

基于注意力机制的Bi-LSTM结合CRF的新闻命名实体识别及其情感分类

胡甜甜^{1*}, 但雅波², 胡杰³, 李想², 李少波²

(1. 贵州大学 计算机科学与技术学院, 贵阳 550025; 2. 贵州大学 机械工程学院, 贵阳 550025;

3. 贵州财经大学 大数据统计学院, 贵阳 550025)

(* 通信作者电子邮箱 2506062823@qq.com)

摘要:针对搜狐coreEntityEmotion_train语料核心实体识别和核心实体情感分析的任务,提出了基于注意力机制的长短期记忆神经网络结合条件随机场模型(AttBi-LSTM-CRF)。首先,对文本进行预训练,将每个字映射为维度相同的低维向量;然后,把这些向量输入到基于注意力机制的长短期记忆神经网络(AttBi-LSTM)中,以获取长远的上下文信息并集中注意力到与输出标签高度相关的信息上;最后,通过条件随机场(CRF)层获取整个序列的最优标签。将AttBi-LSTM-CRF模型与双向长短期记忆神经网络(Bi-LSTM)、AttBi-LSTM和双向长短期记忆神经网络结合条件随机场(Bi-LSTM-CRF)模型进行对比实验。实验结果表明,AttBi-LSTM-CRF模型的准确率达到0.786,召回率达到0.756, F1值达到0.771,优于对比模型,验证了AttBi-LSTM-CRF性能的优越性。

关键词:核心实体识别;情感分类;条件随机场;注意力机制;双向长短期记忆神经网络

中图分类号: TP391.1 **文献标志码:** A

News named entity recognition and sentiment classification based on attention-based bi-directional long short-term memory neural network and conditional random field

HU Tiantian^{1*}, DAN Yabo², HU Jie³, LI Xiang², LI Shaobo²

(1. School of Computer Science and Technology, Guizhou University, Guiyang Guizhou 550025, China;

2. School of Mechanical Engineering, Guizhou University, Guiyang Guizhou 550025, China;

3. College of Big Data Statistics, Guizhou University of Finance and Economics, Guiyang Guizhou 550025, China)

Abstract: Attention-based Bi-directional Long Short-Term Memory neural network and Conditional Random Field (AttBi-LSTM-CRF) model was proposed for the corpus core entity recognition and core entity sentiment analysis task of Sohu coreEntityEmotion_train. Firstly, the text was pre-trained, each word was mapped into a low-dimensional vector with the same dimension. Then, these vectors were input into the Attention-based Bi-directional Long Short-Term Memory neural network (AttBi-LSTM) to obtain the long-term context information and focus on the information highly related to the output label. Finally, the optimal label of the entire sequence was obtained through the Conditional Random Field (CRF) layer. The comparison experiments were conducted among AttBi-LSTM-CRF model, Bi-directional Long Short-Term Memory neural network (Bi-LSTM), AttBi-LSTM and Bi-directional Long Short-Term Memory neural network and Conditional Random Field (Bi-LSTM-CRF) model. The experimental results show that, the accuracy of AttBi-LSTM-CRF model is 0.78, the recall is 0.667, and the F1 value is 0.553, which are better than those of the comparison models. The superiority of AttBi-LSTM-CRF performance is verified.

Key words: core entity recognition; sentiment classification; Conditional Random Field (CRF); attention mechanism; Bi-directional Long Short-Term Memory neural network (Bi-LSTM)

0 引言

21世纪是一个数据爆炸式增长的时代,也是各种媒体应运而生的时代。新闻、教育、医疗等不同领域每时每刻都在产生大量的文本数据,往往这些数据中蕴藏着大量有价值的信

息,对社会性的研究有着非常重要的意义^[1]。面对这些海量文本数据,如何准确高效地进行信息抽取和数据挖掘成为学术界和工业界关注的热点问题,作为其中主要技术的命名实体识别(Named Entity Recognition, NER)技术受到研究者们的高度重视。NER是指从文本中识别出命名性指称项,为关

