DOI:10.3969/j.issn.1005-202X.2021.11.021

医学生物信息

基于ALBERT与BILSTM的糖尿病命名实体识别

马诗语，黄润才
上海工程技术大学电子电气工程学院，上海 201620

【摘要】糖尿病命名实体识别技术能够从糖尿病文献中识别出关键信息，为糖尿病的诊断和治疗工作提供帮助。为此，本研究提出一种基于轻量型动态词向量模型（ALBERT）与双向长短记忆神经网络的命名实体识别方法，该方法旨在解决BERT语义单一、词汇量有限的问题。除此之外，还针对动态词向量训练耗时长、资源成本高的缺点进行了改进。本实验在糖尿病数据集上展开，并与现有主流模型进行对比。结果表明，融合ALBERT的实体识别效果均高于现有主流模型，且ALBERT较BERT训练速度有所提升。

【关键词】糖尿病；命名实体识别；轻量型动态词向量模型；双向长短记忆网络；条件随机场
【中图分类号】TP183；TP391；R319 【文献标志码】A 【文章编号】1005-202X（2021）11-1438-06

Named entity recognition of diabetes based on ALBERT and BILSTM

MA Shiyu, HUANG Runcai
School of Electrical and Electronic Engineering, Shanghai University of Engineering and Technology, Shanghai 201620, China

Abstract: Named entity recognition of diabetes is able to identify useful key information from the literatures related to diabetes, which is helpful for the diagnosis and treatment of diabetes. A named entity recognition method based on a lite BERT (ALBERT) and bidirectional long short-term memory network is proposed for solving the problems of BERT such as single semantics and limited vocabulary. In addition, the shortcomings of time consuming and high resource cost of dynamic word vector training are also improved. The experiment is carried out on the diabetes data set and then compared with the existing mainstream models. The results show that the entity recognition effect of the model with ALBERT is higher than that of the existing mainstream models, and that the training speed of ALBERT is faster than that of BERT.

Keywords: diabetes; named entity recognition; a lite BERT; bidirectional long short-term memory network; conditional random field

前 言

糖尿病是当前威胁全球人类健康的最重要的非传染性疾病之一，根据国际糖尿病联盟统计，2011年全球糖尿病患者已达3.7亿，其中80%在发展中国家，预计到2030年全球将有5亿多糖尿病患者[1]。中国糖尿病患者在全球占比最高，成人糖尿病患者约1.298亿，平均每10个成年人中有1个糖尿病患者，因此，防治糖尿病成为重点难题之一。然而，我国糖尿病血糖控制状况不佳，有研究显示，糖尿病知晓率、治疗率、控制率均偏低[1]。为此，需要加强居民对糖尿病知识的关注，通过分析糖尿病相关知识来引导广大市民及医疗卫生机构提早预防或延缓这一疾病的发生[2]。人们对如何迅速从众多文献中获取专业知识给予了极大的关注。医学文本挖掘技术在文本知识的自动获取中起着重要作用，作为这项技术的任务之一，糖尿病命名实体识别（Named Entity Recognition, NER）旨在从糖尿病文献中识别特定类型的名称，如1型糖尿病、2型糖尿病、血糖、胰岛素促分泌素等[3]。NER为下一步关系抽取，构建知识图谱提供了前提；为引导广大市民了解糖尿病相关知识，指导糖尿病患者加强健康管理提供了技术帮助。

1 相关工作

百度研究院在2015年提出了深度学习应用NER的经典模型，即BILSTM-CRF[4]，凭借双向长短期记忆网络（Bidirectional Long Short-Term Memory Network, BILSTM）对上下文信息进行深度建模，条件随机场

【收稿日期】2021-07-05
【作者简介】马诗语，硕士，主要研究方向为自然语言处理，E-mail: 781052830@qq.com
【通信作者】黄润才，博士，副教授，主要研究方向为计算机网络与信息系统、人工智能与大数据，E-mail: hrc@sues.edu.com

