基于多层次特征提取的中文医疗实体识别

李正辉[1],廖光忠[2]
(1. 武汉科技大学 计算机科学与技术学院,湖北 武汉 430065;
2. 武汉科技大学 智能信息处理与实时工业系统湖北省重点实验室,湖北 武汉 430065)

摘　要:中文医疗实体识别是医疗领域文本信息处理的基础,但中文医疗文本中常常存在语法不规范、实体嵌套和类型易混淆等问题易造成实体识别精度下降,因此确保中文医疗实体识别的准确度具有较大的理论研究和实际应用价值。为此,提出一种融合 BERT 预训练、双向长短期记忆网络(BILSTM)和结合注意力机制的空洞卷积网络(IDCNN)的实体识别模型来提升中文医疗实体识别的精度。起先,使用 BERT 预训练语言模型使中文字符转换为词向量并增强其语法语义特征;而后将训练好的词向量分别通过 BILSTM 网络和加入注意力机制的 IDCNN 网络获取上下文信息和更大的感受野;最终将蕴含语法语义特征、上下文信息和更大的感受野信息的特征融合并输入到条件随机场(CRF)中进行实体预测。在两个公开的医疗数据集 CMeEE/Yidu-S4K 上的实验表明,该模型的 F1 值分别达到了 0.711 6 和 0.820 6,较主流模型分别提高了 1.40 百分点和 2.29 百分点,验证了此模型在中文医疗实体识别上的有效性。

关键词:实体识别;BERT 预训练;空洞卷积网络;注意力机制;感受野
中图分类号:TP391　　**文献标识码**:A　　**文章编号**:1673-629X(2023)09-0119-07
doi:10.3969/j.issn.1673-629X.2023.09.018

Chinese Medical Entity Recognition Based on Multi-level Feature Extraction

LI Zheng-hui[1], LIAO Guang-zhong[2]
(1. School of Computer Science and Technology, Wuhan University of Science and Technology, Wuhan 430065, China;
2. Hubei Province Key Laboratory of Intelligent Information Processing and Real-time Industrial System, Wuhan University of Science and Technology, Wuhan 430065, China)

Abstract: Chinese medical entity recognition is the basis of text information processing in the medical field, but there are often grammatical irregularities, entity nesting and type confusion in Chinese medical texts that may cause the decrease of entity recognition accuracy, so it is of great theoretical research and practical application value to ensure the accuracy of Chinese medical entity recognition. To this end, we propose an entity recognition model that combines BERT pre-training, bi-directional long and short-term memory network (BILSTM) and IDCNN with attention mechanism to improve the accuracy of Chinese medical entity recognition. At first, the BERT pre-trained language model is used to convert Chinese characters into word vectors and enhance their grammatical-semantic features. The trained word vectors are then passed through the BILSTM network and the IDCNN network with attention mechanism to obtain contextual information and a larger perceptual field, respectively. Finally, the features containing grammatical-semantic features, contextual information and a larger perceptual field are fused and fed into the conditional random field (CRF) for entity recognition. The features containing syntactic semantic features, contextual information and larger receptive field information are finally fused and fed into conditional randomization (CRF) for entity prediction. Experiments on two publicly available medical datasets, CMeEE/Yidu-S4K, showed that the F1 values of the model reached 0.711 6 and 0.820 6 respectively, which were 1.40 and 2.29 percentage points higher than that of the mainstream models, validating the effectiveness of this model for Chinese medical entity recognition.

Key words: entity recognition; BERT pre-training; IDCNN; mechanism of attention; receptive field

收稿日期:2022-11-29　　**修回日期**:2023-03-30
基金项目:国家自然科学基金项目(61502359)
作者简介:李正辉(1998-),男,硕士研究生,研究方向为机器学习、自然语言处理;通讯作者:廖光忠(1969-),男,硕士,副教授,CCF 会员(E4235M),研究方向为物联网技术、信息安全。

0 引 言

命名实体识别(Named Entity Recognition,NER)[1]作为自然语言处理(Nature Language Processing,NLP)的一项基本任务,旨在精准定位与识别文本信息中的预定义实体类型。近年来,随着信息技术的快速发展,自然语言处理的研究也逐渐融入各行各业,其中医疗领域的命名实体识别受到业界的广泛关注。

