山西大学学报(自然科学版)47(2):260—268,2024
Journal of Shanxi University(Nat. Sci. Ed.)

DOI:10.13451/j.sxu.ns.2023165

基于SoftLexicon和对抗训练的中文医疗命名实体识别

潘世鹏[1,2],吐尔地·托合提[1,2*],梁毅[1,2],艾斯卡尔·艾木都拉[1,2]
(1.新疆大学 计算机科学与技术学院,新疆 乌鲁木齐 830017;
2.新疆多语种信息技术重点实验室,新疆 乌鲁木齐 830017)

摘 要:现有的医疗实体识别模型当中,多数模型不能充分提取和利用文本序列当中词汇信息,且模型结构复杂,使得模型在面临医疗领域的文本时存在实体边界识别不准、鲁棒性较差等问题,并且多数基于字粒度的命名实体识别(Named Entity Recognition, NER)方法对信息遗漏此类问题解决不够完善。针对此类问题,本文提出了一种基于字词融合和对抗训练的命名实体识别模型。模型使用预训练模型BERT(Bidirectional Encoder Representation from Transformers)获取文本序列的字向量;然后使用SoftLexicon引入词典信息并在字向量中添加对抗训练生成的扰动样本;最后使用BiLSTM-CRF(Bi-Long Short-Term Memory-Condition Random Field)进行特征提取并获取序列标注结果。所提出模型在数据集CCKS2019和CCKS2020上进行实验,*F*1值分别到达了85.07%和90.39%。实验结果表明,与基准模型相比,该模型的*F*1值提升了2.31%和2.88%,说明字词融合方法和对抗训练相结合能够有效识别医疗实体。

关键词:命名实体识别;字词融合;对抗训练;PGD
中图分类号:TP391 **文献标志码**:A **文章编号**:0253-2395(2024)02-0260-09

Chinese Medical Named Entity Recognition Based on Soft-Lexicon and Adversarial Training

PAN Shipeng[1,2], Turdi Tohti[1,2*], LIANG Yi[1,2], Askar Hamdulla[1,2]
(1. *School of Computer Science and Technology, Xinjiang University, Urumqi* 830017*, China;*
2. *Xinjiang Key Laboratory of Multilingual Information Technology, Urumqi* 830017*, China*)

Abstract: In existing medical entity recognition models, most of them cannot fully extract and utilize the lexical information in the text sequence, and their model structures are complex. This makes these models face problems such as inaccurate entity boundary recognition and poor robustness when dealing with medical texts. Additionally, most word-granularity based named entity recognition (NER) methods are not perfect in solving the problem of information omission. To address these problems, a named entity recognition model based on word fusion and adversarial training is proposed in this paper. The model uses a pre-trained model BERT to obtain word vectors of text sequences. Then the SoftLexicon is used to introduce lexical information and add perturbation samples generated by adversarial training to the word vectors. Finally, the BiLSTM-CRF is used to extract features and obtain sequence annotation results. The proposed model is experimented on the datasets CCKS2019 and CCKS2020, where the *F*1 values reach 85.07% and 90.39%, respectively. The experimental results show that compared with the baseline model, the *F*1 value of this model has increased by 2.31% and 2.88%, indicating that the combination of word fusion method and adversarial training can effectively identify medical entities.

收稿日期:2023-08-30;**接受日期**:2023-10-30
基金项目:国家自然科学基金(62166042;U2003207);新疆维吾尔自治区自然科学基金(2021D01C076);国防科技基金加强计划(2021-JCJQ-JJ-0059)
作者简介:潘世鹏(1995-),男,河南郑州人,硕士研究生,研究方向为自然语言处理。E-mail:helloworldp2p@163.com
* **通信作者**:吐尔地·托合提(Turdi Tohti),E-mail:turdy@xju.edu.cn
引文格式:潘世鹏,吐尔地·托合提,梁毅,等.基于SoftLexicon和对抗训练的中文医疗命名实体识别[J].山西大学学报(自然科学版),2024,47(2):260-268. DOI:10.13451/j.sxu.ns.2023165

Key words: named entity recognition; word and character fusion; adversarial training; Projected Gradient Descent (PGD)

0 引言

在医疗领域,人工智能辅助医务人员进行工作的相关技术已经被广泛应用。目前,电子病历(Electronic Medical Record,EMR)在协助医务人员工作以及临床医疗中发挥着重要作用,充分利用电子病历的数据并对其中存储的数据内容进行处理,是构建医疗领域知识图谱的重要一环。文本信息是电子病历当中所存储的主要信息,但是目前多数电子病历当中的文本信息由于其专业领域的特殊性,导致这些信息难以被充分提取。因此,能够快速、准确识别文本中所存在的医疗实体是命名实体识别(Named Entity Recognition,NER)任务的关键应用。

命名实体识别是自然语言处理领域的一项子任务,该技术被广泛应用于信息抽取[1-2]、问答系统[3-4]和机器翻译[5-6]等领域。医疗实体识别是命名实体识别任务中的一个应用,其任务内容需要医疗专业的信息加持,其目的是从医学文本中自动识别出具有特定意义的医学实体。医疗实体识别技术常常被用于医学问答系统[7]、临床信息提取[8]和医疗知识管理[9]等领域。医疗实体识别技术的发展有助于加快医学信息的数字化、智能化进程,提高医学信息处理的效率和质量。