收稿日期:2019-11-18;修回日期:2020-01-05;录用日期:2020-01-16。

基金项目:国家自然科学基金重大研究计划培育项目(91746116)。

作者简介:胡甜甜(1993—),女,湖南岳阳人,硕士研究生,CCF会员,主要研究方向:自然语言处理、大数据挖掘; 但雅波(1993—),男,湖北天门人,硕士研究生,主要研究方向:材料信息学、大数据挖掘; 胡杰(1990—),男,湖南岳阳人,副教授,博士,主要研究方向:自然语言处理; 李想(1994—),男,重庆人,硕士研究生,主要研究方向:材料信息学; 李少波(1973—),男,湖南岳阳人,教授,博士,主要研究方向:智能制造、大数据挖掘。

系抽取等任务做铺垫,是构建知识图谱的最初步骤^[2]。狭义上,NER是识别出人命、地名和组织机构名这三类命名实体(时间、货币名称等构成规律明显的实体类型可以用正则表达式等方式识别)。NER方法主要用三种:基于规则的方法、基于特征模板的方法、基于神经网络的方法。基于规则的方法常用的是利用手工编写的规则,将文本与规则进行匹配来识别出命名实体^[3]。例如,对于中文来说,“说”“老师”等词语可作为人名的下文,“大学”“医院”等词语可作为组织机构名的结尾,还可以利用到词性、句法信息。在构建规则的过程中往往需要大量的语言学知识,不同语言的识别规则不尽相同,而且需要谨慎处理规则之间的冲突问题;此外,构建规则的过程费时费力、可移植性不好。基于特征模板的方法利用大规模语料来学习出标注模型,从而对句子的各个位置进行标注,通常与条件随机场(Conditional Random Field, CRF)^[4]结合使用。特征模板通常是人工定义的一些二值特征函数,试图挖掘命名实体内部以及上下文的构成特点。对于句子中的给定位置来说,提取特征的方式是采用一个特征模板在该位置上进行数学运算。而且,不同的特征模板(窗口)之间可以进行组合来形成一个新的特征模板。CRF的优点在于其为一个位置进行标注的过程中可以利用到此前已经标注的信息,利用Viterbi解码来得到最优序列^[5]。对句子中的各个位置提取特征时,满足条件的特征取值为1,不满足条件的特征取值为0;然后,把特征喂给CRF,训练阶段建模标签的转移,进而在预测阶段为测试句子的各个位置做标注。近年来,人工智能(Artificial Intelligence, AI)取得了突破性的进展。其中,机器学习(Machine Learning, ML)和深度学习(Deep Learning, DL)技术的应用为人类在各个领域的任务带来了优异表现,包括图像识别^[6]、语音识别^[7]和自然语言处理(Natural Language Processing, NLP)^[8]。神经网络方法使得模型的训练成为一个端到端的整体过程,而非传统的Pipeline,不依赖特征工程,是一种数据驱动的方法。

研究者们用神经网络在自然语言处理方面做了大量研究: Huang等^[9]提出将多种神经网络模型应用到自然语言处理中的序列标注问题上,并证明双向长短期记忆神经网络结合条件随机场(Bi-directional Long Short-Term Memory neural network and Conditional Random Field, Bi-LSTM-CRF)模型在序列标注上能取得很好的结果,在CoNLL corpus语料库上进行命名实体识别时,F1值达到了90.10%。Ma等^[10]将长短期记忆神经网络(Long Short-Term Memory neural network, LSTM)、卷积神经网络(Convolutional Neural Network, CNN)、CRF结合构建出Bi-LSTM-CNNs-CRF模型,并应用在命名实体识别任务上,在CoNLL 2003 corpus语料库取得F1得分为0.9121的成绩。Luo等^[11]提出了一个自我注意力的双向长短期记忆网络(Self-Attentive Bi-LSTM)来预测情绪挑战中的多种情绪,Self-Attentive Bi-LSTM模型能够捕获文本之间的上下文依赖关系,有助于对模糊情绪进行分类,它在Friends和EmotionPush测试集中分别获得了59.6和55.0的未加权准确性分数。虽然这些方法在NLP的NER任务和情感分析方面取得了显著的效果,但这些方法都是将实体识别与情感分析分别建立预测模型,没有挖掘这两个任务之间的联系。

