doi:10.3969/j.issn.1672-1292.2021.01.006

基于 ALBERT 的中文医疗病历命名实体识别

陈　杰[1],奚雪峰[1,2],皮　洲[1],盛胜利[3],崔志明[1,2]

(1.苏州科技大学电子与信息工程学院,江苏 苏州 215009)
(2.苏州智慧城市研究院,江苏 苏州 215009)
(3.Computer Science Department,Texas Tech University,Texas 79431,USA)

[摘要]　医疗病历命名实体识别的主要任务是将临床电子病历中的非结构化文本转化为结构化数据,进而为面向医疗领域任务开展的数据挖掘提供基础支撑. 提出一种基于 ALBERT 模型融合学习的中文医疗病历命名实体识别模型. 首先,采用人工标注方式扩展样本数据集,结合 ALBERT 模型对数据集进行微调;其次,采用双向长短记忆网络(BiLSTM)提取文本的全局特征;最后,基于条件随机场模型(CRF)命名实体的序列标记. 在标准数据集上的实验结果表明,该方法进一步提高了医疗文本命名识别精度,减少了时间开销.

[关键词]　ALBERT,命名实体识别,电子医疗病历,双向长短记忆网络,条件随机场
[中图分类号]TP181　[文献标志码]A　[文章编号]1672-1292(2021)01-0036-08

ALBERT-Based Named Entity Recognition of Chinese Medical Records

Chen Jie[1],Xi Xuefeng[1,2],Pi Zhou[1],Victor S Sheng[3],Cui Zhiming[1,2]

(1.School of Electronic and Computer Engineering,Suzhou University of Science and Technology,Suzhou 215009,China)
(2.Suzhou Smart City Research Institute,Suzhou 215009,China)
(3.Computer Science Department,Texas Tech University,Texas 79431,USA)

Abstract:The main task of named entity recognition on medical record is to convert unstructured text into structured data, and then provide an important fundamental support for data mining for medical field tasks. This paper proposes a named entity recognition method for Chinese medical records based on ALBERT and fusion model. Firstly,we use manual labeling to expand the sample dataset,and fine-tune the dataset in conjunction with the ALBERT. Secondly,the Bi-directional Long Short-Term Memory(BiLSTM)is used to extract the global features of the text. Finally,on the basis of the conditional random field model(CRF),sequence tags for named entities are made. The experimental results on the standard dataset show that the proposed method further improves the accuracy of name entity recognition on medical text and greatly reduces the time overhead.

Key words:ALBERT,named entity recognition,clinical electronic medical records,BiLSTM,CRF

随着电子病历的应用范围越来越广,各级医院和研究机构产生了海量的电子病历. 如何对电子病历中的自然语言文本进行分析,从而生成有效的医疗信息,逐渐成为近年来的研究热点. 其中,医疗病历命名实体识别(named entity recognition,NER)就是该领域的基础性研究工作,其研究成果能够为医学知识库的构建、药物安全性检测、临床决策及个性化患者精准医疗服务提供支撑.

目前国内外关于命名实体识别的绝大部分研究都是基于英文[1-3],面向中文的命名实体识别相对较少[4-6],尤其在电子病历文本分析处理领域[7-8]. 针对电子医疗病历命名实体识别任务,本文基于 ALBERT 模型,融合双向长短记忆网络(BiLSTM)及条件随机场模型,构建了一个 ALBERT-BiLSTM-CRF(ALBBC)模型,并针对中文医疗病历命名实体识别任务,开展了多个模型的比对实验,发现所构建的 ALBBC 模型在识别精度及时间开销上均优于同类模型.