在中国知识图谱与语义计算大会(China Conference on Knowledge Graph and Semantic Computing,CCKS)发布中文电子病历测评任务中,仅从CCKS2019和CCKS2020的参赛模型来看,对于中文医疗领域的实体识别任务多数参赛者和研究人员使用最多的模型依然是BERT-BiLSTM-CRF;多数研究者在模型当中加入不同的融合策略来提升识别效果。借助外部词典以及人工构造模型并处理规则也被研究人员所认可,但是针对中文医疗实体识别当中出现的鲁棒性差、模型可解释性差、召回率低等问题,半监督训练和对抗训练也被应用到此项任务当中。

虽然多数研究者在之前的研究工作中提出的模型获得了一定的效果,但是以上的研究中,多数模型的词嵌入层对于实体边界的检测不够精准。即使使用了字粒度级别的信息,依然存在较为明显的信息遗漏问题,因此需要外部词典信息的加持;并针对医疗文本的专业性,模型需要针对文本杂乱的问题有较强的鲁棒性以使得模型在具体任务中表现得更加可靠。

本文所做出的贡献主要有:

(1)提出了一个基于字词融合与对抗训练的中文医疗实体识别方法,利用软词典将词汇信息融入字向量当中。

(2)设计了一个基于SoftLexicon与PGD(Projected Gradient Descent)对抗训练的医疗实体识别模型,通过将单词的词典信息融入到字符表示当中,并与预训练模型BERT相结合优化字向量信息的丰富度。并在使用SoftLexicon融合字词信息的基础上,加入了对抗训练来快速生成对抗样本,以优化模型整体的鲁棒性。

(3)设计实验证明了该模型在使用不同对抗训练算法,如FGM(Fast Gradient Method)、PGD、Free Large Batch Adversarial Training(FreeLB)的场景下,使用PGD对抗训练可以在CCKS2019和CCKS2020数据集中获得较好结果。

1 相关工作

目前中文命名实体识别任务的主要工作方法是基于深度学习的方法,在此方向的优化策略上,研究者们所提出的方法各有不同。在对基础模型的改进上,也都加入了人工制定的规则或者使用集成模型来完成任务。例如Jie等[10]用句法树改进了原本的LSTM的结构,使用依赖树结构捕捉句子中单词间的远距离和句法关系进提出了依赖导向的LSTM-CRF模型;Ronran等[11]等基于现有的Glove(Global Vectors)和Fasttext词嵌入对提高NER任务性能的影响,并基于此将额外的词汇和字符输入特征组合到BiLSTM中,探索了多种预训练向量作为输入对NER任务的影响;Yan等[12]等使用ResNet(Residual Neural Network)和扩张残差网

络捕捉到了更多的局部特征信息;Li等[13]根据网格(Lattice)结构改进的模型可以从文本序列中抽取晶格结构并加速模型的计算。Lin等[14]设计了能够使用门控机制来动态选择词向量和字符级表示的特征。

由于中文医疗实体存在许多专业词汇如“奥沙利铂”“乙状结肠癌”等,在进行实体识别任务的过程中会遇到实体边界信息提取不充分等问题。在此问题上有许多研究人员进行相关的研究工作。例如罗凌等[15]基于汉字内部结构特征信息的笔画ELMo(Embeddings from Language Models)向量作为输入特征,并构建了多任务学习的神经网络模型,提升了模型的性能;Liu等[16]引入额外的标签信息并借助Lattice-LSTM和优化CharCNN(Character-level Convolutional Networks for Text Classification)引入了额外的分词信息,使用多个互补模型融合提升了模型的性能;乔锐等[17]基于BERT的多个模型融合为单一模型以提升精度,并使用频繁模式挖掘等技术建立规则限制,以解决识别实体边界模糊、合并或分裂错误等难题。董哲等[18]融合了BERT和对抗训练,在食品领域进行NER任务,提高了识别实体边界的精准率。孔令巍等[19]借鉴了前者的工作,将对抗训练结合BERT应用到了医疗实体识别工作当中。

然而,以上研究中存在着词嵌入层实体边界检测不明确,词汇信息遗漏等问题。具体而言,接近实体边界的样本比远离边界的样本更容易导致识别错误,这会影响模型的实体识别性能,并且都存在着模型鲁棒性相对不够强的问题。为解决这些问题,本文提出了一种基于字词融合与对抗训练的模型,以下简称CWAT模型。在CWAT模型当中,字词融合策略上借鉴了Ma等[20]的工作,并提出了将对抗训练融合入模型中,以提高命名实体识别的准确性,并借助对字词向量的拼接完成词典信息的引入以提升模型对边界特征的获取能力,进而提升模型的整体识别效果。

2 CWAT模型

CWAT模型主要由4个基本部分组成,分别为:嵌入层、对抗训练、BiLSTM层、CRF层;模型的整体结构如图1所示,以“右半结肠癌”为输入文本为例。在嵌入层中,被输入文本序列的每个字都通过预训练语言模型BERT将每个字表示成所对应的字向量,获得字特征的同时,使用软词典的方法引入字符所对应的词典信息以获取更加全面的语义特征;在小样本数据集情景下,对抗训练能够充分提升模型鲁棒性;在BiLSTM层时,使用对抗训练获取经过对抗训练的扰动因子与字向量相结合,并将对抗样本送入BiLSTM网络中进行特征提取。在最后的CRF层当中,使用条件随机场进行学习序列标签的约束规则以确保模型的最终输出为正确的序列标签。

