doi:10.3969/j.issn.1673-9833.2020.04.009

基于 BERT 的医疗电子病历命名实体识别

梁文桐[1, 2]，朱艳辉[1, 2]，詹　飞[1, 2]，冀相冰[1, 2]

（1. 湖南工业大学 计算机学院，湖南 株洲 412007；
2. 湖南工业大学 智能信息感知及处理技术湖南省重点实验室，湖南 株洲 412007）

摘　要：针对中文医疗电子病历命名实体识别中，传统的字或词向量无法很好地表示上下文语义以及传统 RNN 并行计算能力不足等问题，提出了一个基于 BERT 的医疗电子病历命名实体识别模型。该模型中的 BERT 预训练语言模型可以更好地表示电子病历句子中的上下文语义，迭代膨胀卷积神经网络（IDCNN）对局部实体的卷积编码有更好的识别效果，多头注意力（MHA）多次计算每个字和所有字的注意力概率以获取电子病历句子的长距离依赖。实验结果表明，BERT-IDCNN-MHA-CRF 模型能够较好地识别电子病历中的医疗实体，模型的精确率、召回率和 F_1 值相比于基线模型分别提高了 1.80%, 0.41%, 1.11%。

关键词：电子病历；命名实体识别；BERT；膨胀卷积神经网络；多头注意力

中图分类号：TP391　　**文献标志码**：A　　**文章编号**：1673-9833(2020)04-0054-09

引文格式：梁文桐，朱艳辉，詹　飞，等 . 基于 BERT 的医疗电子病历命名实体识别 [J]. 湖南工业大学学报，2020，34(4)：54-62.

Named Entity Recognition of Electronic Medical Records Based on BERT

LIANG Wentong[1, 2]，ZHU Yanhui[1, 2]，ZHAN Fei[1, 2]，JI Xiangbing[1, 2]

（1. College of Computer Science，Hunan University of Technology，Zhuzhou Hunan 412007，China；
2. Hunan Key Laboratory of Intelligent Information Perception and Processing Technology，
Hunan University of Technology，Zhuzhou Hunan 412007，China）

Abstract：In view of the poor performance exhibited by traditional words or word vectors in expressing context semantics, as well as the insufficiency of traditional RNN parallel computing ability in Chinese medical EMR named entity recognition, a named entity recognition model of medical EMR based on Bert has thus been proposed. In this model, the BERT pre-training language model can better represent the context semantics in electronic medical records, with the iterative expanded convolutional neural network (IDCNN) characterized with a better recognition effect on convolutional coding of local entities, and with the multiple head attention (MHA) computing the attention probability of each word and all words for many times to obtain the long-distance dependence of EMR sentences. The experimental results show that the BERT-IDCNN-MHA-CRF model can better identify medical entities in electronic medical records, and compared with the baseline model, the precision, recall and F_1 values of the model are increased by 1.80% , 0.41%

收稿日期：2019-10-09

基金项目：科技创新 2030——“新一代人工智能”基金资助重大项目（2018AAA0100400），国家自然科学基金资助项目（61702177），湖南省自然科学基金资助项目（2018JJ2098，2020JJ6089），湖南省教育厅基金资助重点项目（19A133）

作者简介：梁文桐（1996-），男，山东枣庄人，湖南工业大学硕士生，主要研究方向为自然语言处理和知识工程，E-mail：wtliang625@163.com

通信作者：朱艳辉（1968-），女，湖南湘潭人，湖南工业大学教授，主要从事自然语言处理和知识工程方面的教学与研究，E-mail：swayhzhu@163.com

and 1.11% respectively.

Keywords：electronic medical record；named entity recognition；BERT；iterated dilated convolution neural network；multi-head attention

1　研究背景

在自然语言处理（natural language processing，NLP）的任务中，命名实体识别（named entity recognition，NER）是具有挑战的基础性工作 [1]。从狭义上来说，一般的命名实体识别任务的目的，是从文本中识别出 3 种类型的实体提及，包括人名、地名和机构名。在医学领域中，医务工作人员通过医疗机构信息系统将病人的临床诊断信息存储在计算机中，得到电子病历（electronic medical records，EMR）。电子病历命名实体识别是命名实体识别在电子病历文本分析研究中的重要应用和扩展，其目的是自动地识别并且分类电子病历中的医疗命名实体。这些命名实体对象能够被用于后续医疗电子病历信息的分析和研究中，比如构建临床信息决策系统、构建医疗领域的知识图谱等。

早期的电子病历命名实体识别方面的研究主要运用基于词典和规则的方法，仅仅依赖于现有的词典和手工编辑的规则来识别医疗命名实体 [2]。后来，基于统计机器学习的方法被运用到电子病历命名实体识别中。如于楠等 [3] 采用基于多特征融合的 CRF（conditional random fields）模型进行了中文电子病历 NER 的研究。A. Kulkarni[4] 从生物医学文本中完成 DNA、RNA 和蛋白质等生物医学术语的识别，该任务使用 CRF 统计模型完成。许源等 [5] 基于 CRF 以及 RUTA（rule-based text annotation）规则，建立了一个医学命名实体识别模型，该模型在识别脑卒中患者入院记录的医学命名实体时取得了良好的效果。王润奇等 [6] 利用半监督学习方法，将 Tri-Training 算法进行了改进，使得中文电子病历实体识别模型的效果得到了提升。