针对上述问题,本文提出了一种基于注意力机制的双向长短期记忆神经网络结合条件随机场(Attention-based Bi-directional Long Short-Term Memory neural network and Conditional Random Field, AttBi-LSTM-CRF)的深度神经模型来同时完成对新闻中核心实体的提取和核心实体的情感分类,并与双向长短期记忆神经网络(Bi-directional Long Short-Term Memory, Bi-LSTM)、带注意力机制的双向长短期记忆神经网络(Attention-based Bi-directional Long Short-Term Memory neural network, AttBi-LSTM)、Bi-LSTM-CRF等算法进行了比较。实验结果表明,本文采用的AttBi-LSTM-CRF在同时识别核心实体及分类其情感时取得了最好的结果,准确率为0.786,召回率为0.756,F1值为0.771。

1 数据预处理与数据标定

1.1 数据来源

本文模型实验数据集来源于2019搜狐校园算法大赛中的比赛数据(<https://biendata.com/competition/sohu2019/data/>),其中包含4万条训练数据(带标签)和4万条测试数据(不带标签)。该数据集文本主要由中文构成,偶尔会出现几个英文单词,每条数据包含一篇不限定主题的新闻标题和新闻主体内容,训练数据同时会给出新闻主题内容对应的1到3个的核心实体以及这些核心实体相对应的情感标签。该数据集来自于搜狐新闻文本,包括娱乐、情感、体育、旅游、时政、时尚、财经等各种常见新闻题材,涵盖内容广泛,种类丰富,语法规则,数据中的核心实体标注和情感分析都由人工标注。

1.2 数据预训练

本文选取北京师范大学中文信息处理研究所与中国人民大学DBIR实验开源的中文词向量语料库Chinese Word Vectors(<https://github.com/Embedding/Chinese-Word-Vectors>)中的搜狐新闻(Sogou News)(<http://www.sogou.com/labs/resource/cs.php>)中文预训练词向量包对每个句子进行预训练,将每个字映射成300维度的向量。搜狐中文预训练词向量包是通过大量搜狐新闻文章用ngram2vec工具包训练得到。由于本文的数据来自于搜狐新闻文本和搜狐新闻文章在各个方面都有非常大的相似度,因此选用搜狐新闻中文预训练词向量包对数据进行预训练会比较适合。最后,每个句子将会用矩阵 $T^{n \times d}$ 表示, d 表示词向量的维度(300), $n=[Max_length \times 0.8]$, $[\cdot]$ 表示取数值的整数部分, Max_length 表示语料库中最大句子长度。对于句子长度小于 Max_length 的用0填充。

1.3 数据标注

在本文实验中,模型训练集的标签是给每个字向量标注上对应类别标号,标注标签一共有10类,分别是:①Null,②B_pos,③I_pos,④E_pos,⑤B_neg,⑥I_neg,⑦E_neg,⑧B_norm,⑨I_norm,⑩E_norm。其中:空标签(Null)表示非核心实体;B表示核心实体的第一个字;I表示核心实体词中间的字;E表示核心实体的最后一个字;pos表示积极的情感;neg表示消极的情感;norm表示中立的情感。例如,B_pos标签表示的是积极情感的核心实体的第一个字。标注实例如表1所示。

表 1 标注实例
Tab. 1 Labeling example

字序列	标签序列	标签标号
易	B_pos	2
烱	I_pos	3
千	I_pos	3
玺	E_pos	4
真的	Null	1
的	Null	1
非	Null	1
常	Null	1
适	Null	1
合	Null	1
出	Null	1
现	Null	1
在	Null	1
这	Null	1
个	Null	1
位	Null	1
置	Null	1

2 模型构建

2.1 AttBi-LSTM

AttBi-LSTM 模型将 Attention 机制^[12]融合到 Bi-LSTM^[13]中。深度学习中,Attention 机制可以理解为将注意力放在更重要的信息上,它与 Bi-LSTM 融合的基本思想是:打破了传统 Bi-LSTM 结构在编解码时都依赖于内部一个固定长度向量的限制。AttBi-LSTM 机制的实现是通过保留 Bi-LSTM 编码器对输入序列的中间输出结果,然后训练一个模型来对这些输入进行选择性的学习,并且在模型输出时将输出序列与之进行关联。图 1 为 AttBi-LSTM 的模型框架。