2.1 嵌入层

为了捕捉文本当中的上下文信息,CAWT模型在嵌入层构造了字词融合机制。使用BERT输入文本序列,输出为文本序列单个字符所逐一对应的特征向量。以图1中的“右半结肠癌”为例,定义模型的输入文本为$S=(\boldsymbol{c}_1,\boldsymbol{c}_2,\cdots,\boldsymbol{c}_n)\in\mathcal{V}_c$,其中$\mathcal{V}_c$表示字符表,输入BERT后可以使得文本中的每一个字符$\boldsymbol{c}_i$表示成一个向量,其联合表示如式(1)式(2)所示:

$$x_i^c=e^c(\boldsymbol{c}_i), \tag{1}$$

$$x_i^c=[e^c(\boldsymbol{c}_i);e^b(\boldsymbol{c}_i,\boldsymbol{c}_{(i+1)})], \tag{2}$$

其中e_c是字符向量的查找表,e_b为bigram的字符向量查找表。

另外,单纯基于字符的命名实体识别方法的缺点在于单词信息未被充分利用,虽然Lattice-LSTM在模型结构当中引入了词典信息,但是依旧存在信息缺失问题,主要的原因就在于模型对于文本的分词信息缺失情况较为严重。针对上述不足,本文模型主要解决分词信息缺失的问题。

为了在模型中保留分词信息,对每个字符$\boldsymbol{c}_i$匹配到的词语进行结果分类,凡匹配到词中的字符都会被分类到集合当中。对于字符$\boldsymbol{c}_i$集合的构成词集可以用式(3)—式(6)表示:

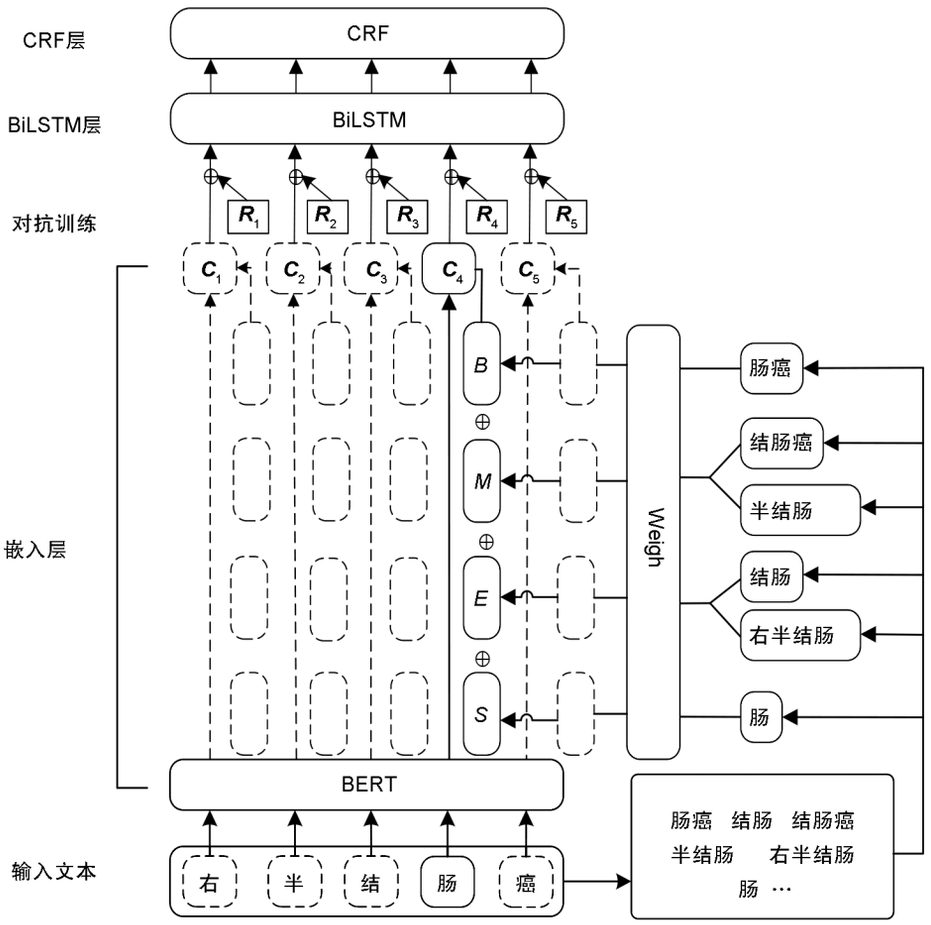
$$B(\boldsymbol{c}_i)=\{w_{i,k},\forall_{w_{i,k}}\in L,i<k\leqslant n\}, \tag{3}$$

$$M(\boldsymbol{c}_i)=\{w_{j,k},\forall_{w_{j,k}}\in L,1\leqslant j<i<k\leqslant n\}, \tag{4}$$

$$E(\boldsymbol{c}_i)=\{w_{j,i},\forall_{w_{j,i}}\in L,1\leqslant j\leqslant i\}, \tag{5}$$

$$S(\boldsymbol{c}_i)=\{\boldsymbol{c}_j,\exists_{\boldsymbol{c}_i}\in L\}, \tag{6}$$