收稿日期:2020-08-08.
基金项目:国家自然科学基金项目(61673290、61876217)、江苏省"六大人才高峰"高层次人才项目(XYDXX-086)、苏州市科技发展计划产业前瞻性项目(SYG201817)、2020 年江苏省研究生科研创新计划项目(KYCX20_2762).
通讯作者:奚雪峰,副教授,研究方向:自然语言处理、高性能并行计算、面向对象技术应用. E-mail:xfxi@usts.edu.cn

1　相关工作

Bikel 等[1]最早提出了基于隐马尔可夫模型的英文命名实体识别方法,其在 MUC-6 测试文本集的测试结果为:英文地名、机构名和人名的识别精度分别达到了 97%、94%和 95%,召回率分别达到了 95%、94%和 94%. Liao 等[2]提出了基于条件随机场模型,采用半监督的学习算法进行命名实体识别. Ratinov 等[3]采用未标注文本训练词类模型(word class model)的办法,有效提高了 NER 系统的识别效率,并针对 CoNLL-2003 数据集开发出精确度和召回率的调和平均数 $F1$ 值达到 90.8%的命名实体识别系统. 中文的命名实体识别也获得了广泛关注. Tsai 等[4]提出基于最大熵的混合的方法. 陈钰枫等[5]提出了汉英双语命名实体识别的对齐交互式方法. 面向中文病历文本命名实体识别的工作不多,杨锦锋等[7]于 2016 年开始研究中文电子病历的命名实体方法,并构建了一个实体关系语料库.

目前针对命名实体识别任务的常用方法包括支持向量机(support vector machines,SVM)[9]、条件随机场(conditional random field,CRF)[10]、结构化支持向量机(structured support vector machines,SVM)[11]、递归神经网络(recurrent neural network,RNN)及其变体模型[12]、卷积神经网络(Convolutional Neural Networks,CNN)及其变体模型[13]等. Liu 等[14]设计不同特征模板和上下文窗口进行条件随机场的学习训练,并比对分析模型实体识别效率,以寻找最佳的电子病历特征模板和上下文窗口. Liu 等[15]通过实验对比了 BiLSTM-CRF 与传统的 CRF 实体识别算法的性能,结果表明前者性能较好. Qiu 等[16]为提高循环神经网络模型的训练速度,提出了残差卷积神经网络条件随机场模型,在 CCKS 开放测试语料上获得了更好的训练速度和 $F1$ 值. 随着近年 Contextualized embedding 研究的突飞猛进,出现了 ELMO[17]、BERT 等[18]可以生成上下文相关的词向量表示模型. 相较于 word2vec 而言,上述模型表达准确率有大幅提高,同时克服了 word2vec 存在的多意词无法用单一词向量表示的问题.

2　问题定义

对于给定的一组电子病历纯文本文档,目标是识别并抽取出与医学临床相关的命名实体并将其归类到预定义类别中. 本文给出了 1 500 份标注好的训练数据用于模型的训练和调优,共需识别包括疾病和诊断、影像检查、实验室检验、手术、药物及解剖部位在内的 6 种实体. 一般而言,中文电子病历命名实体识别是一个序列标注问题. 本文使用 BIO(Begin,Inside,Other)标签方案将数据集给出的标签映射到每一个字符上,进行字符级别(char level)的标记,如表 1 所示.

输入:

1.电子病历的自然语言文本 $D=\{d_1,\cdots d_N\}$,$d_i=\langle w_{i1},\cdots,w_{in}\rangle$;

2.预定义类别:$C=\{c_1,\cdots c_m\}$;

输出:

实体提及和所属类别对的集合:$\{\langle m_1,c_{m_1}\rangle,\langle m_2,c_{m_2}\rangle,\cdots,\langle m_p,c_{m_p}\rangle\}$.

其中,$m_i=\langle d_i,b_i,e_i\rangle$,是出现在文档中的医疗实体提及(mention);b_i 和 e_i 分别表示 m_i 在 d_i 中的起止位置;$c_{m_i}\in C$ 表示所属的预定义类别. 要求实体提及之间不重叠,即 $e_i<b_{i+1}$.

预定义类别如表 2 所示.

表 1　标注方案

Table 1　Labeling schme

患	者	有	轻	微	腹	痛
O	O	O	O	O	B_Dnd	I_Dnd

表 2　预定义类别

Table 2　Predefined categories

类别	描述
疾病与诊断	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.
检查	影像检查(X 线、CT、MR、PETCT 等)+造影+超声+心电图,为避免检查操作与手术操作过多冲突,不包含此外其他的诊断性操作,如胃镜、肠镜等.
检验	在实验室进行的物理或化学检查,本期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.
手术	医生在患者身体局部进行的切除、缝合等治疗,是外科的主要治疗方法.
药物	用于疾病治疗的具体化学物质.
解剖部位	指疾病、症状和体征发生的人体解剖学部位.