(Conditional Random Fields, CRF)利用特征矩阵解码整个句子的标签。Strubell等[5]提出迭代膨胀卷积神经网络(Iterated Dilated Convolutional Neural Network, IDCNN),与传统的CNN相比,此模型在大文本和结构化预测中具有更好的表现能力。对于文本的表示,传统的词表示方法有one-hot、词袋模型、n-gram[6],但这些离散的表示无法考虑词向量之间的关系。为提高模型精度,利用深度神经网络提取特征得到越来越多的关注。Mikoliv等[7]提出的word2vec词嵌入模型是目前最常用的词嵌入模型之一,然而词嵌入模型只提供一层表征,无法解决一词多义的问题。随着深度学习的发展,自然语言处理(Natural Language Processing, NLP)不再是一个任务一个模型,而是预先在大量语料上训练好一个模型,再对模型在特定的下游任务上进行微调,微调后的模型在众多NLP任务上均取得了不错的效果,如ELMO[8]、GTP[9]、BERT[10]等。李妮等[11]利用BERT与IDCNN-CRF的融合提高了NER的准确率。然而,目前这些预训练模型关注的焦点在于将模型变得更复杂,依赖越来越多的参数,很少考虑训练耗时长、成本高等问题。对此,Lan等[12]在2019年提出BERT的轻量级模型,即基于轻量型动态词向量模型(A Lite BERT, ALBERT),两个模型架构几乎一样,但ALBERT参数量相比BERT大幅度减少。即便如此,模型性能不但没有下降,反而有所提升。

鉴于先验知识对实体识别任务有良好的帮助,本研究将ALBERT与经典BILSTM-CRF相结合,提出融合ALBERT的糖尿病NER方法。

2 ALBERT-BILSTM-CRF

本研究提出的ALBERT-BILSTM-CRF模型的主要结构如图1所示。

该模型主要分为3个部分,输入表示层、序列建模层和预测解码层。输入表示部分未使用传统的人工特征,而是选择了拥有先验知识的预训练模型提取字符级表征。本研究使用的是ALBERT模型,该模型通过Embedding层把每个字映射为字向量;然后采用Transformer双向综合的考虑上下文特征进行编码,将学到的知识加到token的表示上,获得字符级别的语义信息。获取的字向量输入到序列编码层的BILSTM模块中,BILSTM考虑上下文信息并进行高维特征抽取。最后在CRF语义解码模块中预测出真实标签序列。

本研究在经典模型BILSTM-CRF的基础上进行改进,引入ALBERT模型。ALBERT模型通过对字符的掩码学习可以捕捉到字符上下文之间的语法和语义层面信息,增强字符级向量的语义表征能力。

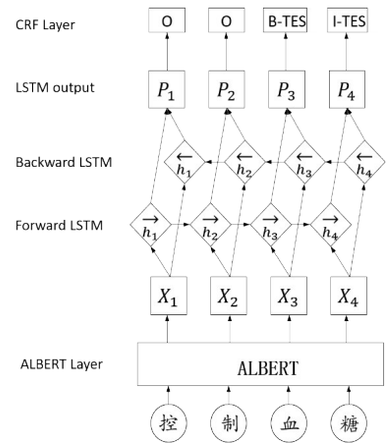

图1 ALBERT−BILSTM−CRF的模型框架

Fig.1 ALBERT−BILSTM−CRF model framework

2.1 ALBERT预训练模型

ALBERT是一种轻量级的BERT[13]。该模型架构与BERT几乎没有区别,但其所占内存仅为BERT的十分之一。为了不大幅度降低模型性能,ALBERT提出跨层参数共享机制,使用句子顺序预测(Sentence Order Prediction, SOP)训练方法代替下一句子预测(Next Sentence Prediction, NSP)训练方法。ALBERT预训练模型的结构如图2所示。

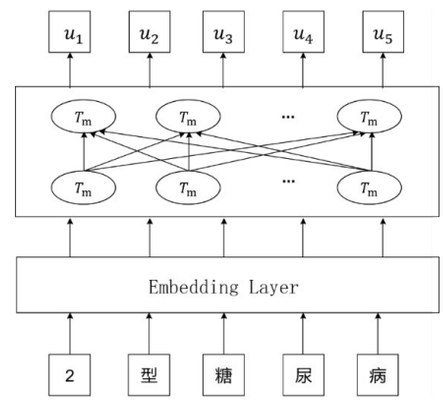

图2 ALBERT预训练模型结构

Fig.2 ALBERT pre−training model structure

ALBERT的结构是由Embedding层和Transformer encoder层组成。Embedding层通过字典将每个字符映射成字向量,输入串联的Transformer encoder层,通过预训练去捕捉语法和语义层面的信息,把文本中包含的语言知识编码到Transformer中以参数的形式体现出来。