虽然命名实体识别在诸如机构名、人名、地点和职务等实体识别上取得了很高的成就[2],但是从中文医疗文本中提取实体是一个更复杂的任务[3]。首先,提取的实体类型易混淆,例如“左肺上叶”是属于“病理”类型还是“影像”类型;其次,数据中存在嵌套问题也是导致医疗实体识别精度下降的原因,如“患者呼吸中枢受累”这句话中,“呼吸中枢受累”的实体类型是“症状”,而“呼吸中枢”是“部位”类型;最后,中文医疗实体识别某些类型比较长,这会造成识别这类实体时边界定位错误,从而导致整体效果变差。此外,中文医疗文本中的标注错误和错别字等问题也会影响命名实体识别模型。针对以上问题,该文提出一种融合 BERT 预训练[4]、双向长短期记忆网络(BILSTM)[5]和结合注意力机制[6]的一维空洞卷积神经网络(Iterated Dilated Convolutional Neural Network,IDCNN)[7]的医疗实体识别模型,相较于传统模型,主要贡献如下:

(1)使用 BERT 预训练加强字与句子的联系,使词向量在拥有位置信息的同时,语法语义特征也得到强化,降低数据中易混淆实体类型对模型的影响。

(2)从多层次提取医疗文本特征。在 BERT 语言模型训练好词向量的基础上,使用 BILSTM 模块提取上下文信息;使用 IDCNN 模块捕获更长距离的特征信息,并且为了不遗漏细节特征加入注意力机制。最后在输入条件随机场(CRF)[8]预测实体前进行特征融合,得到蕴含多层次特征的词向量。

(3)分层设置学习率和学习率衰减策略。为了使模型效果更好,在训练时利用学习率衰减策略和分层设置学习率,得到更好的结果,另外为了避免模型过拟合,每个模块都加入随机失活(Dropout)层。

1 相关工作

命名实体识别研究包括统计机器学习方法和深度学习方法。随着近年来人工智能技术的发展,统计机器学习方法费时费力,深度学习方法已成为业内研究的焦点。当前基于深度学习的 NER 的思路主要分为序列标注和基于分类两类。其中,序列标注方法最为常见。BILSTM-CRF 是实体识别中主流的模型,对中文和英文的数据都有良好的效果。李妮等人[9]使用改进传统的卷积后的空洞卷积,在实体识别方面取得不错的成果。Li 等人[10]提出 FLAT 模型,改良 lattice 结构,在性能和效率上都优于其他基于词汇的模型。为了进一步提高实体识别的精度,Li 等人[11]采用负采样的思想,可以有效降低未标记实体带来的误导。崔少国等人[12]融合汉字图形和五笔等特征更进一步提升了模型的效果。另外,在 2018 年由 Google AI 研究院提出的预训练模型 BERT 也对实体识别精度的提升有很大的帮助。

受到以上研究的启发,该文为解决中文医疗实体识别中存在的实体类型易混淆、数据嵌套和实体类型过长等问题,提出一种融合 BERT 预训练、BILSTM 和结合注意力机制的 IDCNN 模型,可以从多层次提取医疗文本的特征,有效提高医疗实体识别的准确度。

2 模型设计

2.1 模型框架

该文设计了一种新的中文医疗实体识别模型,其整体结构框架如图 1 所示。在嵌入层,模型将输入的医疗文本通过 BERT 层化为词向量表示;在特征提取层,将词向量分别通过 BILSTM 层和融合注意力机制的 IDCNN 层,再将这三者词向量融合得到多层次语义特征向量;在输出层,通过 CRF 得到预测的结果。