3　数据集与预处理

3.1　数据集

本论文使用的数据集是医渡云公开的电子病历数据集(http://openkg.cn/dataset/yidu-s4k),包含 1 500 条标注数据、1 000 条非标注数据、6 个类别的实体词词表,如表 3 所示.

表 3　数据集

Table 3　Dataset

类别	训练集	测试集
文本	1 500	1 000
疾病与诊断	6 211	—
检查	1 490	—
检验	1 885	—
手术	1 327	—
药物	2 841	—
解剖部位	12 660	—
总数	26 414	—

3.2　数据预处理

首先数据集(以下称原始数据集)有部分标注在某些词上存在一定程度上的标注标准不统一问题,手动对数据集进行了修正(以下称修正数据集). 同时考虑到标注标准应可由模型从原始数据集中自动学习到,泛化能力也会更强,因此也保留了原始数据集. 此时产生了原始数据集和修正数据集这两个独立的数据源. 由于医学领域的用词造句较为特殊,目前的公共分词工具在医学术语中表现不佳,因而本文使用字符而不是词语作为序列标注模型的单位. 本文将原始数据集和修正数据集中的每一条数据分别拆分为单个字符,按照 BIO 标签方案将数据集给出的标签映射到每一个字符上. 至此数据预处理流程结束.

4　基于 ALBERT 的中文医疗病历命名实体识别模型

4.1　ALBERT

ALBERT[19] 由谷歌 AI 团队于 2019 年发布. BERT 模型凭借 Masked Language Model(Masked LM)、双向 Transformer encoder 以及句子级别的负采样得到了一个强大的、深度双向编码的、包含着充分描述了字符级、词级、句子级甚至句间关系的特征的预训练模型,针对特定任务,只需简单针对任务数据对模型进行微调,即可完成整个模型的构建. BERT 因其庞大的参数量,在实际应用中常常受到硬件内存的限制;而增加 BERT-large 等模型的隐藏层大小也会导致性能下降. 如表 4 所示,BERT-large 的隐藏层大小增加一倍,该模型在 RACE 基准测试上的准确率显著降低,ALBERT 的准确率基本无影响,ALBERT-large 相较 BERT-large 的参数小得多,只有 18M 个参数.

表 4　预训练模型对比

Table 4　Comparison of pre-trained models

Model	Hidden Size	Parameters	RACE(Accuracy)
BERT-large(Devlin et al.2019)	1 024	334M	72.0%
BERT-large(ours)	1 024	334M	73.9%
BERT-xlarge(ours)	2 048	1 270M	54.3%
ALBERT-large(Lan et al.2020)	1 024	18M	68.4%
ALBERT-xlarge(ours)	2 048	59M	70.2%

4.2　条件随机场(CRF)

条件随机场是一种无向概率图模型,是一种判别模型,长期以来广泛应用于序列标注问题[2-3,13]. 给定字符序列 $z=\{z_1,\cdots,z_n\}$,z_i 代表第 i 个字符及其特征所组成的输入向量;给定 z 的标签序列 $y=\{y_1,\cdots,y_n\}$,$\gamma(z)$ 代表 z 的所有可能标签. CRF 模型定义了在给定字符序列 z 时,标签序列为 y 的概率公式:

$$p(y|z;\theta)=\frac{\sum_{t=1}^{n}\exp(S(y^t,z^t,\theta))}{\sum_{t=1}^{n}\sum_{j\in\gamma(z)}\exp(S(y_j,z^t,\theta))},$$

式中,$S(y^t,z^t,\theta)$ 为势函数;θ 为 CRF 模型的参数.

4.3　BiLSTM-CRF

Hochreiter 和 Schmidhuber[20] 于 1997 年提出了 LSTM,最初是为了解决递归神经网络(recurrent neural network,RNN)训练伴随的梯度缓慢和梯度爆炸,为保持信息完整引入了记忆细胞[15],记录历史上下文信息. 近年 LSTM 方法被广泛应用于自然语言处理领域,目前 NER 的主流模型是 BiLSTM-CRF,其结构如图 1

所示. Lample 等[21]提出了 2 种神经网络方法,一种基于 BiLSTM 和 CRF,另一种是受移位归约解析器(shift-reduce parser)启发提出的基于转换(transition)的方法构建和标记分段.