其中L表示词典,w表示在输入的文本序列中



注:B,M,E,S均为包含关键字的词集;C_i为特征向量;R_i为扰动向量; CRF为条件随机场;BERT为预训练模型。

图1 基于SoftLexicon与对抗训练的模型结构

Fig. 1 Model structure based on SoftLexicon with adversarial training

以c_i开始,并以c_k结尾的存在于词典当中的词语。如果在词典中没有找到符合$\{B,M,E,S\}$结构的词,则使用空集。对于以上词集,分别是以c_i开头的词集$B(c_i)$,c_i为中间字的词集$M(c_i)$,以c_i为结尾字的词集$E(c_i)$,无c_i的词集$S(c_i)$。例如,在所列举的例子中,B词集包含了“肠癌”;M词集包含了“结肠癌”;E词集包含了“右半结肠”;S为None。

当模型得到字符的词集之后,统计集合词当中的每个词语的频率使得词集被压缩成一个固定维度的向量。本文在模型当中引入了预先训练好的词向量,单词集中的每个单词都会转化成对应的词向量;然后对四个单词集中的所有单词执行权重归一化,此处使用基于统计的静态加权的压缩方法。压缩方法是使用每个单词的频率作为权重,由于单词的频率是一个可以离线获取的值,因此可以大大加快单词权重的计算。采用这种方法的目的是防止一些较短词汇的频率总是小于覆盖它的较长词汇的频率问题,如公式(7)所示:

$$v^s(S)=\frac{4}{|Z|}\sum_{w\in S}e^w(w)z(w)。\tag{7}$$

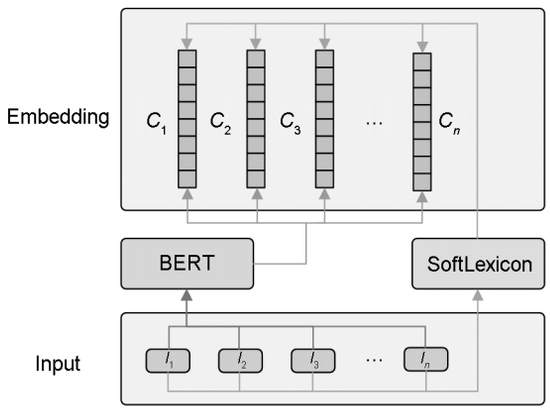
最后将集合$\{B,M,E,S\}$的词向量组成一个固定维度的特征向量,并加入到每个字符当中,将所有的字符对应的词集带入式(7)后,得到公式(8)和公式(9):

$$e^s(B,M,E,S)=[v^s(B);v^s(M);v^s(E);v^s(S)],\tag{8}$$

$$x^c\leftarrow[x^c;e^s(B,M,E,S)]。\tag{9}$$

式(8)和式(9)中,x^c表示字符c对应的字嵌入;$e^s(B,M,E,S)$表示字符c所匹配到的词集经过加权后的词嵌入。经过对输入的文本序列获取字向量以及词向量的融合之后,字符与其对应的词汇信息能够一一对应融合。具体融合方式

如图2所示。



注:I_i为输入的文本序列;C_i为字向量与词向量的拼接后结果。

图2 向量融合方法示意图

Fig. 2 Schematic diagram of vector fusion method

2.2 对抗训练

当完成了对文本信息的特征向量充分编码之后,需要考虑模型在该类数据集上的泛化性和鲁棒性问题。对抗训练作为一种增强网络鲁棒性的手段,主要在分类器对样本数量较小或者有损的情景下使用。对抗训练通过在嵌入层的字向量中添加一些较小的扰动,将获得的对抗样本再送入给模型。对抗训练的定义可简化为公式(10):

$$x^c \leftarrow [\, x^c;\, e^s(B,M,E,S)\,], \tag{10}$$

其中:y为真实信息,x为原始样本信息,θ为模型参数,Δx为扰动,其中扰动的定义如公式(11):

$$\Delta x = \varepsilon \cdot \mathrm{sign}(\nabla_x L(x,y;\theta)), \tag{11}$$

其中sign为符号函数,L为损失函数。

在本文当中,经过模型生成对应的字向量,然后根据字向量以及字向量所对应的标签及模型参数获得扰动值,将扰动值与字向量相加即可得到对抗样本。

2.3 BiLSTM层

为使模型能够充分提取文本上下文信息,此时将特征信息输入BiLSTM当中,其中特征向量$\boldsymbol{H}_i$为添加扰动与词汇信息的字向量$\boldsymbol{C}_i$和对抗样本R_i,X^i_{BiLSTM}为$\boldsymbol{H}_i$经过BiLSTM层输出的特征向量,表示为式(12)和式(13):

$$H_i \in \{\boldsymbol{C}_i, \boldsymbol{R}_i\}, \tag{12}$$

$$X^i_{\mathrm{BiLSTM}} = [\, \overrightarrow{BiLSTM(H_i)};\, \overleftarrow{BiLSTM(H_i)}\,]。 \tag{13}$$