为了训练Transformer模型中的encoder层,ALBERT设计了两个任务:掩码学习和SOP。掩码学习的基本想法是随机遮挡一或两个单词,让encoder层根据前后文预测被遮盖的单词。SOP的基本思想

是将两个句子放在一起，encoder层通过学习去判断两句话是否是原文中相邻且顺序正常的两句话。BERT的NSP任务只需要判断是否为相邻的两句话，而ALBERT在相邻的基础上更侧重于句子之间的连贯性，所以SOP在一定程度上能够解决NSP任务。ALBERT任务较BERT难度增加，也提高了多语句编码的性能。

2.2 Transformer

ALBERT是一个流程，采用双向多层的编码器Transformer。为了使模型轻量化，ALBERT在Transformer层采用参数共享的方法来减少模型存储参数量。Transformer发表于2017年，是seq2seq模型，包括encoder与decoder两部分[14]。ALBERT只采用encoder部分。encoder网络由多头Self-attention和全连接层搭建而成。Self-attention的计算公式如下：

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{1}$$

其中，$\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$是输入向量通过线性变换得到的3个输入矩阵，$d_k$为输入字向量维度。通过计算每一个输入字向量与序列其它字向量之间的关系比重大小，得到不同的权重，再将权重与所有序列的表示进行加权求和，最终获得新的字符表征。

2.3 BILSTM

由于ALBERT中的encoder部分采用自注意力结构，因此输出的特征缺少顺序性。为了得到糖尿病文本中的序列特征，本研究采用BILSTM模型对糖尿病文本的上下文信息进行建模，BILSTM网络结构与BIRNN类似，在隐含层单元采用LSTM[15]结构，其单元结构如图3所示。

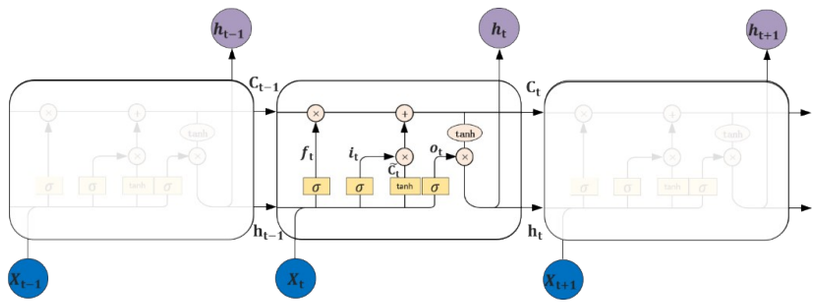

图3 LSTM单元结构

Fig.3 LSTM unit structure

LSTM是对Simple RNN模型的改进，可以避免梯度消失的问题，拥有更长的记忆。LSTM内部结构中，最重要的结构是传输带C，过去的信息直接通过传输带送到下一时刻，传输带可以避免梯度消失。

LSTM由遗忘门、输入门、$\tilde{C}_t$计算及输出门这4部分组成，它们可以选择性地让信息通过。遗忘门f_t介于0~1之间，可以有选择地让传输带C_{t-1}的值通过，假设f_t中有元素0，那么C_{t-1}中对应元素无法通过，即选择性遗忘掉一些元素；并且输入门i_t向传输带C中添加了新的信息，从而对传输带进行一轮更新。输出门o_t依赖于旧的状态向量$\boldsymbol{h}_{t-1}$和新的输入$\boldsymbol{X}_t$，计算类似于遗忘门。最后，对传输带C_t的每一个元素求双曲正切，将元素压到-1与+1之间，然后将o_t与$\tanh[C_t]$的对应元素相乘，运算符记为$\circ$，得到状态向量$\boldsymbol{h}_t$。整体流程为：