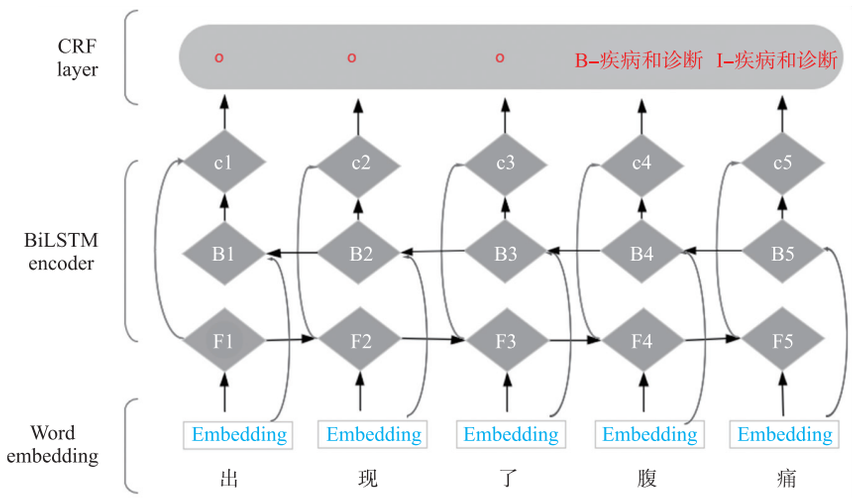


图 1　BiLSTM-CRF 模型

Fig. 1　BiLSTM-CRF model

4.4　ALBBC 模型

针对中文医疗病历原始数据集和修正数据集,本文构建了 ALBERT-BiLSTM-CRF(ALBBC)模型. 模型所使用的特征为纯字符特征. 利用 ALBERT 模型自带的词典将数据集的单个字符映射为 ID 后,再经 ALBERT 模型的 Embedding 层对字符 ID 进行字向量建模,随后进入网络预训练. 首先对大量的电子病历进行命名实体的标注规范;然后在中文数据集上测试 ALBERT 预训练模型,再进入 BiLSTM 提取深层信息,并使用 CRF 进行解码;最后,分析 ALBERT 模型的不足之处,并针对医疗数据集的特点,微调该模型以优化整个网络在医疗数据集上的表现. 模型框架如图 2 所示.