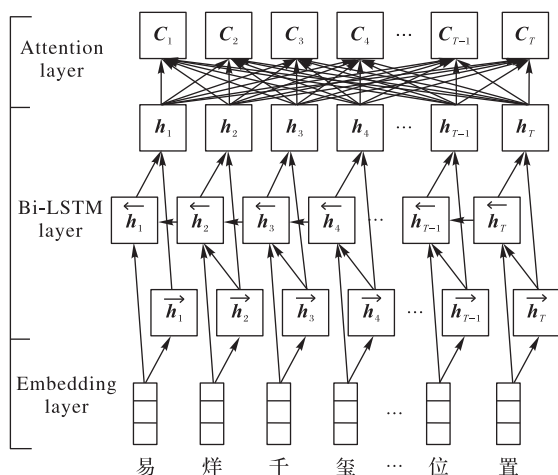


图 1 AttBi-LSTM 模型框架

Fig. 1 AttBi-LSTM model framework

对于时间序列标记任务来说,每一个时刻的历史信息和未来信息的特征对于当前实体标签的预测都很重要,然而标准的 LSTM 并不能捕获未来信息的特征。Bi-LSTM 模型将前向的 LSTM 和后向的 LSTM 结合,具有能够捕获前后信息特征的作用,因此,本文采用了 Bi-LSTM 模型,其输出可以表示为 $h_i = [\vec{h}_i \otimes \overleftarrow{h}_i]$ 。Bi-LSTM 层输入的向量结合表示为 $H: [h_1, h_2, \dots, h_T]$ 。Attention 层权重矩阵由 $M = \tanh(H)$ 、 $\alpha =$

$\text{softmax}(W^T M)$ 、 $\gamma = H\alpha^T$ 的方式得到。其中, $H \in \mathbb{R}^{d \times T}$, d 为词向量的维度, W^T 是一个训练学习得到的参数向量的转置,最后 AttBi-LSTM 的输出为 $h^* = \tanh(r)$ 。

2.2 CRF

AttBi-LSTM 模型最终的输出是相互独立的,AttBi-LSTM 学习到了输入中前后信息的特征,但是没有利用输出标签的作用。CRF 也是一种序列建模算法,它综合了隐马尔可夫模型和最大熵模型的优点。它根据给定观察序列推测对应的状态序列,可以利用相邻前后的标签关系来获取当前的最优的标记。因此,本文在 AttBi-LSTM 的输出层后叠加一层线性 CRF 来标注核心实体及其情感分析的类别。

定义矩阵 $P^{n \times m}$ 为 AttBi-LSTM 层的输出, n 在 1.2 节中已经进行了定义, m 表示标签类别的个数, P_{ij} 表示句中第 i 个字是第 j 个标签的概率。定义状态转移矩阵 $A^{(m+2) \times (m+2)}$, 其中, A_{ij} 表示在连续的一段时间内,第 i 个标签转移到第 j 个标签的概率。

对于预测序列 y 的概率可以表示为: $K(h, y) = \sum_{i=0}^n A_{y_i, y_{i+1}} + \sum_{i=1}^n P_{i, y_i}$, 再通过 softmax 层计算出所有类别标签的概率。

2.3 核心实体识别与情感分析框架

图 2 为本文的 AttBi-LSTM-CRF 中文核心实体识别及其情感分析模型框架。AttBi-LSTM-CRF 框架共分为 4 步: 1) 预训练^[14]部分, 将待训练的文本序列进行文本向量化, 将其每个字转换为对应的有特定意义的固定长度的向量。2) 将处理好的词向量序列输入 Bi-LSTM, 提取文本双向长距离依赖特征。3) 通过 Attention 机制, 提取输入和输出之间的相关性进行重要度计算, 根据重要度获取文本整体特征。4) 用线性 CRF 层处理标签之间的状态关系, 得到全局最优标注序列。