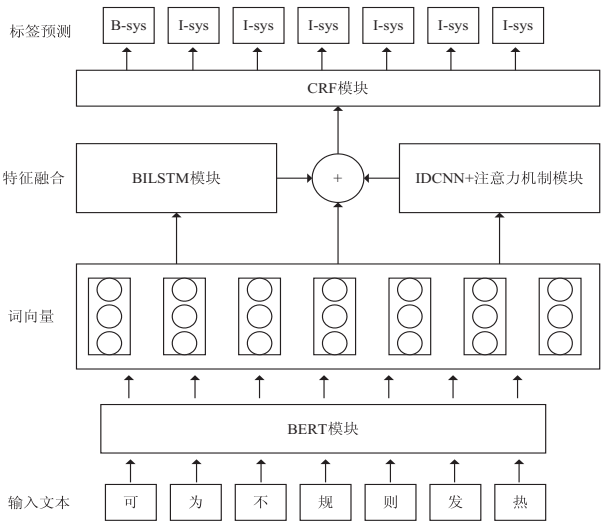

图 1 模型整体结构

2.2 嵌入层

由于输入嵌入层的医疗文本中存在易混淆的类别并且可能有错字等问题,易影响模型性能。而 BERT 语言模型历经几代预训练模型的迭代,克服了 Word2Vec 模型训练的缺点——词向量是静态的,无法表示一词多义;综合 ELMO 和 GPT 模型的优势,做到获取每个字词在当前句子中的上下文信息[13]。因此选用 BERT 作为嵌入层。

BERT 的总体结构如图 2 所示,主要应用

Transform[14]中的编码器(encoder),使用多个 encoder 组成图 2 中的 Trm 单元,使得最终的词向量具有很好的位置特征、句法特征和语义特征。

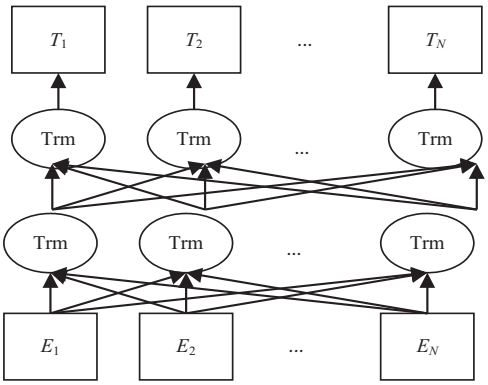

图 2　BERT 整体结构

Trm 单元结构如图 3 所示。对于输入序列 $s=(C_1,C_2,\cdots,C_n)$,其中 C_i 表示医疗文本输入句子的第 i 个字。首先通过 Word2Vec 得到该字的特征向量表示 $e_i=[\boldsymbol{e}^w(C_i)]$,$\boldsymbol{e}^w$ 是 Word2Vec 初始化矩阵。并且需要叠加位置编码,这是因为对于文本序列来说,获取句子中每个字的相对位置很重要,其计算如公式(1)所示:

$$\begin{aligned}\mathrm{PE}_{(\mathrm{pos},2i)}&=\sin(\mathrm{pos}/10\ 000^{2i/d_{\mathrm{model}}})\\ \mathrm{PE}_{(\mathrm{pos},2i+1)}&=\cos(\mathrm{pos}/10\ 000^{2i/d_{\mathrm{model}}})\end{aligned}\tag{1}$$

其中,pos 表示单词的位置,i 表示单词的维度。然后到了 encoder 的关键部分,采用注意力思想使输入序列中每个词向量获取其与其他词向量的关联程度。具体计算如公式(2)所示:

$$\mathrm{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^T}{\sqrt{d_k}}\right)\boldsymbol{V}\tag{2}$$

其中,$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 是词向量矩阵,d_k 是词向量的维度。而这里使用的多头注意力机制是基于注意力机制的基础,先将每个词向量拆分成多个词向量,然后对每个拆分的词向量单独做自注意力机制,最后将不同的结果拼接起来,计算如公式(3)所示:

$$\begin{aligned}&\mathrm{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\mathrm{Concat}(\mathrm{head}_1,\cdots,\mathrm{head}_n)\boldsymbol{W}^o\\ &\mathrm{head}_i=\mathrm{Attention}(\boldsymbol{Q}\boldsymbol{W}_i^Q,\boldsymbol{K}\boldsymbol{W}_i^K,\boldsymbol{V}\boldsymbol{W}_i^V)\end{aligned}\tag{3}$$

因此,词向量就能得到不同空间的句法特征,其中 $\boldsymbol{W}$ 是权重矩阵。最后经过全连接前馈网络(FFN)做非线性变换,计算如公式(4)所示:

$$\mathrm{FFN}(\boldsymbol{Z})=\mathrm{Relu}(\boldsymbol{Z}\boldsymbol{W}_1+\boldsymbol{b}_1)\boldsymbol{W}_2+\boldsymbol{b}_2\tag{4}$$