$$f_t = \sigma[\boldsymbol{W}_f \cdot [\boldsymbol{h}_{t-1},\boldsymbol{X}_t] + \boldsymbol{b}_f] \tag{2}$$

$$i_t = \sigma[\boldsymbol{W}_i \cdot [\boldsymbol{h}_{t-1},\boldsymbol{X}_t] + \boldsymbol{b}_i] \tag{3}$$

$$\tilde{C}_t = \tanh[\boldsymbol{W}_C \cdot [\boldsymbol{h}_{t-1},\boldsymbol{X}_t] + \boldsymbol{b}_C] \tag{4}$$

$$C_t = C_{t-1} \circ f_t + i_t \circ \tilde{C}_t \tag{5}$$

$$o_t = \sigma[\boldsymbol{W}_o \cdot [\boldsymbol{h}_{t-1},\boldsymbol{X}_t] + \boldsymbol{b}_o] \tag{6}$$

$$\boldsymbol{h}_t = o_t \circ \tanh[C_t] \tag{7}$$

其中，σ为Sigmoid激活函数；$\tilde{C}_t$、f_t、i_t、C_t、o_t、$\boldsymbol{h}_t$、$\boldsymbol{X}_t$分别为t时刻的输入的中间状态、遗忘门、输入门、传输带、输出门、状态向量、输入向量；$\boldsymbol{b}$为偏置向量；tanh为双曲正切函数；$\boldsymbol{W}$为模型参数矩阵。

BILSTM训练两条双向LSTM，一条从左往右，一条从右往左，两条LSTM是完全独立的，不共享参数及状态向量。BILSTM层的输出向量是由两条LSTM输出的状态向量拼接后得到的。

2.4 CRF

BILSTM层能够学习上下文信息，但不能限制前后两个标签之间的关系，输出结果相互之间没有影响。BILSTM在每一步挑选出最大的概率值作为输出标签，这样可能会出现B-label1后接入B-label2的情况；而CRF中有特征转移矩阵，可以考虑输出标签之间的顺序性，从而提高预测的准确率。

CRF是由Lafferty等[16]在2001年首次提出。CRF用于序列标注问题，是通过输入序列来预测输出序列的判别式模型。给定一组观测序列$X=\{x_1,x_2,\cdots,x_n\}$，得到预测序列标签$y=\{y_1,y_2,\cdots,y_n\}$。文本X对应的标签y的分数由转移矩阵$\boldsymbol{A}$和发射矩阵$\boldsymbol{P}$相加得到。

$$\text{score}(X,y)=\sum_{i=1}^{n}\boldsymbol{P}_{i,y_i}+\sum_{i=0}^{n}\boldsymbol{A}_{y_i,y_i+1} \tag{8}$$

其中，$\boldsymbol{A}_{y_i,y_i+1}$为从标签$y_i$到标签$y_{i+1}$的转移分数；$\boldsymbol{P}_{i,y_i}$为发射矩阵，表示第$i$个字符预测为$y_i$标签的分数。

CRF的优化目标是正确序列的概率最大化，给定一个线性链条件随机场$P(y|X)$

$$P(y|X)=\frac{\exp(\text{score}(X,y))}{\sum_{y'}\exp(\text{score}(X,y'))} \tag{9}$$

其中，y'为所有可能的状态序列集合，y为真实序列。

让真实序列的分值在所有序列的分值和中最大。最终做预测时，采用Veterbi算法，在所有状态序列中寻找分值最大的序列，即为CRF层的最终标注序列。

3 实 验

3.1 实验环境与数据

本实验在Ubuntu18.04 LTS上进行，使用语言为Python 3.7，GPU版本为华硕1070Ti GPU，显存8 G，CPU为E3-1281-V3，系统内存为16 G，tensorfolw版本为1.15GPU版。

实验数据由阿里云天池大数据平台提供，数据内容主要为基于糖尿病的相关研究论文以及糖尿病临床指南。其中共有363篇文档，约250万字，训练集与验证集以8:2的比例划分。测试集由平台独立提供的59篇糖尿病文本构成。由于数据存在一定的噪声，需要对数据进行清洗数据、句子划分等一系列预处理操作。