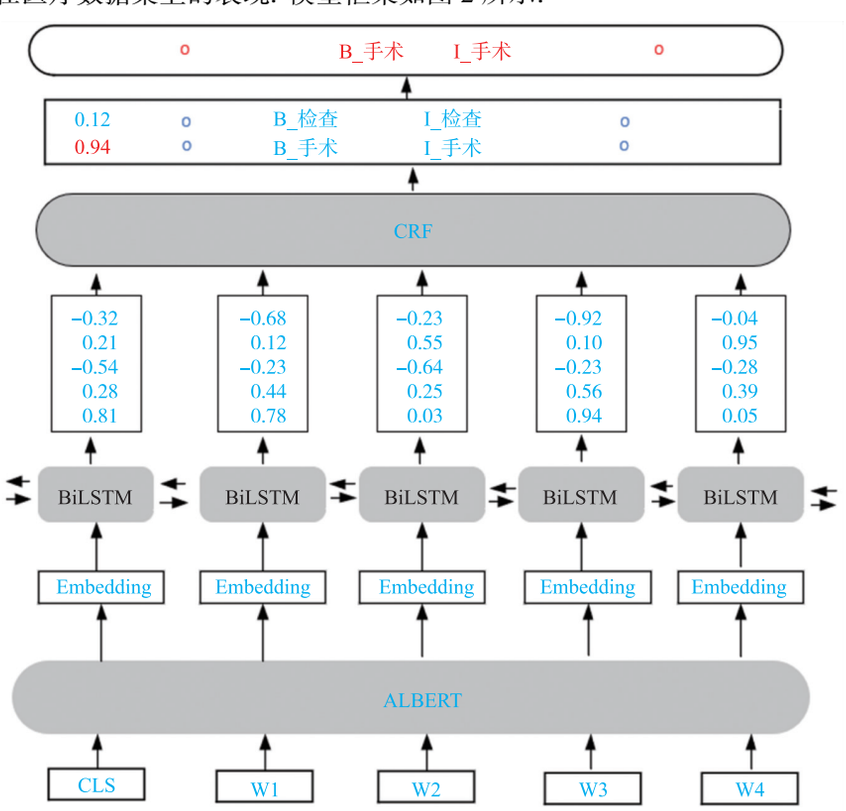


图 2　ALBERT-BiLSTM-CRF 模型

Fig. 2　ALBERT-BiLSTM-CRF model

5　实验结果与比较

5.1　实验设置

实验软件模型基于 TensorFlow-gpu1.15.2 深度学习框架和 keras、numpy、seqeval 等第三方库,使用参数大小为 1.8M 的中文模型 ALBERT_TINY;硬件采用 4 块 NVIDIA GeForce GTX 2080Ti 显卡加速训练与预测.

中文电子病历数据集共包含疾病与诊断、检查、检验、手术、药物、解剖部位 6 类命名实体. 将数据集中的 80%作为训练集,10%作为验证集,10%作为测试集. ALBERT 模型的最长序列长度选择为 512,优化算法为 Adam 算法,学习率设为 1e-5,batch_size 为 16,epoch 为 60,MAX_SEQ_LEN=128,双向 LSTM 层的隐层节点数均为 32(这里指单个方向的隐藏层节点数).

本论文所有实验结果均由多次实验取均值所得.

5.2　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)、$F1$ 值(F-measure)对电子病历命名实体识别结果进行性能衡量.

具体计算公式为:

$$\text{正确率 } P=\frac{\text{正确识别的命名实体个数}}{\text{命名实体总个数}},$$

$$\text{召回率 } R=\frac{\text{正确识别的命名实体个数}}{\text{测试集中出现的实体个数}},$$

$$F1=\frac{2*P*R}{P+R}.$$

5.3　实验比对

如图 3 所示,不断调整参数进行试验,分别训练了当 epoch 值为 10,15,20,30,50 时的 $F1$ 值,此处 $F1$ 值为各个实体 $F1$ 值的平均值. 可以看到,当 epoch 值小于 30 时,模型明显为欠收敛;epoch 为 50 左右,基本可以收敛;$F1$ 值在 epoch 值为 60 时达到最高,当 epoch 值为 70 时已经开始下降. 进一步调整 epoch 值,将 $F1$ 值提高到将总体识别效果提高了 1.3 个 $F1$ 值,总体识别的 $F1$ 值为各个实体 $F1$ 值的平均值,效果最佳的 epoch 值为 60,其 $F1$ 值为 85.38%,故选择该 epoch 值为综合实验参数. 以下实验均在 epoch 值为 60 的基础上展开,其中对“实验室检验”类别实体的识别效果最为突出,$F1$ 值从原来的 81.76% 提高到 84.85%. epoch 值和 $F1$ 值实验中均采用最佳优化器 Adam.

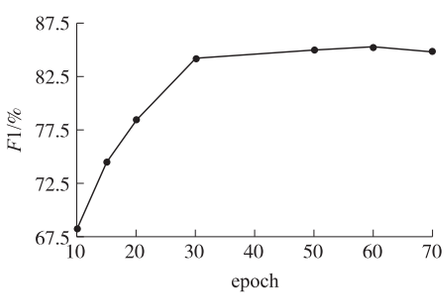


图 3　epoch 值与 $F1$ 值关系图

Fig. 3　Graph of epoch value and $F1$ value

为了得到更适合的参数值,对丢弃率(Dropout)和优化器(Optimizer)也进行了类似的实验. 对 Dropout 进行的实验如表 5 所示.