其中,Relu()是激活函数,$\boldsymbol{W}_1$、$\boldsymbol{W}_2$ 是权重矩阵,$\boldsymbol{Z}$ 是经过多头注意力机制的词向量表示,$\boldsymbol{b}_1$、$\boldsymbol{b}_2$ 是偏置量。另外在流程中,多头注意力机制和全连接前馈网络分别经过一层残差网络和归一化是为了避免梯度消失和梯度爆炸。

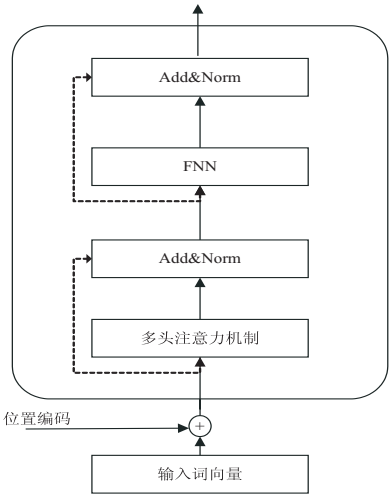

图 3　Trm 单元结构

2.3　特征提取层

特征提取层的目的是基于 BERT 预训练,并结合 BILSTM 和融入注意力机制的 IDCNN 对医疗文本进行多层次的特征提取,特征提取层有三大部分:BILSTM 模块、IDCNN+注意力机制模块和特征融合模块。

2.3.1　BILSTM 模块

BILSTM 是由双向 LSTM 网络组合而成,通过对输入的医疗文本序列做顺序和逆序的计算,提取序列中的上下文信息,最终得到训练好的词向量序列。实现其功能的主要单元为 LSTM,其结构如图 4 所示。LSTM 记忆单元是由输入门、细胞状态、临时细胞状态、隐层状态、遗忘门、记忆门、输出门组成。其核心思想为,通过对细胞状态中信息遗忘和记忆新的信息使得对后续时刻计算有用的信息得以传递,而无用的信息被丢弃,并在每个时间步都会输出隐层状态,其中遗忘、记忆与输出由通过上个时刻的隐层状态和当前输入计算出来的遗忘门、记忆门、输出门来控制。计算如式(5)所示:

$$\begin{aligned}f_t&=\sigma(\boldsymbol{W}_f\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_f)\\ i_t&=\sigma(\boldsymbol{W}_i\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_i)\\ \tilde{C}_t&=\tanh(\boldsymbol{W}_C\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_C)\\ C_t&=f_t*C_{t-1}+i_t*\tilde{C}_t\\ o_t&=\sigma(\boldsymbol{W}_o\cdot[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_o)\\ \boldsymbol{h}_t&=o_t*\tanh(C_t)\end{aligned}\tag{5}$$

式中,σ 是激活函数,$\boldsymbol{W}$ 是权重矩阵,$\boldsymbol{b}$ 是偏置向量,$\boldsymbol{h}_{t-1}$ 是前一时刻的输入,i_t、f_t、o_t 分别是输入门、遗忘门及输出门的输出结果,$\boldsymbol{h}_t$ 是当前时刻的输出,$\boldsymbol{x}_t$ 是当前输入词向量,C_t、C_{t-1}、$\tilde{C}_t$ 分别代表当前时刻、上一时刻和临时的细胞状态。最终得到包含上下文信息的词