2.4 CRF层

经过BiLSTM输出的序列定义为$x=\{x_1, x_2, \cdots, x_n\}$,对应的文本序列标签为$y=\{y_1, y_2, \cdots, y_n\}$。预测得分函数$S$的表达式如公式(14)所示:

$$S(X_{\mathrm{BiLSTM}}, y) = \sum_{i=1}^{n} P_{i,y_i} + \sum_{i=0}^{n} A_{y_i, y_{i+1}}, \tag{14}$$

其中$P_{i,y}$表示序列中第i个位置的元素输出标签为y_i的概率值;$A_{y_i,y_{i+1}}$表示元素标签y_i转移当前标签y_{i+1}的转移概率;对于每个样本X_{BiLSTM},使用Softmax函数进行归一化,求出标签序列y的概率最大值,如式(15)所示:

$$P(y|X_{\mathrm{BiLSTM}}) = \frac{e^{S(X_{\mathrm{BiLSTM},y})}}{\sum\limits_{\tilde{y}\in Y_{X_{\mathrm{BiLSTM}}}} e^{S(X_{\mathrm{BiLSTM},\tilde{y}})}}。 \tag{15}$$

3 实验与结果分析

3.1 中文病历数据集

本实验中所采用的数据集为公开数据集CCKS2019与CCKS2020。该数据集是中国知识图谱与语义计算大会提供的CCKS2019和CCKS2020医疗电子病历数据。每条数据都包含有原始文本、预定义的实体标签,在CCKS2019和CCKS2020数据集当中大致分为以下6类标签,具体介绍如下:

疾病和诊断(Dis):包括直肠癌、白血病等,是医学疾病名称以及专业医疗人员病理分析得到的实体结果。

影像检查(Che):包括腹部B超、CT等,但不包含一些手术当中常见的操作,例如“胃镜”等。

解剖部位(Ana):包括幽门下淋巴结、肠壁二站等;是医疗领域中常见的人体解剖学部位。

手术(Opr):包括乙状结肠癌根治术等;是常见的外科医疗术语。

药物(Dru):包括奥沙利铂、多西他赛等药物名称。

检验(Test):包括CEA、CA125、白细胞计数等在医学实验室中进行的物理或化学检查。

CCKS2019数据集和CCKS2020数据集训练集分别为1 000条和1 200条,测试集分别为379条和300条。各类实体的分布情况和数量统计见表1。

在构建外部词典时,选用了天池竞赛中药

表1 CCKS数据集各实体数量统计

Table 1 Statistics on the number of entities in the CCKS dataset

标签	CCKS2019		CCKS2020	
	训练集	测试集	训练集	测试集
Dis	2 116	682	4 345	1 866
Che	222	91	1 002	488
Ana	1 486	447	8 811	3 849
Opr	765	140	923	404
Dru	456	263	1 935	906
Tes	318	193	1 297	588

说明书实体数据集和CCKS历年测评任务进行去重融合,将其视为文本实验所用数据集。

3.2 评价指标

在实体识别任务当中,通常使用精准率、召回率以及$F1$来评估一个模型的优越程度,计算精准率、召回率的混淆矩阵和含义如表2所示。

表2 混淆矩阵

Table 2 Confusion matrix

标签	预测值为正	预测值为负
真实值为正	TP	FN
真实值为负	FP	TN

精准率、召回率以及$F1$值的计算方式如公式(16)—(18)所示:

$$precision=\frac{TP}{FP+TP}, \tag{16}$$

$$recall=\frac{TP}{FN+TP}, \tag{17}$$

$$F1=\frac{2\times precison\times recall}{precion+recall}。 \tag{18}$$

如果精确率,召回率,$F1$值的数值越趋近于1,则说明模型的效果越好。

3.3 实验环境

实验所用的硬件环境:CPU为Intel(R) Xeon(R) Gold 6330 CPU @ 2.00 GHz;GPU为NVIDIA GeForce GTX 3090(24 GB);操作系统为Linux version 5.13.0-52-generic (gcc (Ubuntu 9.4.0-1ubuntu1~20.04.1))。软件环境为:CUDA 11.6,深度学习框架Pytorch 1.9.0, Python版本为3.8.2。

本文模型参数设置如表3所示。

3.4 对比实验与分析

为验证本文模型的有效性,实验选取了命名实体识别领域当中常用的模型以及基于字词融合的常用模型进行对比实验分析。本文实验

表3 参数列表

Table 3 List of parameters

项目	参数值
字向量维度	100
词向量维度	50
Dropout	0.5
学习率	2.0×10^5
Batch size	32
优化器	Adam
Epoch	50
隐藏层数	128

在数据集CCKS2019和数据集CCKS2020上进行,实验所获得的结果如表4和表5所示。

表4 CCKS2019对比实验数据

Table 4 Data of comparison experiment on CCKS2019 dataset

模型	CCKS2019		
	Precision/%	Recall/%	$F1$/%
BiLSTM-CRF	77.43	80.22	78.80
Lattice-LSTM[21]	78.88	80.76	79.81
FLAT[22]	81.77	83.43	82.59
BERT-BiLSTM-CRF[23]	81.43	84.12	82.75
ME-CNER[24]	83.56	82.91	83.13
BERT-GCN-CRF[25]	85.05	84.14	84.65
CWAT	83.27	86.94	85.07

