	83.78
BERT+BiLSTM+CRF	82.77	84.33	83.54
BTC	83.78	84.92	84.35
BTC+radical+dictionary	**83.81**	**85.52**	**84.65**

表6 CCKS2021各类型实体的识别比较

Tab.6 Recognition and comparison of different types of entities on CCKS2021

实体类型	*P*/%	*R*/%	*F*1/%
解剖部位	83.08	84.25	83.66
手术	84.13	85.48	84.80
药物	93.48	93.14	93.31
影像检查	86.21	90.85	88.34
疾病和诊断	81.22	81.22	81.22
实验室检验	81.04	85.07	83.01
总体	84.86	86.62	85.71

4.3.2 与现有方法比较

CCKS2017数据集其他的测试结果可以在表7中看到，体现了本文最佳模型和最先进的深度模型之间的比较结果。QIN等[21]在中文电子病历领域提出了一个基于RoBERTa−BiGRU−CRF的命名实体识别方法，将其应用于脑血管疾病领域，通过将电子病历转化为低维向量输入BiGRU层捕获上下文特征，总体*F*1值达到90.38%。罗熹等[22]提出一种融合领域词典的字符级表示方法，结合多头注意力机制和BiLSTM−CRF捕捉字符间的潜在依赖权重、语义和结构特征等多方面特征。WU等[23]利用RoBERTa中的全词掩码获取词向量表示，同时通过BiLSTM捕捉提取部首信息后捕捉特征的内在关联性，并拼接RoBERTa生成的特征向量。李丹等[24]设计BiLSTM与CRF的联合模型并引入BERT模型，预测的时候考虑了上下文特征，同时将部首信息与字向量编码相结合，利用部首信息在标签矩阵中加入部首以修改CRF层得分函数，*F*1分数可以增加到93.81%。张云秋等[25]将RoBERTa−wwm中的各编码层生成的语义表示进行动态融合，BiLSTM层用来捕获序列信息，再输入条件随机场保证各标签之间的顺序关系。实验表明，本文模型在CCKS2017数据集上取得了96.22%的精度，相较于其他模型，总体识别精度提高了2.14%—5.84%，优于其他模型。

表7 与CCKS2017现有的深度模型比较

Tab.7 Comparison with the existing deep model on CCKS2017 dataset

团队	模型	*F*1/%
QIN et al[21]	RoBERTa-BiGRU-CRF	90.38
罗熹等[22]	MHA-BiLSTM-CRF	91.97
WU et al[23]	RoBERTa+RC	93.26
李丹等[24]	BERT+BiLSTM-CRF+dictionary	93.81
张云秋等[25]	RoBERTa-wwm-BiLSTM-CRF	94.08
本文	**BTC+racial+dictionary**	**96.22**

从表8可以看出，BTC+radical+dictionary模型取得了总体*F*1值为84.65%的成绩；而BiLSTM+CRF、BERT+CRF、BERT+BiLSTM+CRF的总体*F*1值分别为80.26%、83.78%、83.54%。相较于以上模型，本文的*F*1值分别提高了4.39%、0.87%、1.11%，充分证明了本文模型的有效性。BERT+CRF与BiLSTM+CRF相比，预训练BERT获取字向量特征时具有非常好的并行性质，总体识别的精度为83.78%。将BERT+BiLSTM+CRF与BiLSTM+CRF进行对比发现，*F*1提高了3.28%，充分说明在临床语料库预训练的BERT模型能获得更加丰富的语义特征，证明预训练BERT模型的有效性。BTC相比BERT+BiLSTM+CRF，各实体识别精度均有提升，总体识别精度提升了0.81%，表明Transformer获得实体间长距离依赖能力优于BiLSTM。引入部首特征和手术、药物术语词典，模型的性能进一步提升，但受限于词典的质量，结果提升不显著，在解剖部位、药物、疾病和诊断、实验室检验实体类别中识别的精确度分别提升了0.52%、0.14%、0.24%、1.23%。本文模型在解剖部位、手术、影像检查三种实体的识别效果均为最好。同时，从表8中可以发现，BiLSTM+CRF模型在疾病和诊断、实验室检查中取得了最好的成绩，即88.48%、87.91%，这证明了有些模型即使总体*F*1值并非最高，在特定实体上也能获得出色的性能。

表8 不同模型在CCKS2021数据集上对每个实体和整体的*F*1值的预测结果

Tab.8 Prediction results of different models on CCKS2021 dataset for *F*1 value of each entity and the whole