近年来，随着硬件计算能力的大幅度提高，基于深度神经网络的方法已经被成功地应用到电子病历命名实体识别中，该方法是一种端到端的方法，不需要特殊的领域资源（如词典）或者构建本体，可以从大规模的标注数据中自动地学习和抽取文本特征。在电子病历 NER 任务中，杨红梅等 [7] 利用一种基于 BiLSTM（bidirectional long short-term memory）与 CRF 的实体识别模型，抽取了入院记录和出院小结中的医学命名实体。万里等 [8] 提出了一种基于字词联合训练的 BiLSTM 模型，能够有效识别中文电子病历中疾病、症状等相关实体。Wang Q. 等 [9] 将词典特征加入深度神经网络中，提出了 5 种不同特征的表示方式和基于 BiLSTM 两种不同的神经网络结构。S. Chowdhury 等 [10] 提出了一种新型的、多任务的双向循环神经网络（recurrent neural network，RNN）模型，该模型可以从中文的电子病历中抽取出医疗实体。杨文明等 [11] 使用 BiLSTM-CRF 和 IndRNN-CRF 等模型，抽取了在线医疗问答文本中疾病、治疗、检查和症状 4 类医疗实体。与此同时，也有很多学者利用卷积神经网络（convolutional neural network，CNN）的方法，将其应用到医疗电子病历 NER 任务中。如 Gao M. 等 [12] 利用一种结合词序和局部上下文特征的基于注意力的 IDCNN（iterated dilated convolution neural networks）-CRF 模型，完成了对临床电子病历中医学实体术语的抽取。

但是，以上基于深度神经网络的 NER 方法，都存在无法准确表示字符或者词语多义性的问题。例如，“张三和李四的身高差得很远”和“小明的学习成绩很差”，两个句子中的“差”字在各自的语境中是两个完全不同的含义，但是在上下文无关的词嵌入表示方法（如 Word2Vec）中，两个“差”字映射成完全相同的向量，因此这种向量无法考虑到句子的上下文语义。近年来，学术界提出了许多与上下文有关的词嵌入表示方法，比如 EMLo（embeddings from language models）方法和 OpenAI-GPT（generative pre-training）方法 [13]。但是，上述两种语言模型的语言表示都是单向的，无法同时获取前后两个方向电子病历文本的语义信息。

当前，医疗电子病历的命名实体识别面临着训练语料不足和标注质量不高的问题，由于医疗领域的专业性，导致其缺少高质量的标注语料 [14]。此外，医疗电子病历中的命名实体有着特殊和严谨的语言结构，使得该领域命名实体识别具有一定的挑战性。

为了解决上述问题，本研究拟将可以表示双向丰富语义的 BERT（bidirectional encoder representations from transformers）预训练语言模型引入电子病历 NER 任务中，提出了 BERT-IDCNN-MHA（multi-head

attention）-CRF 命名实体识别模型。并利用该模型对医疗电子病历中预定义的疾病和诊断、影像检查、实验室检验、手术、药物以及解剖部位 6 类实体进行命名实体识别，并且将该 6 类实体正确归类到预定义类别中。

2 BERT-IDCNN-MHA-CRF 命名实体识别模型

BERT-IDCNN-MHA-CRF 命名实体识别模型的整体结构如图 1 所示。

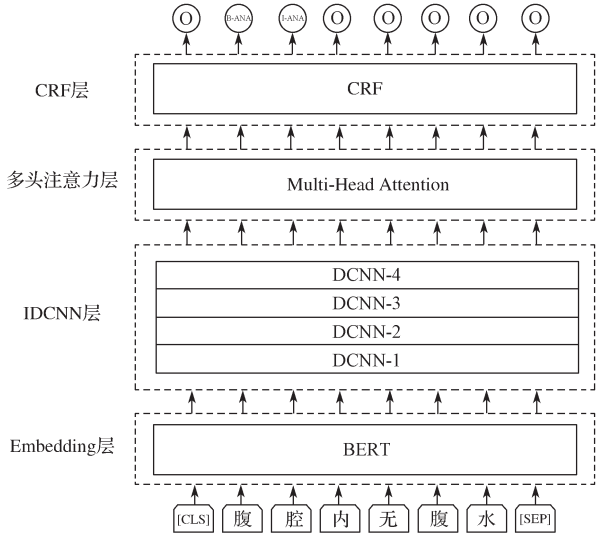


图 1 BERT-IDCNN-MHA-CRF 命名实体识别模型结构图

Fig. 1 BERT-IDCNN-MHA-CRF NER model structure diagram

整个识别模型由 4 个部分组成：首先，输入电子病历中的每一个字，经过 Embedding 层即 BERT 模型，得到与每个字的上下文相关的向量表示。其次，经过 IDCNN 层，将上层输入的每个字的向量进行膨胀卷积编码来提取局部特征，再将获取到的特征向量输入到多头注意力层，多次计算每个字和所有字的注意力概率来获取电子病历句子的长距离特征，得到新的特征向量。因为多头注意力层无法考虑标签之间的依赖关系，比如“I-ANA”标签不能紧接在“B-DIS”标签的后面，所以最后经过 CRF 层约束预测标签之间的依赖关系，对标签序列进行建模，从而获取全局最优序列。为了提高该模型的泛化能力，在 Embedding 层与 IDCNN 层之间加入了 dropout 层。

本研究通过上述命名实体识别模型识别电子病历中的医疗命名实体，具体步骤如下：

1）预处理原始电子病历文本数据集。将电子病历文本集合 $\boldsymbol{D}=\{d_1, d_2, \cdots, d_N\}$ 及其对应的预定义类别 $\boldsymbol{C}=\{c_1, c_2, \cdots, c_M\}$ 按照字符级别进行分割并进行标注，标注时字符和预定义类别用空格隔开。