表5 CCKS2020对比实验数据

Table 5 Data of comparison experiment on CCKS2020 dataset

模型	CCKS2020		
	Precision/%	Recall/%	$F1$/%
BiLSTM-CRF	81.93	84.11	83.01
Lattice-LSTM[21]	84.34	85.95	85.14
FLAT[22]	85.89	87.76	86.81
BERT-BiLSTM-CRF[23]	86.27	88.79	87.51
RoBERTa-wwm-BiLSTM[26]	87.00	89.86	88.70
ME-CNER[24]	90.10	90.17	90.15
CWAT	89.86	90.92	90.39

由表4和表5的数据可得:

(1)经过对表中数据分析发现,采用了word2vec的BiLSTM-CRF模型与使用了预训练模型的BERT-BiLSTM-CRF相比,在CCKS2019和CCKS2020数据集上的$F1$值分别相差3.95%和4.50%。原因可能在于BiLSTM-CRF使用的是word2vec这种静态词嵌入,相比于使用BERT经过多次transformer输出后的字向量而言,模型难以捕捉丰富的语义文本信息。因此采用预训练模型可以获得相对较为丰

富的语义特征,以增强命名实体识别的效果。

(2)与BiLSTM-CRF模型相比,采用了网格结构的Lattice-LSTM能够将词汇信息引入到字向量当中,丰富了语义特征,在数据集CCKS2019和CCKS2020上的$F1$值分别提升了1.01%和2.13%。但是相比于预训练模型提升不大,主要原因在于未能对输入的文本序列完成丰富的字嵌入,字向量信息获取较少。

(3)FLAT相比于Lattice-LSTM而言,由于FLAT使用了Transformer与位置编码对Lattice-LSTM进行改进,提升了模型获取长距离依赖的能力,同时能够无损地引入词汇信息。因此相比于Lattice-LSTM结构,模型在CCKS2019和CCKS2020上的$F1$值分别提升了2.78%和1.67%。可见注意力机制与位置编码对于实体识别任务有较为显著的提升作用。

(4)本文模型用字词融合的方式对基准模型BERT-BiLSTM-CRF进行改进,使输入序列能将每个字符映射到一个密集向量当中,并将词典特征添加进每个字符表示。在提升计算速度的同时能够兼顾字词融合的准确性。加入对抗训练在针对专业领域的词汇时能够提升模型的鲁棒性。因此在数据集CCKS2019和CCKS2020上的$F1$值相比于BERT-BiLSTM-CRF提升了2.31%和2.88%。

3.5 消融实验与分析

为了验证模型各个模块对模型整体的影响,本文设计了消融实验。将BERT-BiLSTM-CRF模型作为基准模型,并分别使用SoftLexicon算法和对抗训练进行实验。实验结果如表6所示。

表6 CCKS2020消融实验数据

Table 6 Ablation experiment dataset on CCKS2020

模型	CCKS2020		
	Precision/%	Recall/%	$F1$/%
BERT-BiLSTM-CRF	86.27	88.79	87.51
BERT-BiLSTM-CRF (SoftLexicon)	87.44	90.22	88.81
BERT-BiLSTM-CRF (PGD)	87.93	91.03	89.45
CWAT	89.86	90.92	90.39

由表6可以看出:

(1) BERT-BiLSTM-CRF (SoftLexicon)相比于基准模型BERT-BiLSTM-CRF,使用了SoftLexcion算法,充分提取了字符和词汇的特征并融合,精准率、召回率以及$F1$值都有所提升。同时也证明了在专业领域,如医疗领域使用字词融合方法进行实体识别任务是能够提升模型各项指标的。

(2)BERT-BiLSTM-CRF(PGD)相比于基准模型BERT-BiLSTM-CRF使用了对抗训练,尤其是使用了PGD算法。PGD作为一种迭代攻击策略进行了多次反复的迭代,能够将多次迭代扰动映射从而提升模型的鲁棒性。

(3)本文模型在基准模型的基础上同时使用了SoftLexicon进行字词融合以及PGD算法进行对抗训练。使用了字词融合与对抗训练的模型相比于基准模型的Precision、Recall、$F1$值都分别提升了2.31%和2.88%,证明了各个模块在模型中的有效性。

除此之外,本文也探索了使用不同对抗训练算法对模型的影响。选用的对抗训练算法为FGM,PGD,FreeLB,并且在CCKS2020数据集进行了实验。由于FreeLB的作者认为对抗训练和Dropout不能同时使用。但是目前依然有许多模型在使用对抗训练的同时使用Dropout。因此考虑Dropout可能对模型的影响也可能起到一定作用,于是分别在实验中对使用Dropout=0.5和不使用Dropout进行对比实验,并结合不同的对抗训练算法进行实验,探究不同对抗训练算法在使用及不使用Dropout下的结果。结果如表7所示。

表7 不同对抗训练算法实验数据

Table 7 Experimental data of different adversarial training algorithms

模型	Dropout=0.5		
	Precision/%	Recall/%	$F1$/%
Our-FGM	88.38	90.11	89.24
Our-FreeLB	89.67	90.41	90.04
Our-PGD	89.86	90.92	90.39
模型	No Dropout		
	Precision/%	Recall/%	$F1$/%
Our-FGM	89.26	89.61	89.43
Our-PGD	89.89	90.01	89.95
Our-FreeLB	90.01	90.66	90.28