实体类别	*F*1/%				
	BiLSTM+CRF	BERT+CRF	BERT+BiLSTM+CRF	BTC	BTC+radical+dictionary
解剖部位	81.55	83.63	83.53	83.66	**84.18**
手术	76.55	83.46	82.21	**84.80**	83.46
药物	79.66	**93.60**	93.28	93.31	93.45
影像检查	81.10	84.10	85.92	**88.34**	87.14
疾病和诊断	**88.48**	79.77	79.02	81.22	81.46
实验室检验	**87.91**	83.37	82.96	83.01	84.24
总体	80.26	83.78	83.54	84.35	**84.65**

图5表明本文模型对身体部位类别的识别效果较差。根据图6的混淆矩阵显示，本文的模型对影像检查、手术、解剖部位实体识别效果较好。本文的最佳模型对实体的中间部位识别率较高，起始部位识别有偏差。其中，手术的起始位置可能被划分为手术的中间位置，比如“小肠切除术”被错误划

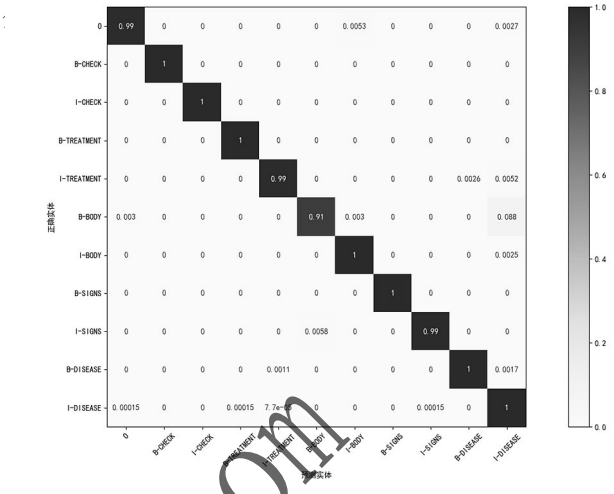

图5 CCKS2017实体类别混淆矩阵

Fig.5 Entity category confusion matrix on

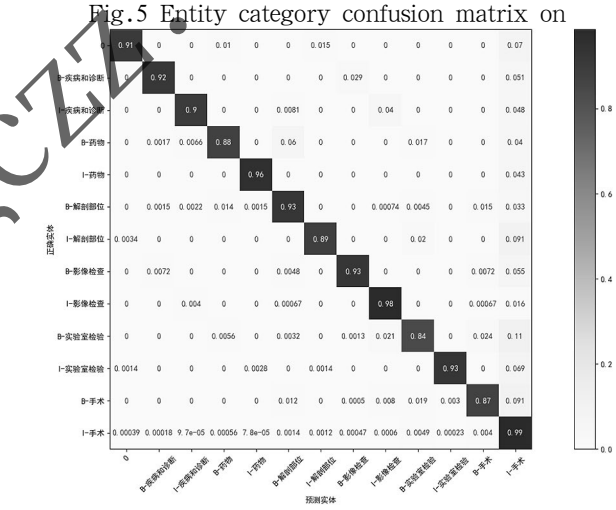

图6 CCKS2021实体类别混淆矩阵

Fig.6 Entity category confusion matrix on CCKS2021 dataset

5 结论(Conclusion)

本文提出一种基于微调的BERT-Transformer-CRF实体识别模型。该模型通过在中文临床语料库预训练的BERT获得字符级的增强语义信息，与部首语义信息融合输入Transformer。Transformer中的多头注意力机制和前馈神经网络能够更好地捕捉字符间的长距离依赖关系，CRF能保证临近标签间的顺序关系，有效提升了医疗领域命名实体的识别能力。同时，添加手术、药物等术语字典特征进一步提升性能。实验结果表明，该模型能有效识别手术、影像检查、解剖部位等领域实体，在CCKS2017和CCKS2021数据集中获得96.22%和84.65%的*F*1值，优于现有模型的结果。在未来工

作中，考虑扩充学习高质量的领域词典和构建更大规模的语料库，可以将其应用到医学命名实体的信息抽取和医疗知识图谱的构建等后续工作中。

参考文献(References)

[1] 陈杰,奚雪峰,皮洲,等.基于ALBERT的中文医疗病历命名实体识别[J].南京师范大学学报(工程技术版),2021,21(01):36-43.

[2] 任明,许光,王文祥.家谱文本中实体关系提取方法研究[J].中文信息学报,2020,34(6):45-54.

[3] 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(Z1):251-262.

[4] 刘峰,高赛,于碧辉,等.基于Multi-head Attention和Bi-LSTM的实体关系分类[J].计算机系统应用,2019,28(6):118-124.