表 5　不同的 Dropout 对模型效果的影响

Table 5　Effect of different Dropout

Entity	$F1$				
	Dropout=0.1	Dropout=0.2	Dropout=0.3	Dropout=0.5	Dropout=0.8
药物	85.01	83.33	78.63	81.75	66.85
手术	86.70	83.42	83.67	64.07	77.04
影像检查	94.34	84.62	82.46	75.00	86.00
实验室检验	100.00	90.24	90.24	94.74	91.67
疾病和诊断	79.74	85.85	87.27	74.43	83.92
解剖部位	90.88	92.90	92.54	65.32	88.57

表 5 是不同的 Dropout 值对于模型精度的影响实验,可以看出 Dropout 值对实验识别效果存在较为明

显的影响. 除原先默认参数值 0.5 之外,从 0.1 到 0.8,无论是单个实体识别效果还是整体识别效果均有所提升,说明本实验中 Dropout 值设置越小对实验效果提高越有帮助. 故此,本实验选择 Dropout 值为 0.1 作为综合模型参数.

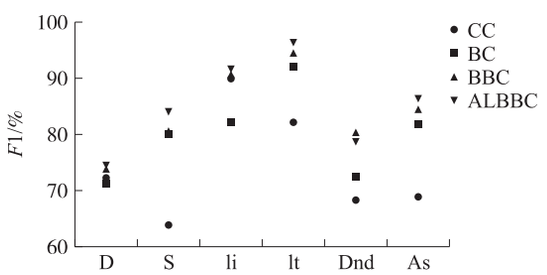
表 6 所示为不同优化器对模型精度的影响. 由表 6 结果可以看出,Adam 优化器的识别效果最为突出. Adam、Momentum 和 SGD 都是常用的神经网络优化器. Adam 是一类对每个参数的自适应学习率进行计算的算法,结合了 Adadelta 和 Momentum 的优点. 实验结果显示,Adam 的效果优于其他两个优化器. 最传统的 SGD 相较于 Adam 和 Momentum 在结果上有一定差距,经过分析 SGD 多次陷入鞍点,从而只有局部最优解. 因此,本文实验选择 Adam 作为模型优化器.

通过与 CNN-CRF(CC)[22]、BiLSTM-CRF(BC,该模型还使用了 attention 机制[23])、BERT-BiLSTM-CRF(BBC)3 种模型的实验结果对比,本文采用的方法在速度与精度上均有所提升. 如图 4 所示,ALBERT-BiLSTM-CRF(ALBBC)模型在除疾病与诊断这一个实体上的 *F*1 值低于 BBC 外,其余 5 类实体上 *F*1 值均领先于另外 3 种模型. BERT 的参数量远高于 ALBERT,其在数据集中属于疾病与诊断实体的数据量是 6 种实体中最多的,因此可以训练出更好的模型,但 BERT 的训练时间也远超 ALBERT. 从表 4 可以看出,相对于不使用 contextual representation 的模型来说,ALBERT 和 BERT 这种上下文相关的 word representation 所得的结果提升明显.

表 6　不同的优化器对模型效果的影响

Table 6　The influence of different optimizers on the model performance

Entity	*F*1		
	SGD	Adam	Momentum
药物	74.53	74.59	79.12
手术	81.27	84.19	82.14
影像检查	85.30	91.74	85.21
实验室检验	74.07	96.47	80.17
疾病和诊断	57.96	78.83	67.31
解剖部位	83.90	86.46	84.08



D:药物;S:手术;li:影像检查;lt:实验室检验;Dnd:疾病和诊断;As:解剖部位

图 4　4 个模型 *F*1 值散点对比图

Fig. 4　*F*1 scatter comparison chart of the four models

词之间广泛存在的语义相关对于提升深度学习模型对于语义的理解至关重要. 含有信息更加丰富的 word embedding 对于下游的任务来说,更容易让模型得到更优的结果.

本论文的主要实验是 CC、BC、BBC、ALBBC 4 个模型的对比,其中 CC、BBC 两个模型是本文实验的对照组. 模型效果对比如表 7 所示,实验结果对比如表 8 所示.