2）构建电子病历文本训练数据集。按照比例，将分割并标注好的电子病历训练数据分为训练集、验证集和测试集。

3）训练生成命名实体识别模型。基于深度学习技术，训练 BERT-IDCNN-MHA-CRF 命名实体识别模型。

4）识别电子病历文本测试数据集，计算识别率。以电子病历测试文本集合 $\boldsymbol{D}_{\text{test}}=\{d_1, d_2, \cdots, d_N\}$ 为输入，文本中医疗实体提及和所属预定义类别的集合 $\left\{\left\langle m_1, c_{m_1}\right\rangle, \left\langle m_2, c_{m_2}\right\rangle, \cdots, \left\langle m_p, c_{m_p}\right\rangle\right\}$（其中，$m_i$ 是出现在文档 d_i 中的实体提及，c_{m_i} 表示所属的预定义类别）为输出，再根据精确率、召回率和 F_1 值来计算其识别率。

2.1 BERT 预训练语言模型

BERT 模型是一个深度双向编码的包含字符级、词语级和句子级特征的预训练语言模型 [15]。针对医疗电子病历的 NER 任务，只需要调用该预训练模型的相应接口，就能够得到电子病历中每个字的嵌入表示，且能更准确地表示电子病历中与上下文相关的语义信息。本文构建 BERT 预训练语言模型的网络结构如图 2 所示。

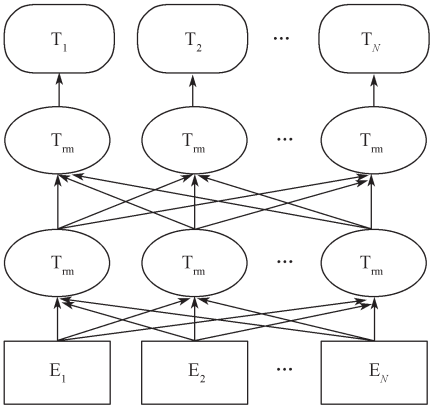


图 2 BERT 预训练语言模型的网络结构图

Fig. 2 Network structure diagram of BERT pre-training language model

BERT 模型使用“Masked 语言模型”来预训练该语言模型，以获取字词级别的上下文相关语义表示。“Masked 语言模型”的核心思想来自于完形填空。传统的语言模型以句子中某个给定词语的下一个词语来预测该词语，而“Marked 语言模型”则是把句子中随机选择的 15% 的词语盖住，通过上下文的内容预测被盖住的词语，但是这一方法会导致微调时模

型无法准确地预测某些 100% 被盖住的词语。为解决这一问题，本研究在 BERT 预训练实验中采取了如下策略：

1）80% 的时间，用“[MASK]”标记来替换被盖住的词语。

2）10% 的时间，用一个任意的词语来替换被盖住的词语。

3）剩余 10% 的时间，保持被盖住的词语不变。

同时，BERT 模型的预训练利用“下一个句子预测”任务来获取句子级别的上下文相关语义表示。该任务的目标，是判断句子 N 是否是句子 M 的下一句。传统的语言模型不能直接反映两个句子之间的关系，在 NLP 领域的许多任务中，都需要在理解两个句子之间关系的基础上进行，如问答和自然语言推理等，因此无法直接使用传统的语言模型。两个句子之间的关系通过 BERT 预训练一个模型学习得到，训练的输入是句子 M 和 N 两个句子，然后利用模型来预测句子 N 是否是 M 的下一句。

BERT 预训练语言模型的输入是电子病历文本中的每一个字，输出是该字的总特征向量，总特征向量由字 (词) 向量、句子切分向量和位置向量 3 种不同的特征向量相加得到，位置向量的计算公式如式（1）和（2）所示。其中，编码使用的是正弦函数和余弦函数，pos 代表的是电子病历文本中第几个字，i 代表第几维，编码后的向量维度是 d_{model}。

$$\boldsymbol{PE}(pos,\ 2i)=\sin\left(pos/10\ 000^{2i/d_{\text{model}}}\right), \tag{1}$$

$$\boldsymbol{PE}\left(pos,\ 2i+1\right)=\cos\left(pos/10\ 000^{2i/d_{\text{model}}}\right)。 \tag{2}$$

BERT 模型输入示例如图 3 所示，第一个标记的标签是一种特殊嵌入 [CLS]，代表电子病历文本的开始位置；其后的特殊嵌入 [SEP]，代表电子病历文本的结束位置。

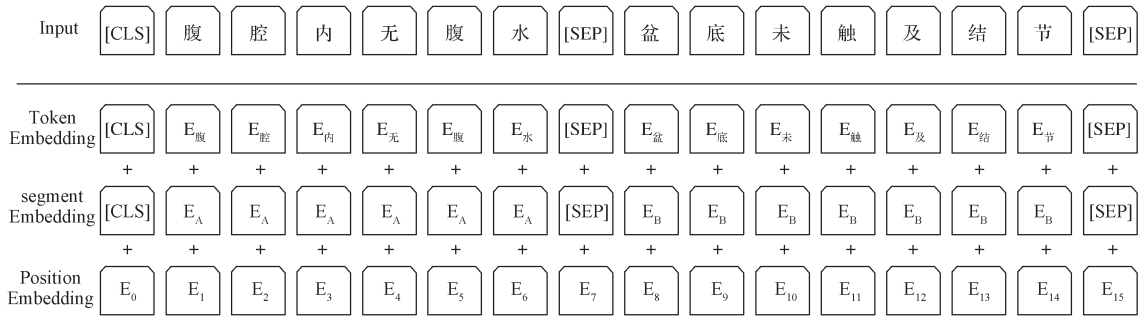
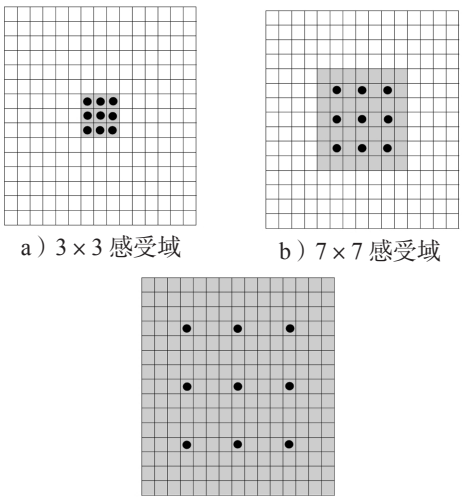