根据表7分析可知:

当使用Dropout并设置为Dropout=0.5时,模型在使用FGM算法相比于使用PGD算法的

$F1$值降低了1.15%。分析其原因,可能是FGM算法只进行了一次迭代对抗。但PGD采用的是多次迭代、小批量下降的结果,并逐步将新的扰动累加到原始梯度上。在使用FreeLB算法进行对抗训练时,相比于使用PGD算法的$F1$值下降了0.35%。分析其原因,可能是FreeLB每轮计算并不是在每次梯度提升的时候都会对参数更新,而是将参数梯度累加起来。

当不使用Dropout时,模型在使用FreeLB时候的效果相比较于PGD更好。并且在不使用Dropout时,各个模型的召回率均有上升,精确率均有下降。$F1$值相比于设置Dropout=0.5时,FreeLB的效果获得0.29%提升,同时PGD的$F1$下降了0.44%。可能的原因是,同时使用Dropout和对抗训练造成了模型的梯度计算和反向传播不稳定。

4 结语

对于中文医疗实体识别任务,我们结合字词融合和对抗训练提出了一种应用于中文医疗实体识别模型。对模型进行改进在于,在字嵌入层使用SoftLexicon算法将字符特征和词典进行融合,充分提取文本序列中的实体信息;使用对抗训练在融合后的特征向量上增加扰动,有助于模型在面临像CCKS这样的小样本数据集上的表现,提升模型的鲁棒性。本文设计的模型效果理想,但能力也较为有限,例如由于使用PGD对抗训练可能会导致训练模型的时间耗费巨大,以及使用SoftLexcion需要引入大量的词典信息。而后续的工作也可以从如下几个方面改进:

(1)可以参考文献[27]采用分级识别思路,分细粒度和粗粒度多级识别,提高模型对不同数据的适应性。

(2)可以考虑在BiLSTM-CRF结构中加入注意力机制,利用文档级的信息来解决序列标签不一致的问题。

(3)尝试多种序列标注模型,也可以提升模型的整体性能。

参考文献:

[1] LU Y J, LIU Q, DAI D, *et al*. Unified Structure Generation for Universal Information Extraction[C]//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dubin:Association for Computation Linguistics, 2022: 5755-5772. DOI: 10.18653/v1/2022.acl-long.395.

[2] BANERJEE P S, CHAKRABORTY B, ANAND U, *et al*. Trainable Framework for Information Extraction, Structuring and Summarization of Unstructured Data, Using Modified NER[J].*Wirel Pers Commun*, 2021, **117**(2): 769-807. DOI: 10.1007/s11277-020-07896-w.

[3] SARROUTI M, OUATIK EL ALAOUI S. SemBioNLQA: a Semantic Biomedical Question Answering System for Retrieving Exact and Ideal Answers to Natural Language Questions[J]. *Artif Intell Med*, 2020, **102**: 101767. DOI: 10.1016/j.artmed.2019.101767.

[4] YIN D D, CHENG S Y, PAN B X, *et al*. Chinese Named Entity Recognition Based on Knowledge Based Question Answering System[J]. *Appl Sci*, 2022, **12**(11): 5373. DOI: 10.3390/app12115373.

[5] JAIN A, PARANJAPE B, LIPTON Z C. Entity Projection via Machine Translation for Cross-lingual NER[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 1083-1092. DOI: 10.18653/v1/d19-1100.

[6] LI Z, QU D, XIE C J, *et al*. Language Model Pre-training Method in Machine Translation Based on Named Entity Recognition[J]. *Int J Artif Intell Tools*, 2020, **29**(**7**n08): 2040021. DOI: 10.1142/s0218213020400217.

[7] JIANG Z X, CHI C Y, ZHAN Y Y. Research on Medical Question Answering System Based on Knowledge Graph [J]. *IEEE Access*, 2021, **9**: 21094-21101. DOI: 10.1109/ACCESS.2021.3055371.

[8] HAHN U, OLEYNIK M. Medical Information Extraction in the Age of Deep Learning[J]. *Yearb Med Inform*, 2020, **29**(1): 208-220. DOI: 10.1055/s-0040-1702001.

[9] GONG F, WANG M, WANG H F, *et al*. SMR: Medical Knowledge Graph Embedding for Safe Medicine Recommendation[J]. *Big Data Res*, 2021, **23**: 100174. DOI: 10.1016/j.bdr.2020.100174.

[10] JIE Z M, LU W. Dependency-guided LSTM-CRF for Named Entity Recognition[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 3862-3872. DOI: 10.18653/v1/d19-1399.

[11] RONRAN C, LEE S. Effect of Character and Word Features in Bidirectional LSTM-CRF for NER[C]//2020 IEEE International Conference on Big Data and Smart Com-

puting (BigComp). Busan: IEEE, 2020: 613−616. DOI: 10.1109/BigComp48618.2020.00132.

[12] YAN C X, SU Q, WANG J. MoGCN: Mixture of Gated Convolutional Neural Network for Named Entity Recognition of Chinese Historical Texts[J]. *IEEE Access*, 2020, **8**: 181629−181639. DOI: 10.1109/ACCESS.2020.3026535.

[13] LI X N, YAN H, QIU X P, *et al*. FLAT: Chinese NER Using Flat-lattice Transformer[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2020: 6836−6842. DOI: 10.18653/v1/2020.acl-main.611.