表 7　模型效果对比

Table 7　Comparison of models' effect

	CC	BC	BBC	ALBBC
患者 3 月前因“直肠癌”在我院于全麻上行直肠癌根治术(DIXON 术),手术过程顺利,术后给予抗感染及营养支持治疗,患者恢复好,切口愈合良好	Dnd:直肠癌	Dnd:直肠癌	Dnd:直肠癌 S:全麻	Dnd:直肠癌 S:全麻
患者于 2015-07 出现左下侧胸部疼痛,不伴发热、皮肤瘙痒、乏力等症状,自行药膏涂抹,未予重视,后左下侧胸壁肿物逐渐增大,患者于 2016-01 就诊于 ***** 医院	As:胸部 Dnd:胸壁	As:下侧胸部 Dnd:胸壁	As:下侧胸部 Dnd:下侧胸壁	As:下侧胸部 Dnd:后左下侧胸壁
患者因“咳嗽、消瘦伴胸背胀痛 2 个月”于 2016.05.31 日第 1 次入我院. 入院查体:右肺语颤减弱,右肺中叶叩诊浊音. 双肺呼吸音粗,右肺呼吸音低,轻度湿性啰音.	As:胸背;双肺	As:胸背;双肺	As:胸背,双肺 Dnd:双肺呼吸音粗;右肺呼吸音低	As:胸背;双肺 Dnd:右肺语颤减弱;双肺呼吸音粗;右肺呼吸音低
患者于 2 月前无明显诱因出现左侧肢体不自主抽搐,伴有意识不清发作,就诊于我院神经内三科,考虑脑梗死症状性癫痫,给予德巴金 0.52/日口服,控制癫痫发作.	Dnd:脑梗;癫痫	Dnd:脑梗;癫痫 D:德巴金	Dnd:脑梗死症状性癫痫 D:德巴金	Dnd:脑梗死症状性癫痫 D:德巴金

表 8　实验结果对比

Table 8　Comparison of experimental results

模型	命名实体	正确率/%	召回率/%	*F*1 值/%
CC	药物	74.11	70.65	72.34
	手术	66.58	61.76	64.08
	影像检查	87.64	92.34	90.02
	实验室检验	66.65	80.58	82.32
	疾病和诊断	66.65	70.58	68.56
	解剖部位	71.25	70.58	69.21
BC	药物	81.22	64.52	71.32
	手术	84.56	79.13	80.31
	影像检查	90.12	84.37	82.36
	实验室检验	93.67	88.64	92.16
	疾病和诊断	74.45	76.54	72.56
	解剖部位	88.90	76.65	81.87
BBC	药物	83.32	72.89	74.01
	手术	86.32	80.56	80.56
	影像检查	90.56	89.12	90.62
	实验室检验	98.67	92.12	94.45
	疾病和诊断	80.42	82.34	80.45
	解剖部位	90.12	83.12	84.52
ALBBC	药物	85.01	66.44	74.59
	手术	86.70	81.82	84.19
	影像检查	94.34	89.29	91.74
	实验室检验	96.53	93.18	96.47
	疾病和诊断	79.74	77.93	78.83
	解剖部位	90.88	82.45	86.46

6　结论

针对中文电子病历文本,本文提出一种模型融合方法用于文本实体识别任务. 其核心包括基于 ALBERT 模型微调数据集,采用双向长短记忆网络(BiLSTM)来提取文本的全局特征,以及基于条件随机场模型(CRF)进行命名实体的序列标记. 在标准数据集上的实验结果表明,本文所提方法进一步提高了医疗文本命名识别精度,并大大减少了时间开销.

BERT 和 ALBERT 等结合了上下文信息的预处理模型意义重大,但均需大量的训练数据来支持. 相对于其他领域拥有大规模数据的任务来讲,电子病历的数据集相对较小,所以准确率不高,但专业领域上的实体识别更具有研究和现实意义. 针对数据量不足的问题,未来将考虑扩充数据集和结合小样本学习的方法来进一步提高准确率.

[参考文献](References)

[1] BIKEL D M,SCHWARTA R,WEISCHEDEL R M. An algorithm that learns what's in a name[J]. Machine Learning,1999,34(1/2/3):211-231.

[2] LIAO W H,VEERAMACHANENI S. A simple semi-supervised algorithm for named entity recognition[C]//The Proceedings of NAACL HLT 2009. Boulder,USA:ASL,2009:58-65.