图 3　BERT 模型输入示例

Fig. 3　Samples of BERT model input

2.2　IDCNN 层

相关研究结果表明，相对于 BiLSTM 的长距离依赖关系编码，IDCNN 对局部实体的卷积编码可以达到更好的医疗实体识别效果，同时其训练速度和预测的效率都有所提高 [12]。因此，本研究采用 IDCNN 模型对电子病历文本的特征进行提取。一般的 CNN 滤波器，都是通过在输入矩阵的区域上不断地滑动来做卷积运算，且这种区域通常是连续的。而 DCNN（deep convolutional neural networks）则是因在滤波器上添加了膨胀宽度，导致此时输入矩阵的区域不再连续，每次做卷积运算时都会跳过所有膨胀宽度中间的输入数据。在膨胀卷积运算过程中，输入矩阵上更多的数据被滤波器获取，但是滤波器本身的大小并没有发生变化，反而扩大了其感受域，看上去像是“膨胀”了一般，因此称作膨胀卷积神经网络。与一般的 CNN 相比，DCNN 没有通过池化操作也可以获得较大的感受域，而且减少了信息损失。DCNN 的膨胀示意图如图 4 所示。



a）3 × 3 感受域　　b）7 × 7 感受域

c）15 × 15 感受域

图 4　DCNN 的膨胀示意图

Fig. 4　Dilated schematic diagram of DCNN

图 4 中，图中心点的 1 × 1 区域是开始的感受域，卷积核的大小为 3，从感受域的中心点出发，以步长为 1 向外部扩散，得到图 a 中大小为 3 × 3 的新感

受域；再从新感受域的中心点出发，以步长为 2 向外部扩散，得到图 b 中大小为 7 × 7 的新感受域；接下来从这一新感受域的中心点出发，以步长为 4 向外扩散，得到图 c 中大小为 15 × 15 的新感受域。膨胀卷积的感受域计算公式见式（3），式中 i 代表步长。

$$F_{i+1}=\left(2^{i+2}-1\right)\times\left(2^{i+2}-1\right)。\tag{3}$$

在逐步扩大感受域、层数不断增加的过程中，神经网络参数呈线性增加，而感受域呈指数级增加。如图 4 所示，仅经过 3 步膨胀变化后，感受域就已扩散至输入矩阵中的全部数据。这种膨胀卷积神经网络结构，每层的参数都是相互独立且数量相同的，可有效减少训练时的参数，从而可加快训练速度。

IDCNN 模型则是将 4 个结构相同的膨胀卷积块进行堆叠，相当于进行了 4 次迭代，每次迭代将前一次的结果作为输入，这种参数共享可有效防止模型过拟合，每个膨胀卷积块有膨胀宽度分别为 1, 1, 2 的 3 层膨胀卷积。通过 IDCNN 模型，将电子病历中的每个字进行膨胀卷积编码，自动提取文本中特征，输出为对应的特征向量。虽然 IDCNN 可使感受域变大，但提取的特征仅是局部的，因此还需经多头注意力层进行电子病历文本的长距离特征提取。

2.3 多头注意力层

注意力机制（attention mechanism）首先被应用在数字图像处理领域，后来逐渐被应用于 NLP 领域的多种任务中。可以将注意力函数看作一个查询（$\boldsymbol{Q}$）到一系列键（$\boldsymbol{K}$）– 值（$\boldsymbol{V}$）对的映射。在 NLP 领域的多种任务中，$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 通常取相等值。在计算自注意力时，通常取 $\boldsymbol{Q}$=$\boldsymbol{K}$=$\boldsymbol{V}$，可以计算输入句子中每个字符和所有字符的注意力概率。本研究利用注意力机制中的多头注意力，从电子病历文本的内部结构中得到字符之间的长距离依赖关系。多头注意力模型的结构如图 5 所示，其中，拼接 k 次自注意力计算结果，将拼接结果进行线性变换后，即可以得到本次注意力计算结果。

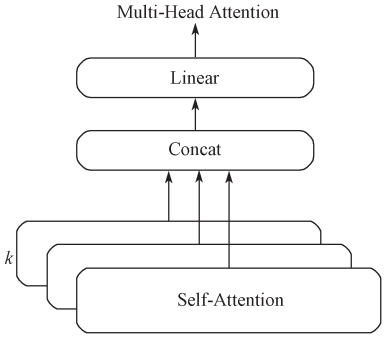


图 5 多头注意力模型结构图

Fig. 5 Multi-Head Attention model structure diagram

与自注意力模型相比，多头注意力模型实质上是进行多次自注意力计算，每一次算一个头，可以使模型在不同的表示子空间里学习到相关的信息而且具有优于 RNN 的并行计算性能。