3.2 概念定义与标注文本

糖尿病NER旨在从大量医学文献中抽取有价值的医学知识，其知识指医学文本中有用的实体，将实体按事先定义的实体类型进行分类，即NER过程。实体类型的定义需要满足知识图谱的应用需求，本研究针对糖尿病知识图谱的应用需求，定义了15类实体类型，如表1所示。

本研究采取BIO序列标注模式，对句子中的每个字符进行标注。“B-实体类型”表示为该实体类型的实体首字符标签，“I-实体类型”表示为该实体类型除了首字符标签外实体其他符标签，“O”表示非实体标签。本次任务有15种实体类型，因此每个字符有31种标注可能性，如表2所示。

表1 实体类型定义

Tab.1 Entity type definition

实体类型	案例	实体类型	案例
程度(LEV)	轻度	病因(REA)	胰岛素抵抗
检查指标值(TSV)	下降5%	用药方法(MET)	注射
检查方法(TES)	总胆固醇	持续时间(DUR)	6个月后
部位(ANT)	肾脏	手术(OPE)	甲状腺手术
用药剂量(AMO)	500 μmol/L	用药频率(FRE)	1日1次
疾病名称(DIS)	糖尿病	临床表现(SYM)	血脂紊乱
药品名称(DRU)	视黄醇	不良反应(SID)	输液反应
非药治疗(TRE)	血液透析		

表2 实体标签定义

Tab.2 Entity tag definition

实体类型	开始标签	中间标签
程度(LEV)	B-LEV	I-LEV
检查指标值(TSV)	B-TSV	I-TSV
检查方法(TES)	B-TES	I-TES
部位(ANT)	B-ANT	I-ANT
用药剂量(AMO)	B-AMO	I-AMO
疾病名称(DIS)	B-DIS	I-DIS
药品名称(DRU)	B-DRU	I-DRU
非药治疗(TRE)	B-TRE	I-TRE
病因(REA)	B-REA	I-REA
用药方法(MET)	B-MET	I-MET
持续时间(DUR)	B-DUR	I-DUR
手术(OPE)	B-OPE	I-OPE
用药频率(FRE)	B-FRE	I-FRE
临床表现(SYM)	B-SYM	I-SYM
不良反应(SID)	B-SID	I-SID

3.3 评价指标

糖尿病NER采用精确率P、召回率R和F1值作为评价指标，具体公式如下。

$$\text{P}=\frac{\text{TP}}{\text{TP}+\text{FP}} \tag{10}$$

$$\text{R}=\frac{\text{TP}}{\text{TP}+\text{FN}} \tag{11}$$

$$\text{F1}=\frac{2\text{PR}}{\text{P}+\text{R}} \tag{12}$$

其中，TP为识别到正确实体的个数；FP为识别到非实体的个数；FN为未识别到正确实体的个数。

3.4 参数设置

本研究采用谷歌发布的ALBERT_BASE模型，其模型嵌入层尺寸128，隐藏层共12层，隐层维度768，采用12头模式，并使用gelu；BILSTM的隐藏层

节点数为100。ALBERT-BILSTM-CRF模型训练参数如下：最大序列长度为128，batch_size为32，epoch为9，ALBERT的学习率为5e-5，其他模块的学习率为0.001，dropout为0.5。

3.5 实验过程

为了证明ALBERT-BILSTM-CRF明显提升了糖尿病NER效果，本研究设计了几种方法进行对比：(1)BILSTM-CRF模型，该模型为NER的经典模型，使用word2vec词嵌入作为文本的输入表示，然后输入BILSTM-CRF模型进行编码以及预测。(2)IDCNN-CRF模型，该模型同样使用预训练好的word2vec词嵌入作为输入表示，语义编码部分使用的是IDCNN，与BILSTM-CRF模型对比。(3)BILSTM-ATT-CRF模型，在BILSTM后添加一层注意力机制，根据相关程度选择序列特征。(4)BERT-BILSTM-CRF模型，模型使用预训练模型BERT提取语义表征。