[3] RATINOV L,ROTH D. Design challenges and misconceptions in named entity recognition[C]//Proceedings of the Thirteenth Conference on Computational Natural Language Learning(CoNLL-2009). Boulder,USA:ASL,2009:147-155.

[4] TSAI T H,WU S H,LEE C W,et al. Mencius:a Chinese named entity recognizer using the maximum entropy-based hybrid model[J]. International Journal of Computational Linguistics and Chinese Language Processing,2004,9(1):65-82.

[5] 陈钰枫,宗成庆,苏克毅. 汉英双语命名实体识别与对齐的交互式方法[J]. 计算机学报,2011,34(9):1688-1696.

[6] 张海楠,伍大勇,刘悦,等. 基于深度神经网络的中文命名实体识别[J]. 中文信息学报,2017,31(4):28-35.

[7] 杨锦锋,关毅,何彬,等. 中文电子病历命名实体和实体关系语料库构建[J]. 软件学报,2016,27(11):2725-2746.

[8] YOUNG T,HAZARIKA D,PORIA S,et al. Recent trends in deep learning based natural language processing[J]. IEEE Computational Intelligence Magazine,2018,13(3):55-75.

[9] ASAHARA M,MATSUMOTO Y. Japanese named entity extraction with redundant morphological analysis[C]//Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology. Association for Computational Linguistics. Sapporo,Japan:ACL,2003:8-15.

[10] CHEN A,PENG F,SHAN R,et al. Chinese named entity recognition with conditional probabilistic models[C]//Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing. Sydney,Australia:ACL,2006:173-176.

[11] CHEN Y,ZHOU C J,LI T X,et al. Named entity recognition from Chinese adverse drug event reports with lexical feature based BiLSTM-CRF and tri-training[J]. Journal of Biomedical Informatics,2019,96:103252.

[12] HUANG Z H,XU W,YU K. Bidirectional LSTM-CRF models for sequence tagging[C]//ACL. Beijing,China:ACL,2015:13-16.

[13] STRUBELL E,VERGA P,BELANGER D,et al. Fast and accurate entity recognition with iterated dilated convolutions[C]//EMNLP. Copenhagen,Denmark:ACL,2017:2670-2680.

[14] LIU K X,HU Q C,LIU J W. Named entity recognition in Chinese electronic medical records based on CRF[C]//2017 14th Web Information Systems and Applications Conference(WISA). Jeju,Korea:IEEE,2017:105-110.

[15] LIU Z J,YANG M,WANG X L,et al. Entity recognition from clinical texts via recurrent neural network[J]. BMC Medical Informatics and Decision Making,2017,17:53-61.

[16] QIU J,QI W,ZHOU Y,et al. Fast and accurate recognition of Chinese clinical named entities with residual dilated convolutions[C]//2018 IEEE International Conference on Bioinformatics and Biomedicine(BIBM). Madrid,Spain:IEEE,2018:935-942.

[17] PETERS M E,NEUMANN M,IYYER M,et al. Deep contextualized word representations[C]//Proceedings of NAACL-HLT. New Orleans,USA:ACL,2018:2227-2237.

[18] DEVLIN J,CHANG M W,LEE K,et al. BERT:pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics:Human Language Technologies. Minneapolis,USA:ACL,2019:278-286.

[19] LAN Z,CHEN M,GOODMAN S,et al. ALBERT:a lite BERT for self-supervised learning of language representations[C]//International Conference on Learning Representations. New Orleans,USA:Elsevier,2019:12-17.

[20] HOCHREITER S,SCHMIDHUBER J. Long short-termmemory[J]. Neural Computation,1997,9(8):1735-1780.

[21] LAMPLE G,BALLESTEROS M,SUBRAMANIAN S,et al. Neural architectures for named entity recognition[C]//NAACL-HLT. San Diego,USA:ACL,2016:260-270.

[22] LUO L,YANG Z,YANG P,et al. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition[J]. Bioinformatics,2018,34(8):1381-1388.

[23] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need[C]//Advances in Neural Information Processing Systems. Long Beach,USA:NeurIPS,2017:6000-6010.

[责任编辑:严海琳]