首先，在电子病历 NER 任务中，对于输入的一个句子 $\boldsymbol{X}$=($\boldsymbol{x}_1$, $\boldsymbol{x}_2$, ···, $\boldsymbol{x}_t$,···, $\boldsymbol{x}_n$)，通过 IDCNN 层后的输出是 $\boldsymbol{Y}$=($\boldsymbol{Y}_1$, $\boldsymbol{Y}_2$,···, $\boldsymbol{Y}_t$,···, $\boldsymbol{Y}_n$)，对于句子中的第 t 个字符的输出状态 Y_t，通过式（4）进行单头自注意力计算。其中，共进行 i 次计算，即有 i 个头，第 i 次计算的结果是 $head_i$。

$$head_i=Attention\left(\boldsymbol{Y}_t\boldsymbol{W}_i^{Q},\boldsymbol{Y}_t\boldsymbol{W}_i^{K},\boldsymbol{Y}_t\boldsymbol{W}_i^{V}\right)=softmax\left(\frac{\left(\boldsymbol{Y}_t\boldsymbol{W}_i^{Q}\right)\left(\boldsymbol{Y}_t\boldsymbol{W}_i^{K}\right)^{\mathrm{T}}}{\sqrt{d_k}}\right)\left(\boldsymbol{Y}_t\boldsymbol{W}_i^{V}\right)。\tag{4}$$

式中：$\boldsymbol{W}_i^{Q}$、$\boldsymbol{W}_i^{K}$和 $\boldsymbol{W}_i^{V}$分别为第 i 次计算的权重参数；

$\sqrt{d_k}$ 为 k 维度的调节平滑项；

$softmax$( ) 为归一化因子。

然后，拼接这 i 次的计算结果，再进行一次线性变换，即可以得到句子中第 t 个字符的多头注意力计算结果，具体的计算公式如式（5）所示，其中 $\boldsymbol{W}^{O}$ 为权重参数。

$$MultiHead_t=Concat\left(head_1,\ head_2,\cdots,\ head_i\right)\boldsymbol{W}^{O}。\tag{5}$$

2.4 CRF 层

CRF 模型是一种经典的判别式概率无向图模型，该模型经常被应用于序列标注任务中，即在给定观察序列 $\boldsymbol{C}$=(c_1, c_2, ···, c_n) 的情况下，计算状态序列 $\boldsymbol{Y}$=(y_1, y_2, ···, y_n) 的条件概率 $P(y|c)$，具体计算公式如式（6）所示，其中，f_k 为特征函数，w_k 为特征函数的权重，$Z(c)$ 为归一化项。

$$P\left(y\mid c\right)=\frac{1}{Z\left(c\right)}\exp\sum_{k=1}^{K}w_k f_k\left(y,\ c\right),\tag{6}$$

$$Z(c)=\sum_{y}\exp\sum_{k=1}^{K}w_k f_k\left(y,\ c\right)。\tag{7}$$

在医疗电子病历 NER 中，多头注意力层无法考虑标签之间的依赖关系，比如“I-ANA”标签不能紧接在“B-DIS”标签的后面。CRF 层可以有效地约束预测标签之间的依赖关系，对标签序列进行建模，从而获取全局最优序列。多头注意力层的输出是电子病历句子中每个字对应的各个标注符号的分数，记矩阵 $\boldsymbol{P}$ 为打分矩阵，$P_{i,\ j}$ 为第 i 个字符分类到第 j 个标签的概率值，$T_{i,\ j}$ 为第 i 个到第 j 个标签的状态转移打分。对于输入句子 $\boldsymbol{X}$=(x_1, x_2, ···, x_n)，句子标签序列 $\boldsymbol{y}$=(y_1, y_2, ···, y_n) 的打分为

$$score(\boldsymbol{X},\ \boldsymbol{y})=\sum_{i=1}^{n}P_{i,\ y_i}+\sum_{i=0}^{n}T_{y_i,\ y_{i+1}}。\qquad(8)$$

使用最大化对数似然函数对CRF模型进行训练，通过式（9）和（10）计算在给定句子 $\boldsymbol{X}$ 的情况下标签序列 $\boldsymbol{y}$ 的条件概率，其中 $\boldsymbol{y}_X$ 为给定的句子 $\boldsymbol{X}$ 全部可能的标签序列，L 为定义的损失函数。

$$P(\boldsymbol{y}\mid\boldsymbol{X})=\frac{\exp\left(score(\boldsymbol{X},\ \boldsymbol{y})\right)}{\sum_{\tilde{y}\in y_X}\exp(\boldsymbol{X},\ \tilde{\boldsymbol{y}})},\qquad(9)$$

$$L=\log\left(P(\boldsymbol{y}\mid\boldsymbol{X})\right)。\qquad(10)$$

在 CRF 模型预测过程中，采用维特比（Viterbi）算法来求解全局最优序列，公式如式（11）所示，其中 $\boldsymbol{y}^*$ 为集合中使得分函数取得最大值的序列。

$$\boldsymbol{y}^*=\arg\max_{\tilde{y}\in y_X}score(\boldsymbol{X},\ \tilde{\boldsymbol{y}})。\qquad(11)$$

3　实验及结果分析

3.1　实验数据及标注策略

本研究采用的电子病历医疗实体识别中文数据集由 CCKS2019 评测任务一“面向中文电子病历的医疗实体识别及属性抽取”提供，所有电子病历语料由专业的医学团队进行人工标注。该标注数据集分为训练集和测试集，其中训练集包含 1 000 份医疗电子病历，共计 7 717 个句子；测试集共包含 379 份医疗电子病历。表 1 是各类医疗实体个数统计信息，总共为 5 363 个文档。