3.6 实验结果与分析

实验设置不同epoch值，研究模型随着迭代次数的拟合状况，从而确定合适的epoch值。实验结果如图4所示，横坐标为epoch值，纵坐标为实体识别的性能百分数。图中折线分别是F1、精确率和召回率的变化情况。由图可知，在第8个epoch时，F1和精确率的值分别为70.02%和71.04%，达到最高；召回率在第9个epoch时达到最优值71.12%。随着训练次数的增加，模型逐渐拟合并趋于平稳状态。综合考虑，本研究选择9为实验的epoch值。

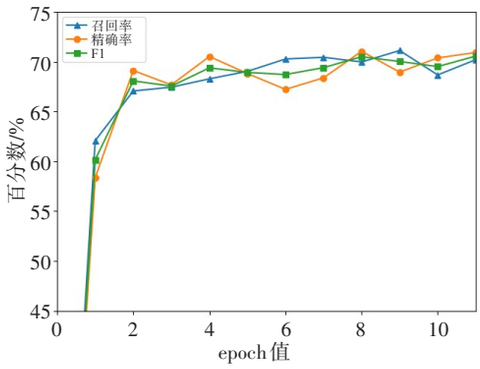

图4 ALBERT模型的各项指标变化

Fig.4 Changes in various indicators of ALBERT model

为了体现出动态词向量模型的优越性，本研究对比了在糖尿病数据集训练过程中不同模型F1值的变化情况。从图5可以看出，BILSTM-CRF和IDCNN-CRF收敛较为缓慢，在第7个epoch时开始收敛，而融合了ALBERT模型的BILSTM-CRF收敛更早速度更快，并且训练期间F1值波动较为平稳，远高于word2vec词嵌入模型的F1值，说明ALBERT提供了更深层次的语义信息，进一步证明了动态词向量的优越性。除此之外，与其他模型的训练结果见表3。

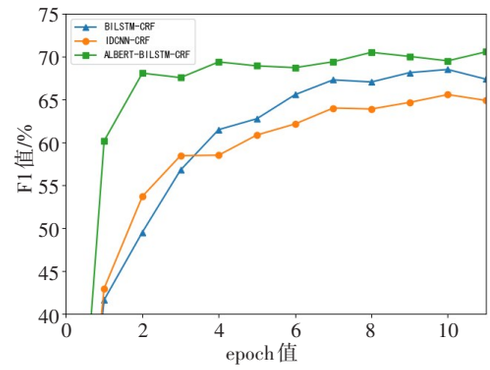

图5 不同模型的F1值对比

Fig.5 Comparison of F1 values of different model

表3 不同模型的精确率、召回率、F1值对比(%)

Tab.3 Comparison of precision, recall rate and F1 value of different models (%)

模型	精确率	召回率	F1值
BILSTM-CRF	68.81	62.04	65.25
IDCNN-CRF	69.75	55.69	61.93
BILSTM-ATT-CRF	72.26	61.54	66.47
BERT-BILSTM-CRF	66.51	68.98	67.72
ALBERT-BILSTM-CRF	68.14	67.70	67.92

由表3可知，在糖尿病数据集中，BILSTM-CRF与IDCNN-CRF实体识别的F1值分别为65.25%和61.93%，BILSTM-CRF添加注意力机制后F1值达到66.47%，较之前提升1.22%。其中，BILSTM最大的优点是具有强大的记忆力，在长文本处理中，距离较远的两个单词仍有依赖关系；IDCNN更关注实体周围的信息；加入一层注意力机制的BILSTM，注意力权重矩阵会根据每个单元特征向量对于单词的重要程度分配不同的权重，对特征向量进行加权。实验结果说明BILSTM在序列特征抽取方面优于IDCNN，基于注意力机制的模型相比无注意力机制的模型实体识别的F1值提高1.22%左右。对比模型ALBERT-BILSTM-CRF与BILSTM-CRF，F1值提高了2.67%，说明ALBERT相对于传统的word2vec词向量，可以更准确地捕捉上下文的语法和语义层面的信息，具有更好的语义表征能力。

ALBERT-BILSTM-CRF与BERT-BILSTM-CRF相比，9轮次的训练后，识别准确率分别为68.14%和66.51%，F1值分别为67.92%和67.72%；另一方面，