向量序列{ $\boldsymbol{h}_1,\cdots,\boldsymbol{h}_{n+1}$ }。

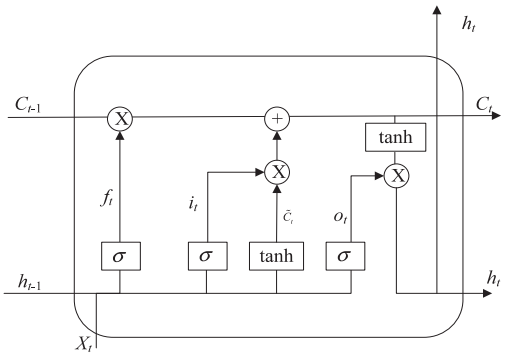

图 4 LSTM 单元结构

2.3.2 IDCNN+注意力机制模块

为了进一步提取医疗文本的特征信息,选用变体的卷积神经网络 IDCNN,这是因为相较于 BILSTM 网络,卷积神经网络对局部特征的提取效果更好,并且 IDCNN 在标准卷积的基础上注入空洞,能够在不通过池化损失信息的情况下增加感受野,对输入中的较长实体能够分词更加准确。另外因 IDCNN 模块的特性,不会造成整体模型参数过大和训练时间过长。

空洞卷积与标准卷积的区别如图 5 所示,通过这种方式,在卷积核大小不变的条件下,就能得到更大的感受域。空洞卷积的感受域计算公式如式(6),其中 i 代表步长。

$$\boldsymbol{F}_{i+1} = (2^{i+2} - 1) \times (2^{i+2} - 1) \tag{6}$$

IDCNN 模块则是将 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台,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 模块,经过 BERT 预训练的词向量将能更好地提取文本中的特征。

虽然 IDCNN 可使感受域变大,能有效提取更长文本的特征,但提取的特征不会都是有用的信息,因此还需经注意力机制进行医疗文本的筛选。这里使用 BERT 模块中的多头注意力机制,经过式(3)的计算,词向量将能得到更好的特征信息。

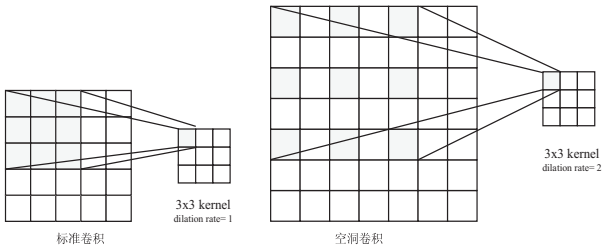

图 5 空洞卷积与标准卷积

2.3.3 特征融合模块

经过上述模块后,分别得到能够准确识别自己的词向量 $\boldsymbol{X}_i$、捕获具有上下文信息的词向量 $\boldsymbol{Y}_i$ 和更长特征的词向量 $\boldsymbol{Z}_i$,为使最终的词向量 $\boldsymbol{C}_i$ 拥有以上所有特征并不造成模型过拟合,先以公式(7)对词向量进行拼接,然后经过 Dropout 层,最后将其输入 CRF 层。

$$\boldsymbol{C}_i = \text{Concat}(\boldsymbol{X}_i,\boldsymbol{Y}_i,\boldsymbol{Z}_i) \tag{7}$$

式中,i 为序号。

2.4 输出层

CRF 模型常用于序列标注任务中的输出层,其可以通过学习给定数据集中标签之间的转移概率从而修正特征提取层的输出,确保最终预测输出的合理性。

3 实验与结果分析

将文中模型与各基线模型在 CMeEE[15] 和 Yidu-S4K 数据集上进行对比实验,评估各模型的输出结果,并进行结果分析。

3.1 实验数据集

采用的数据集为国内公开的中文医疗评测数据集,CMeEE 数据集是中文医学语言理解测评(CBLUE)中命名实体识别任务所使用的数据集,包含训练集 15 000 条句子,验证集 5 000 条句子,每条句子中包含的实体类型和数量不等,总计分为 9 大实体类型,包括:疾病(dis)、临床表现(sym)、药物(dru)、医疗设备(equ)、医疗程序(pro)、身体(bod)、医学检验项目(ite)、微生物类(mic)、科室(dep)。Yidu-S4K 数据集是医渡云(北京)技术有限公司用于医疗命名实体识别评测任务的数据集,其包含:疾病和诊断、检查、检验、手术、药物、解剖部位 6 种实体类型。表 1 和表 2 分别展示了 CMeEE 数据集和 Yidu-S4K 数据集的分布情况。

表 1 CMeEE 数据集分布情况 /个

实体	训练集	验证集	总计
dis	19 083	6 194	25 277
sym	13 808	4 431	18 239
dru	4 348	1 559	5 907
equ	856	162	1 018
pro	9 396	3 238	12 634
bod	15 778	4 759	20 537
ite	3 236	881	4 117
mic	2 227	605	2 832
dep	345	86	431
实体数	69 077	21 915	90 992