表 1　医疗实体类别数据统计

Table 1　Statistics of medical entity categories

预定义类别	文档个数	预定义类别	文档个数
疾病和诊断	2 116	手术	765
影像检查	222	药物	456
实验室检验	318	解剖部位	1 486

每份电子病历详细地标注了医疗实体的名称、起始位置、结束位置和预定义实体类别，并进行脱敏处理。具体分为疾病和诊断、影像检查、实验室检验、手术、药物和解剖部位 6 类预定义类别，各类预定义类别及其含义信息如下：

1）疾病和诊断。即医学上定义的疾病和医生在临床工作中对病因、病生理、分型分期等所作的判断，如胃癌、肠胃炎等。

2）影像检查。包括影像检查、造影、超声、心电图，如 CT、MRI（magnetic resonance imaging）等。

3）实验室检验。指在实验室进行的物理或化学检查，特指临床工作中检验科进行的化验，不含免疫组化等广义实验室检查，如血红蛋白、CA199 等。

4）手术。指医生在患者身体局部进行的切除、缝合等治疗，如腹腔镜根治性全胃切除术、经腹直肠癌切除术（DIXON）等。

5）药物。指用于疾病治疗的具体化学物质，如伊立替康、格列卫等。

6）解剖部位。指疾病、症状和体征发生的人体解剖学部位，如口腔、十二指肠等。

本研究选择字标注方法完成对数据集的标注，采用 BIO（begin, inside, outside）标注体系，其具体格式为 B-X、I-X 和 O。B 代表医疗实体开始位置的字符，I 代表医疗实体剩余部分的字符，O 代表非医疗实体的字符。X 代表医疗实体的类别，记为 DIS、IMG、LAB、OPE、MED 和 ANA，分别代表疾病和诊断、影像检查、实验室检验、手术、药物和解剖部位 6 类医疗实体。该任务共有 13 种不同的标签，分别为 B-DIS、I-DIS、B-IMG、I-IMG、B-LAB、I-LAB、B-OPE、I-OPE、B-MED、I-MED、B-ANA、I-ANA 和 O。各类别的实体标注符号及示例如表 2 所示。

表 2　医疗实体类别标注符号及示例

Table 2　Classification labeling symbols and examples of medical entities

医疗实体类别	标注符号	标注示例
疾病和诊断	B-DIS / I-DIS	胃 B-DIS 癌 I-DIS
影像检查	B-IMG / I- IMG	C B- IMG T I- IMG
实验室检验	B-LAB / I- LAB	乳 B- LAB 酸 I- LAB
手术	B-OPE / I- OPE	剖 B- OPE 腹 I- OPE 探 I- OPE 查 I- OPE
药物	B-MED / I-MED	同 B-MED 澳 I-MED
解剖部位	B-ANA / I-ANA	腹 B-ANA 腔 I-ANA
非医疗实体	O	无 O

虽然电子病历语料由专业的医学团队进行人工标注，但是不可避免地会出现实体类别或者开始、结束位置的标注错误以及标注前后不一致等问题。比如，在一段电子病历文本“直肠癌术后，拟行第 4 次化疗”中，“直肠癌术后”被人工标注为“疾病和诊断”类别的医疗实体，而在另一段电子病历文本“食管癌术后、肝癌介入术后”中，“食管癌”被人工标注为“疾病和诊断”，与前一段文本中的标注存在前后不一致的问题，这种标注不一致会导致实体识别过

程中错误预测实体边界，从而影响实体识别的效果。本研究针对实体类别或者开始、结束位置的标注错误问题，在数据集的预处理中采取人工纠错的方式，将标注错误的实体进行纠正。

3.2 评价指标

医疗电子病历命名实体识别的评价指标采用精确率（precision）P、召回率（recall）R 以及 F_1-*Measure*，其中 F_1-*Measure* 是精确率和召回率的加权调和平均值，具体公式为（12）~（14）。

$$P=\frac{TP}{TP+FP}\times 100\%, \tag{12}$$

$$R=\frac{TP}{TP+FN}\times 100\%, \tag{13}$$

$$F_1\text{-}Measure=\frac{2PR}{P+R}\times 100\%。 \tag{14}$$

式（12）~（14）中：*TP* 为正确识别医疗实体的个数；

FP 为识别到不相关医疗实体的个数；

FN 为未识别到相关医疗实体的个数。

在预测时，判断医疗实体预测完全正确的标准是实体的边界和类别同时预测正确。

3.3 实验环境

本文实验的命名实体识别模型基于 TensorFlow 框架，具体实验环境设置如表 3 所示。

表 3 实验环境设置

Table 3 Experimental environment settings

项 目	环 境
操作系统	Ubuntu 16.04 LTS
CPU	i5-8300H @2.3 GHz
GPU（显存大小）	NVIDIA GeForce RTX 2060(6 GB)
内存	8 GB
硬盘	512 GB
Python 版本	3.6
TensorFlow 版本	1.10.0

3.4 实验参数设置

BiLSTM-CRF 模型参数设置如下：Word2Vec 的预训练字嵌入向量维数为 100，窗口大小为 3，最小词频为 10；LSTM（long short-term memory）隐藏层的单元个数为 128；学习率为 0.000 5，批大小（batchsize）为 20，dropout 为 0.5，clip 为 5，优化算法使用自适应时刻估计法（Adam）。