BERT-BILSTM-CRF 训练耗时逼近 50 min，而 ALBERT-BILSTM-CRF 模型耗时 40 min。总体看来，在各项评价指标相差不大情况下，ALBERT的模型训练效率更为出色。这是由于ABLBERT采用了参数共享机制，模型参数量仅为BERT的十分之一。在训练过程中，需要梯度更新的参数大幅度减少，训练速度加快，在NER任务上达到了相同的表现。从实验结果中可以看出，在削减参数量后，ALBERT模型识别速度提高但性能并没有下降，这也说明BERT中存在冗余参数。因此，将ALBERT引入输入表示层表现更为出色。

4 结束语

针对大多数动态词向量模型训练耗时长、资源成本高的缺点，本研究采用ALBERT模型，提出ALBERT-BILSTM-CRF，ALBERT通过参数共享机制，减少训练时长，提高实体识别的效率。该模型在糖尿病数据集上取得了良好的结果，与传统模型方法相比有所提高，但识别效果仍有提升空间，可以利用文本的其他特征，通过双通道特征融合继续进行研究。在接下来的工作中，将在糖尿病NER的基础上，进行关系抽取，对糖尿病领域知识图谱基本框架进行设计。

【参考文献】

[1] 中国2型糖尿病防治指南(2013年版)[J]. 中国医学前沿杂志(电子版), 2015, 7(3): 26-89.
Guidelines for the prevention and treatment of type 2 diabetes in China (2013 edition)[J]. Chinese Journal of the Frontiers of Medical Science (Electronic Version), 2015, 7(3): 26-89.

[2] 马建忠, 贝贝.《健康中国·2020健康体检白皮书糖尿病防控报告》出炉[EB/OL]. [2021-01-01]. https://www. sohu. com/a/437506522_161795.
MA J Z, BEI B. "Healthy China. 2020 health examination white paper diabetes prevention and control report" released[EB/OL]. [2021-01-01] https://www.sohu.com/a/437506522_161795.

[3] 刘浏, 王东波. 命名实体识别研究综述[J]. 情报学报, 2018, 37(3): 329-340.
LIU L, WANG D B. A review on named entity recognition[J]. Journal of the China Society for Scientific and Technical Information, 2018, 37(3): 329-340.

[4] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging[J]. Comput Sci, 2015, 4(1): 1508-1519.

[5] STRUBELL E, VERCA P, BELANGER D, et al. Fast and accurate entity recognition with iterated dilated convolutions[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017: 2670-2680.

[6] FORNEY G D. The viterbi algorithm[J]. Proc IEEE, 1973(3): 268-278.

[7] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [C]//Proceedings of the International Conference on Intelligent Text Processing and Computational Linguistics. Berlin: ACM, 2013: 3111-3119.

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

[9] RADFORD A, NARASINMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[EB/OL]. [2019-12-03]. https://www. docin. com/p-2176538517.html.

[10] DEVLIN J, CHANG M W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[C]//NAACL-HLT 2019: Annual Conference of the North American Chapter of the Association for Computational Linguistics, Minneapolis, USA, 2019: 4171-4186.

[11] 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.
LI N, GUAN H M, YANG P, et al. BERT-IDCNN-CRF for named entity recognition in Chinese[J]. Journal of Shandong University (Natural Science) , 2020, 55(1): 102-109.

[12] LAN Z, CHEN M, GOODMAN S, et al. ALBERT: a lite BERT for self-supervised learning of language representations [C]// Proceedings of the 8th International Conference on Learning Representations. La Jolla, CA: ICLR, 2020: 1-17.

[13] 徐菲菲, 冯东升. 文本词向量与预训练语言模型研究[J]. 上海电力大学学报, 2020, 36(4): 320-328.
XU F F, FENG D S. A survey of research on word vectors and pretrained language models[J]. Journal of Shanghai University of Electric Power, 2020, 36(4): 320-328.

[14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Advances in Neural Information Processing Systems, 2017: 5998-6008.

[15] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Comput, 1997, 9(8): 1735-1780.

[16] LAFFERTY J, MCCALLUM A, PEREIRA F, et al. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the 18th International Conference on Machine Learning. USA: Morgan Kaufmann Publishers, 2001: 282-289.

（编辑：谭斯允）