表 2 Yidu-S4K 数据集分布情况 /个

实体	训练集	验证集	总计
疾病和诊断	2 465	1 747	4 212
手术	612	417	1 029

续表 2

实体	训练集	验证集	总计
解剖部位	4 822	3 604	8 426
药物	984	838	1 822
影像检查	746	223	969
实验室检验	682	513	1 195
实体数	10 311	7 342	17 653

从上述两个表格可以看出,CMeEE 数据集规模较大,而 Yidu-S4K 数据集规模较小。通过这两个实体类型相似而规模不同的数据集,能进一步验证模型的泛化能力。

3.2　数据集标注与评价指标

实验中两份数据集都使用 BIO 标注体系进行数据的标注。在 BIO 标注体系中,B 代表实体的开始位置,I 代表实体的内部,O 代表非实体部分。并且将两份原始数据集都分别重新划分数据集与测试集,其划分比例为 7∶3。

使用如下三个指标来衡量模型的精准度:精准率(Precision, P)、召回率(Recall, R)和 F1 值(F1)。为计算这三个指标,需使用混淆矩阵,如表 3 所示。该混淆矩阵由以下四个元素组成,真正例:TP;真负例:TN;假正例:FP;假负例:FN。

精准率 P 表示预测为正例的样本中真正正例的比例,表达式如公式(8)所示:

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{8}$$

召回率 R 表示预测为正例的真实正例占所有正例的比例,表达式如公式(9)所示:

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{9}$$

F1 值是精确率和召回率的合成指标,综合了二者的结果,取值范围为[0,1],F1 值越高,代表模型的综合性能越好,表达式如公式(10)所示:

$$\mathrm{F1} = \frac{2 \times P \times R}{P + R} \tag{10}$$

3.3　实验环境与实验参数设置

实验环境为 Window 10 操作系统、CPU Intel Core i5-8400 4.0 GHz、16 GB RAM 以及 NVIDIA GTX 1050Ti 图形处理器。模型框架为 Python3.7.0、Keras 2.3.1 和 tensorflow-gpu 2.4.0。

实验模型中所使用的语言模型 BERT 的版本为 bert-base-chinese,训练方式为自动检查点机制,若当前 epoch 的结果好于当前最优记录,将自动地将此轮 epoch 学习到的参数进行转储,并更新最优记录。并且为了更好地优化模型参数,使用学习率衰减策略和分层设置学习率的方法。学习率衰减策略是每当模型经过 5 轮训练后性能没提升,就缩小当前设置的学习率,再继续训练直到训练轮数结束。使用分层设置学习率是因为该实验需要在预训练模型基础上对下游任务进行微调,因此自定义的模块初始需要较大的学习率。另外的一些其他参数设置见表 3。

表 3　实验参数

参数名	值
梯度下降优化方法	Adam
BILSTM 隐层大小	256
Dropout	0.4
Batch-size	8
初始学习率	$5e^{-4}$
epoch	60
BERT 的 embedding 大小	768
输入句子最大长度	100

3.4　实验方式

为了验证模型的优越性,将文中模型与现有主流模型进行对比实验以及与自身进行的消融实验。

(1)BERT-BILSTM-CRF 模型[16]。该模型是现有常用于序列标注任务的模型之一。

(2)BERT-BILSTM-CRF(*)模型。该模型与模型(1)的不同点为将经过 BERT-BILSTM 的词向量与通过 BERT 预训练的词向量拼接融合后再输入到 CRF 层。

(3)BERT-IDCNN-CRF 模型[9]。该模型将 BILSTM 层换成 IDCNN 层。

(4)BERT-IDCNN-CRF(*)模型。该模型与模型(3)的不同点为将经过 BERT-IDCNN 的词向量与通过 BERT 预训练的词向量拼接融合后再输入到 CRF 层。