IDCNN-CRF 模型的参数设置如下：IDCNN 隐藏层的滤波器个数为 128 个；其余参数的设置与 BiLSTM-CRF 模型保持一致。

BERT-IDCNN-CRF 模型的参数设置如下：采用 BERT-Base 版预训练语言模型，该模型由 Google 提供，为 12 头模式，共有 12 层和 110 M 个参数，隐藏层为 768 维；最大序列长度（max_seq_len）为 128；学习率为 0.000 5，批大小（batchsize）为 20，dropout 为 0.5，clip 为 5，优化算法使用自适应时刻估计法（Adam）。

BERT-IDCNN-MHA-CRF 模型的参数设置如下：采用 BERT-Base 版预训练语言模型，该模型由 Google 提供，为 12 头模式，共有 12 层和 110 M 个参数，隐藏层为 768 维；多头注意力层头数为 4；最大序列长度（max_seq_len）为 128；其余参数设置与 BERT-IDCNN-CRF 模型保持一致。

3.5 实验设计与结果分析

本研究将 CCKS2019 提供的电子病历数据，采用交叉验证的方法，以 7:2:1 的比例划分为训练集、验证集和测试集。为验证 BERT-IDCNN-MHA-CRF 模型的有效性，将该模型和以下模型进行对比：

1）BiLSTM-CRF 模型。即基于 BiLSTM 的特征抽取和 CRF 约束的模型，在该模型中，使用 100 维的 Word2Vec 预训练字向量。

2）IDCNN-CRF 模型。即基于 IDCNN 的特征抽取和 CRF 约束的模型，IDCNN 能够更好地抽取句子的局部特征，且有更快的并行计算速度。在该模型中，使用 100 维的 Word2Vec 预训练字向量。

3）BERT-IDCNN-CRF 模型。即在 IDCNN-CRF 模型的基础上加入 BERT 预训练语言模型。

在该项实验中，epoch 默认设置为 80 次，表 4 是不同模型的实验结果。对比表 4 中各模型的实验结果，可以看出 BERT-IDCNN-MHA-CRF 模型的精确率、召回率和 F_1 值相比于 BiLSTM-CRF 基线模型的分别提高了 1.80%, 0.41%, 1.11%，该模型在疾病和诊断、检查、手术、药物和解剖部位 5 类医疗实体上的 F_1 值是最高的。检验实体最高的 F_1 值为 87.82%，出现在 BiLSTM-CRF 模型中。

在所有模型中，“疾病和诊断”类型医疗实体的 F_1 值较低，该类型实体普遍长度较长，而且存在括号等补充说明信息，例如“（直肠）腺癌（中度分化），浸润溃疡型”，因此在预测该类实体时存在边界预测错误的问题，从而导致实体识别错误。此外，一些“疾病和诊断”医疗实体和“手术”医疗实体在文本结构上相似，这会导致该类型实体被错误分类，比如“脾脏切除术后”和“脾脏切除术”，前者属于“疾病和诊断”实体，而后者属于“手术”实体，虽然两个实体仅一字之差，却是预定义类别不同的两类实体。“解剖部位”类型医疗实体 F_1 值也较低，该类实体的数量是 6 类实体中最多的，而且特征众多，识别时存在较大的难度。

BiLSTM-CRF 模型和 IDCNN-CRF 模型的 F_1 值分别为 81.32% 和 81.44%，说明两种模型的识别效果相当。但是，IDCNN 的并行计算能力比 BiLSTM 的要强，IDCNN-CRF 模型与 BiLSTM-CRF 模型相比，训练一轮的时间要少 25 s。因此，本文实验选择在 IDCNN-CRF 模型的基础上加入 BERT 预训练语言模型，相比于 IDCNN-CRF 模型，BERT-IDCNN-CRF 模型的识别效果有小幅度提升，F_1 值提高了约 0.42%，这说明 BERT 预训练语言模型对于电子病历句子中的上下文语义有更准确的表示，从而可以提高实体识别效果。BERT-IDCNN-MHA-CRF 模型是在 BERT-IDCNN-CRF 模型的基础上，加入多头注意力机制，多次计算句子中每个字和所有字的注意力概率，实验结果表明，该模型的精确率为 82.63%，F_1 值为 82.43%，是所有模型中最高的；同时，其召回率为 82.23%，相比于 BERT-IDCNN-CRF 模型的 F_1 值，提高了 0.57%。

综上所述，本研究提出的 BERT-IDCNN-MHA-CRF 模型的总体性能最好，可以被成功地应用于医疗电子病历命名实体识别中。

表 4　不同模型的实验结果

Table 4　Experimental results of different models　　%

模　型	评价指标	疾病和诊断	检查	检验	手术	药物	解剖部位	综合
BiLSTM- CRF (baseline)	P	80.52	86.17	90.84	82.18	82.05	78.45	80.83
	R	78.85	90.00	85.00	82.18	88.89	80.40	81.82
	F_1	79.67	88.04	87.82	82.18	85.33	79.41	81.32
IDCNN-CRF	P	80.91	81.00	91.27	79.17	81.59	78.48	80.46
	R	81.84	90.00	82.14	75.25	91.11	81.00	82.44
	F_1	81.37	85.26	86.47	77.16	86.09	79.72	81.44
BERT-IDCNN-CRF	P	79.81	85.90	91.84	82.26	86.07	81.61	82.29
	R	82.11	93.06	76.27	71.83	91.30	79.61	81.43
	F_1	80.94	89.33	83.33	76.69	88.61	80.59	81.86
BERT-IDCNN-MHA-CRF	P	84.39	88.00	84.31	86.29	86.29	80.48	82.63
	R	81.15	91.67	72.88	93.04	93.04	81.40	82.23
	F_1	82.74	89.80	78.18	89.54	89.54	80.94	82.43