[5] DROVO M D, CHOWDHURY M, UDAY S I, et al. Named entity recognition in bengali text using merged hidden markov model and rule base approach[C]// WANG Y L. Proceedings of 2019 7th International Conference on Smart Computing & Communications (ICSCC). New York: IEEE, 2019:1-5.

[6] LAKSHMI G, PANICKER J R, MEERA M. Named entity recognition in Malayalam using fuzzy support vector machine[C]// XU S. Proceedings of 2016 international conference on information science (icis). Piscataway: IEEE, 2016:201-206.

[7] LEI J, TANG B, LU X, et al. A comprehensive study of named entity recognition in Chinese clinical text[J]. Journal of the American Medical Informatics Association, 2014, 21(5): 808-814.

[8] 扈应,陈艳平,黄瑞章,等.结合CRF的边界组合生物医学命名实体识别[J]. 计算机应用研究,2021,38(07):2025-2031.

[9] WU G, TANG G, WANG Z, et al. An attention-based BiLSTM-CRF model for Chinese clinic named entity recognition[J]. IEEE Access, 2019, 7:113942-113949.

[10] LI X, ZHANG H, ZHOU X. Chinese clinical named entity recognition with variant neural structures based on BERT methods[J]. Journal of biomedical informatics, 2020, 107:103422.

[11] YIN M, MOU C, 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.

[12] KONG J, ZHANG L, JIANG M, et al. Incorporating multi-level CNN and attention mechanism for Chinese clinical named entity recognition[J]. Journal of Biomedical Informatics, 2021, 116:103737.

[13] YAN R, JIANG X, DANG D. Named entity recognition by using XLNet-BiLSTM-CRF[J]. Neural Processing Letters, 2021, 53(5):3339-3356.

[14] QIU J, WANG Q, ZHOU Y, et al. Fast and accurate recognition of Chinese clinical named entities with residual dilated convolutions[C]// HU X H. Proceedings of 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM). New York: IEEE, 2018:935-942.

[15] 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.

[16] ZHENG L, HONG Y, ZHEN G, et al. Medical named entity recognition based on multi feature fusion of BERT[C]// SU S B. Proceedings of 2021 4th International Conference on Big Data Technologies. New York: Association for Computing Machinery, 2021:86-91.

[17] PENG H, CAMBRIA E, ZOU X. Radical-based hierarchical embeddings for Chinese sentiment analysis at sentence level[C]// MARKOV Z. Proceedings of The Thirtieth International Flairs Conference. Palo Alto: AAAI Press, 2018, 148:167-176.

[18] 崔少国,陈俊桦,李晓虹.融合语义及边界信息的中文电子病历命名实体识别[J].电子科技大学学报,2022,51(04):565-571.

[19] GAI R, GAO F, DUAN L M, et al. Bidirectional maximal matching word segmentation algorithm with rules[J]. Advanced materials research, 2014, 926:3368-3372.

[20] 李韧,李童,杨建喜,等.基于Transformer-BiLSTM-CRF的桥梁检测领域命名实体识别[J].中文信息学报,2021,35(04):83-91.

[21] QIN Q, ZHAO S, LIU C. A BERT-BiGRU-CRF model for entity recognition of Chinese electronic medical records[J]. Complexity, 2021, 2021:6631837.

[22] 罗熹,夏先运,安莹,等.结合多头自注意力机制与BiLSTM-CRF的中文临床实体识别[J].湖南大学学报(自然科学版),2021,48(04):45-55.

[23] WU Y, HUANG J, XU C, et al. Research on named entity recognition of electronic medical records based on RoBERTa and radical-Level feature[J]. Wireless Communications and Mobile Computing, 2021, 2021:2489754.

[24] 李丹,徐童,郑毅,等.部首感知的中文医疗命名实体识别[J].中文信息学报, 2020,34(12):54-64.

[25] 张云秋,汪洋,李博诚.基于RoBERTa-wwm动态融合模型的中文电子病历命名实体识别[J].数据分析与知识发现,2022,6(Z1):242-250.

作者简介：

姚 蕾(1997-)，女，硕士生.研究领域：自然语言处理，人工智能.

蒋明峰(1977-)，男，博士，教授.研究领域：深度学习与优化方法，计算机图像处理.

方 贤(1994-)，男，博士，讲师.研究领域：目标检测，聚类分析，数据挖掘.

魏 波(1983-)，男，博士，讲师.研究领域：优化算法理论，人工智能.

李 杨(1986-)，男，博士，讲师.研究领域：模式识别，深度学习.