(5)BERT-IDCNN-CRF(**)模型。该模型在模型(4)的基础上,在 IDCNN 后加入多头注意力机制。

(6)FLAT 模型[10]。该模型改良了 lattice 结构,提出相对位置编码。

(7)BILSTM-CRF 模型。早期泛用的实体识别模型,相较于模型(1),减少了 BERT 层。

对于对比实验,使用上述模型(2)(4)和(5)与文中模型在两个不同数据集上进行对比。对于消融实验,为了验证文中模型结构的有效性,使用上述模型(2)(4)和(5)与文中模型在数据集 Yidu-S4K 上进行对比。

3.5　结果分析

针对上述实验,使用 3.2 小节所述的评价指标 P、R 和 F1 值进行评估。对比实验结果如表 4 所示。

从表 4 的实验结果可以分析出,BEET 预训练对词向量的增强效果是明显的,在 CMeEE 数据集上 F1 从 61.41% 提升到 69.76%,在 Yidu-S4K 数据集上 F1 从 70.97% 提升到 79.77%,说明 BEER 能够增强词向量的语法语义特征。并且当在 BERT 预训练的基础上进一步融合上下文和更广的感受野信息,模型的性能大大提高,在两个数据集上 F1 值分别提高了 1.40 百分点和 2.29 百分点,可见从多层次提取特征信息有效提高了医疗实体识别的精确度。另外,模型在这两个规模不一样的数据集上都有良好表现,验证了模型的泛化性。

表 4 模型对比实验 %

数据集	模型	*P*	*R*	F1
CMeEE	BERT-BILSTM-CRF	69.24	70.28	69.76
	BERT-IDCNN-CRF	68.44	69.28	68.86
	FLAT	60.56	65.17	62.78
	BILSTM-CRF	58.87	64.17	61.41
	文中模型	70.21	72.13	71.16
Yidu-S4K	BERT-BILSTM-CRF	79.91	79.63	79.77
	BERT-IDCNN-CRF	79.31	79.11	79.21
	FLAT	73.76	72.51	73.13
	BILSTM-CRF	69.43	72.58	70.97
	文中模型	82.35	81.77	82.06

为进一步验证模型的可靠性与泛用性,在 CMeEE 数据集上与近年来刚发表的医疗实体识别模型进行对比实验,在使用相同的参数与数据集的前提下,实验结果如表 5 所示。从中可以看出,虽然文中模型对比文献[17-19]的模型提升性能不大,但是却反映出文中模型拥有着不输于近年来新发表的医疗实体识别模型的性能。

表 5 与近几年模型的对比实验 %

数据集	模型	*P*	*R*	F1
CMeEE	BERT-BILSTM-IDCNN-CRF[17]	71.05	71.08	71.06
	BERT-IDCNN-GAT[18]	71.15	70.28	70.71
	ERNIE-BILSTM-IDCNN-CRF[19]	70.10	72.18	71.12
	文中模型	70.21	72.13	71.16

表 6 模型消融实验 %

数据集	模型	*P*	*R*	F1
Yidu-S4K	BERT-BILSTM-CRF(*)	80.98	79.17	80.07
	BERT-IDCNN-CRF(*)	80.12	79.87	79.99
	BERT-IDCNN-CRF(**)	81.15	80.28	80.71
	文中模型	82.35	81.77	82.06

消融实验的结果如表 6 所示。对比两个基于 IDCNN 模块的模型,F1 值从 79.99% 上升到 80.71%,说明多头注意力机制能够在 IDCNN 获取更大感受野的同时不遗漏细节特征。此外,从表中 F1 值的提高可以看出,多层次提取特征后的确可以提升模型的识别效率,并且在多个特征提取融合后能进一步增强其效果,同时也验证了文中模型的有效性和结构的合理性。

4 结束语

中文医疗实体识别是医疗领域文本信息抽取的基础任务,具有较大的理论研究和实际应用价值。该文提出一种融合 BERT 预训练、双向长短期记忆网络和结合注意力机制的一维空洞卷积网络的实体识别模型,在多层次充分提取了医疗文本的语法和语义特征,有效缓解了医疗文本中语法不规范、实体嵌套和类型易混淆等问题,在两个不同的数据集上表现出不错的性能。

但是中文医疗实体识别是一个复杂的任务,还有很多需要面临的挑战,识别精度还有进一步提升的空间。未来将继续增强模型的鲁棒性和考虑引入例如字

典和字形图形等额外信息以增强模型性能。

参考文献:

[1] 张汝佳,代　璐,王　邦,等. 基于深度学习的中文命名实体识别最新研究进展综述[J]. 中文信息学报,2022,36(6):20-35.