4　结语

采用基于 BERT 的医疗电子病历命名实体识别模型，能够较好地识别电子病历中的医疗实体。其中 BERT 预训练语言模型可以更准确地表示电子病历句子中的上下文语义，IDCNN 对局部实体的卷积编码相对于 BiLSTM 的长距离依赖关系编码，可以达到更好的医疗实体识别效果，训练速度和预测的效率都有所提高。多头注意力可以获取电子病历句子中的长距离依赖特征。实验结果表明，模型能够较好地完成医疗电子病历的命名实体识别任务。接下来将该命名实体识别模型进行改进，再应用到其它领域的命名实体识别研究中。

参考文献：

[1] 冀相冰，朱艳辉，李　飞，等 . 基于 Attention-BiLSTM 的中文命名实体识别 [J]. 湖南工业大学学报，2019，33(5)：73−78.
JI Xiangbing，ZHU Yanhui，LI Fei. et al. Entity Recognition of Chinese Names Based on Attention-BiLSTM[J]. Journal of Hunan University of Technology，2019，33(5)：73−78.

[2] 王若佳，赵常煜，王继民 . 中文电子病历的分词及实体识别研究 [J]. 图书情报工作，2019，63(2)：34−42.
WANG Ruojia，CHO Sang Wouk，WANG Jimin. Healthcare Data Mining：Word Segmentation and Named Entity Recognition in Chinese Electronic Medical Record[J]. Library and Information Service，2019，63(2)：34−42.

[3] 于　楠，王　普，翁　壮，等 . 基于多特征融合的中文电子病历命名实体识别 [J]. 北京生物医学工程，2018，37(3)：279−284，324.
YU Nan，WANG Pu，WENG Zhuang，et al. Named Entity Recognition in Chinese Electronic Medical Records Based on Multi-Feature Integration[J]. Beijing Biomedical Engineering，2018，37(3)：279−284，324.

[4] KULKARNI A. CRF Based Bio-Medical Named Entity Recognition[J]. International Journal of Emerging Technology and Computer Science，2018，3(2)：135−139.

[5] 许　源，葛艳秋，王　强，等 . 基于 CRF 与 RUTA 规则相结合的卒中入院记录医学实体识别及应用 [J]. 中山大学学报 (医学版)，2018，39(3)：455−462.

XU Yuan，GE Yanqiu，WANG Qiang，et al. Medical Name Entity Recognition and Application in Chinese Admission Record of Stroke Patients Based on CRF and RUTA Rule[J]. Journal of Sun Yat-Sen University（Medical Sciences），2018，39(3)：455−462.

[6] 王润奇，关 毅 . 基于 Tri-Training 算法的中文电子病历实体识别研究 [J]. 智能计算机与应用，2017，7(6)：132−134，138.
WANG Runqi，GUAN Yi. Named Entity Recognition Research in Chinese Electronic Medical Records Based on Tri-Training Algorithm[J]. Intelligent Computer and Applications，2017，7(6)：132−134，138.

[7] 杨红梅，李 琳，杨日东，等 . 基于双向 LSTM 神经网络电子病历命名实体的识别模型 [J]. 中国组织工程研究，2018，22(20)：3237−3242.
YANG Hongmei，LI Lin，YANG Ridong，et al. Named Entity Recognition Based on Bidirectional Long Short-Term Memory Combined with Case Report Form[J]. Chinese Journal of Tissue Engineering Research，2018，22(20)：3237−3242.

[8] 万 里，罗曜儒，李 智，等 . 基于字词联合训练的 Bi-LSTM 中文电子病历命名实体识别 [J]. 中国数字医学，2019，14(2)：54−56.
WAN Li，LUO Yaoru，LI Zhi，et al. The Recognition of Naming Entity of Bi-LSTM Chinese Electronic Medical Records Based on the Joint Training of Chinese Characters and Words[J]. China Digital Medicine，2019，14(2)：54−56.

[9] 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.

[10] CHOWDHURY S，DONG X，QIAN L，et al. A Multitask Bi-Directional RNN Model for Named Entity Recognition on Chinese Electronic Medical Records[J]. BMC Bioinformatics，2018，19：499.

[11] 杨文明，褚伟杰 . 在线医疗问答文本的命名实体识别 [J]. 计算机系统应用，2019，28(2)：8−14.
YANG Wenming，CHU Weijie. Named Entity Recognition of Online Medical Question Answering Text[J]. Computer Systems & Applications，2019，28(2)：8−14.

[12] GAO M，XIAO Q，WU S，et al. An Attention-Based ID-CNNs-CRF Model for Named Entity Recognition on Clinical Electronic Medical Records[C]//International Conference on Artificial Neural Networks. Cham：Springer International Publishing，2019：231−242.

[13] RADFORD A，NARASIMHAN K，SALIMANS T，et al. Improving Language Understanding by Generative Pre-Training[EB/OL]. [2019−09−22]. https://s3-us-west-2. amazonaws.com/openai-assets/researchcovers/languageunsupervised/language understanding paper.pdf, 2018.

[14] 李 飞，朱艳辉，王天吉，等 . 基于医疗类别的电子病历命名实体识别研究 [J]. 湖南工业大学学报，2018，32(4)：61−66.
LI Fei，ZHU Yanhui，WANG Tianji，et al. Research on Electronic Medical Record Named Entity Recognition Based on Medical Categories[J]. Journal of Hunan University of Technology，2018，32(4)：61−66.

[15] DEVLIN J，CHANG M W，LEE K，et al. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding[EB/OL]. [2019−09−22]. https://arxiv.org/abs/1810.04805.

（责任编辑：廖友媛）