[14] LIN Y, LIU L Y, JI H, *et al*. Reliability-aware Dynamic Feature Composition for Name Tagging[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2019: 165−174. DOI: 10.18653/v1/p19-1016.

[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. DOI: 10.11897/SP.J.1016.2020.01943.
LUO L, YANG Z H, SONG Y W, *et al*. Chinese Clinical Named Entity Recognition Based on Stroke ELMo and Multi-task Learning[J]. *Chin J Comput*, 2020, **43**(10): 1943−1957. DOI: 10.11897/SP.J.1016.2020.01943.

[16] LIU M L, ZHOU X S, CAO Z, *et al*. Team MSIIP at CCKS 2019 Task 1[C]//2019 China Conference on Knowledge Graph and Semantic Computing. Hangzhou: Chinese Information Processing Society of China, 2019: 1−11.

[17] 乔锐, 杨笑然, 黄文亢, 等. 基于BERT与模型融合的医疗命名实体识别[J/OL]. 2022−08−08. https://conference.bj.bcebos.com/ccks2019/eval/webpage/pdfs/eval_paper_1_1_1.pdf.
QIAO R, YANG X R, HUANG W K, *et al*. Medical Named Entity Recognition Based on BERT and Model Fusion[J/OL]. 2022−08−08. https://conference.bj.bcebos.com/ccks2019/eval/webpage/pdfs/eval_paper_1_1_1.pdf.

[18] 董哲, 邵若琦, 陈玉梁, 等. 基于BERT和对抗训练的食品领域命名实体识别[J]. 计算机科学, 2021, **48**(5): 247−253.. DOI: 10.11896/jsjkx.200800181.
DONG Z, SHAO R Q, CHEN Y L, *et al*. Named Entity Recognition in Food Field Based on BERT and Adversarial Training[J]. *Comput Sci*, 2021, **48**(5): 247−253.. DOI: 10.11896/jsjkx.200800181.

[19] 孔令巍, 朱艳辉, 张旭, 等. 基于对抗训练的中文电子病历命名实体识别[J]. 湖南工业大学学报, 2022, **36**(3): 36−43. DOI: 10.3969/j.issn.1673-9833.2022.03.006.
KONG L W, ZHU Y H, ZHANG X, *et al*. Named Entity Recognition of Chinese Electronic Medical Records Based on Adversarial Training[J]. *J Hunan Univ Technol*, 2022, **36**(3): 36−43. DOI: 10.3969/j.issn.1673-9833.2022.03.006.

[20] MA R, PENG M L, ZHANG Q, *et al*. Simplify the Usage of Lexicon in Chinese NER[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 5951−5960. DOI: 10.48550/arXiv.1908.05969.

[21] DEVLIN J, CHANG M , LEE K, *et al*. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[J]. arXiv Prerint: 1810.04805, 2018. DOI: 10.48550/arXiv.1810.04805.

[22] LI X, YAN H, QIU X, *et al*. FLAT: Chinese NER Using Flat-lattice Transformer[J]. arXiv Preprint: 2004.11795, 2020. DOI: 10.18653/v1/2020.acl-main.611.

[23] DAI Z J, WANG X T, NI P, *et al*. Named Entity Recognition Using BERT BiLSTM CRF for Chinese Electronic Health Records[C]//2019 12th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP−BMEI). Suzhou: IEEE, 2020: 1−5. DOI: 10.1109/CISP-BMEI48845.2019.8965823.

[24] XU C W, WANG F Y, HAN J L, *et al*. Exploiting Multiple Embeddings for Chinese Named Entity Recognition[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management. New York: ACM, 2019: 2269−2272. DOI: 10.1145/3357384.3358117.

[25] 景慎旗, 赵又霖. 面向中文电子病历文书的医学命名实体识别研究: 一种基于半监督深度学习的方法[J]. 信息资源管理学报, 2021, **11**(6): 105−115. DOI: 10.13365/j.jirm.2021.06.105.
JING S Q, ZHAO Y L. Recognizing Clinical Named Entity from Chinese Electronic Medical Record Texts Based on Semi-supervised Deep Learning[J]. *J Inf Resour Manag*, 2021, **11**(6): 105−115. DOI: 10.13365/j.jirm.2021.06.105.

[26] 温超杰,陈涛,朱江. 基于预训练模型和领域词典的医疗命名实体识别方法研究[C]//2020年全国知识图谱与语义计算大会.南昌:中国中文信息学会, 2020: 1−11.
WEN C J, CHEN T, ZHU J. Research on medical named entity recognition method based on pre-training model and domain dictionary [C]//2020 National Conference on Knowledge Graph and Semantic Computing. Nanchang: Chinese Information Processing Society, 2020: 1−11.

[27] 盛剑, 向政鹏, 秦兵, 等. 多场景文本的细粒度命名实体识别[J]. 中文信息学报, 2019, **33**(6): 80−87. DOI: 10.3969/j.issn.1003-0077.2019.06.012.
SHENG J, XIANG Z P, QIN B, *et al*. Fine-grained Named Entity Recognition for Multi-scenario[J]. *J Chin Inf Process*, 2019, **33**(6): 80−87. DOI: 10.3969/j. issn. 1003-0077.2019.06.012.