[2] LI J,SUN A,HAN J,et al. A survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering,2020,34(1):50-70.

[3] SHI J, SUN M, SUN Z,et al. Multi-level semantic fusion network for Chinese medical named entity recognition[J]. Journal of Biomedical Informatics,2022,133:104144.

[4] HU P,DONG L,ZHAN Y. BERT pre-training acceleration algorithm based on MASK mechanism[J]. Journal of Physics:Conference Series,2021,2025(1):012038.

[5] 周　康,曲卫东,杨艺琛. 基于增强 BiLSTM 的网络文章核心实体识别[J]. 计算机技术与发展,2021,31(1):7-12.

[6] 陈　明,刘　蓉,张　晔. 基于多重注意力机制的中文医疗实体识别[J]. 计算机工程,2022,48(9):1-7.

[7] KAPOOR A J,FAN H,SARDAR M S. Intelligent detection using convolutional neural network (ID-CNN)[J]. IOP Conference Series:Earth and Environmental Science,2019,234(1):012061.

[8] GAO W, ZHENG X, ZHAO S. Named entity recognition method of Chinese EMR based on BERT-BiLSTM-CRF[J]. Journal of Physics:Conference Series,2021,1848(1):012083.

[9] 李　妮,关焕梅,杨　飘,等. 基于 BERT-IDCNN-CRF 的中文命名实体识别方法[J]. 山东大学学报:理学版,2020,55(1):102-109.

[10] LI X,YAN H,QIU X,et al. FLAT:Chinese NER using flat-lattice transformer[J]. arXiv:2004.11795,2020.

[11] LI Y,LIU L,SHI S. Empirical analysis of unlabeled entity problem in named entity recognition[J]. arXiv:2012.05426,2020.

[12] 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(4):565-571.

[13] 张　栋,陈文亮. 基于上下文相关字向量的中文命名实体识别[J]. 计算机科学,2021,48(3):233-238.

[14] 罗　峦,夏骄雄. 融合 ERNIE 与改进 Transformer 的中文 NER 模型[J]. 计算机技术与发展,2022,32(10):120-125.

[15] ZHANG N,CHEN M,BI Z,et al. Cblue:a chinese biomedical language understanding evaluation benchmark[J]. arXiv:2106.08087,2021.

[16] 谢　腾,杨俊安,刘　辉. 基于 BERT-BiLSTM-CRF 模型的中文实体识别[J]. 计算机系统应用,2020,29(7):48-55.

[17] 郭建勤. 基于深度学习的中文医疗病历命名实体识别研究[D]. 太原:中北大学,2022.

[18] 梁文桐,朱艳辉,詹　飞,等. 基于深度学习多模型融合的医疗命名实体识别[J]. 计算机应用与软件,2022,39(10):162-168.

[19] 邢照野,刘晓群,刘亚军,等. 基于增强字符信息的混合电子病历实体识别模型[J]. 科学与信息化,2021(20):145-148.

(上接第 118 页)

from cheap operations[C]//2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR). Seattle:IEEE,2020:1577-1586.

[13] LIU S,QI L,QIN H,et al. Path aggregation network for instance segmentation[C]//2018 IEEE/CVF conference on computer vision and pattern recognition (CVPR). Salt Lake City:IEEE,2018:8759-8768.

[14] ZHANG Y F,REN W,ZHANG Z,et al. Focal and efficient IOU loss for accurate bounding box regression[J]. Neurocomputing,2022,506:146-157.

[15] REDMON J,DIVVALA S,GIRSHICK R,et al. You only look once:unified,real-time object detection[C]//2016 IEEE conference on computer vision and pattern recognition (CVPR). Las Vegas:IEEE,2016:779-788.

[16] TAN M,PANG R,LE Q V. EfficientDet:scalable and efficient object detection[C]//2020 IEEE/CVF conference on computer vision and pattern recognition (CVPR). Seattle:IEEE,2020:10778-10787.