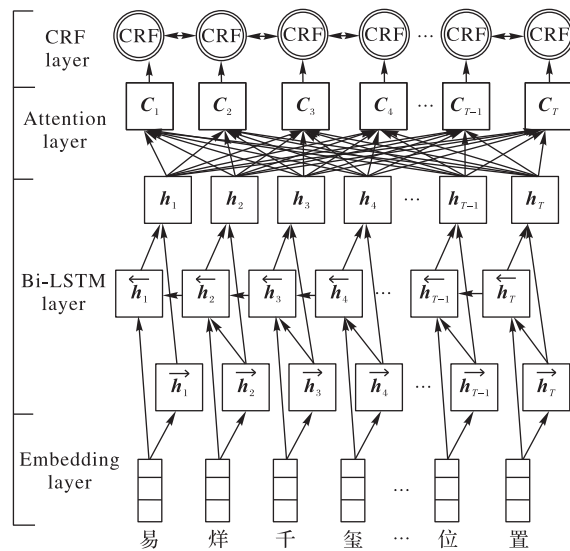


图 2 AttBi-LSTM-CRF 模型框架

Fig. 2 AttBi-LSTM-CRF model framework

3 实验与结果分析

3.1 实验环境

本文的实验环境为: 操作系统 Ubuntu 16.04, CUDA 8.0, cudnn 6; 处理器 4 个 CPU 核心, 1 颗 Nvidia Tesla P100 共享 GPU 核; 内存 60 GB, 显存 16 GB; 编译平台 Pycharm Profession, Python 3.5, TensorFlow 1.5.1。

3.2 实验参数设置

本文采用网格搜索法(grid search)进行主要参数调节,获取模型的最优参数集合,模型参数取值及其说明如表2所示。其中,模型的部分超参数主要是来自于现有研究中的经验,如学习速率、Dropout 比例;一些参数是由数据集的特性而设置,如最长序列长度取值范围根据句子长度统计得到;一些参数根据模型训练和硬件的条件配置,如每批数据量大小、LSTM 隐藏层单元数。

表2 AttBi-LSTM-CRF 模型的超参数

Tab. 2 Hyperparameters of AttBi-LSTM-CRF model

参数	取值	参数	取值
批次大小	32	Dropout	0.9
LSTM size	128	学习率	0.001
代数	20	学习率衰减	1.0
Embedding dim	300		

3.3 实验结果评估方法

训练好的模型输出结果是文本对应的标签,连续的 BIE 标签对应的词表示一个核心实体,其中,可能一个核心实体会几种不同的情感标签,这时候通过投票机制,即少数服从多数,得到这个核心实体的情感分析。

本文模型效果评估指标有准确率(Precision)、召回率(Recall)和 F1 值(F1-score)^[15]。每个指标都有实体词的得分和实体情感的得分两部分,计算每个样本单独的指标值,然后取所有样本指标值的平均数作为最后的结果。情感分析的指标值由实体_情绪的组合标签进行判断,只有实体与情绪都正确才算正确的标签。其中,两条实例数据的预测结果如表3所示,对应的预测得分如表4所示。

表3 预测标注实例

Tab. 3 Predictive labeling examples

文章ID	真实实体	预测实体	真实情感	预测情感
0	a,b,c	a,b,c	pos,pos,pos	neg,neg,neg
1	d,e,f	d,e	pos,pos,neg	neg,pos

表4 预测得分

Tab. 4 Predictive scores

文章ID	准确率		召回率		F1 值	
	实体	情感	实体	情感	实体	情感
0	1.000	0.667	1.000	0.667	1.000	0.667
1	1.000	0.500	0.667	0.333	0.800	0.400

3.4 实验对比分析

3.4.1 不同模型的性能对比

为了验证 AttBi-LSTM-CRF 模型中每个模块的作用,本文选择了 Bi-LSTM、Bi-LSTM-CRF、AttBi-LSTM 这三种模型进行相同的实验作为实验对照。这些模型的参数与 AttBi-LSTM-CRF 模型使用的参数相同。为使实验结果更加稳定可靠,每种模型都将重复实验 10 次,并取 10 次实验的平均值作为最终的结果,如表5所示。

由表5可知,与 AttBi-LSTM 相比,Bi-LSTM 的结果较差,原因是其没有引入注意力机制,无法捕捉到核心实体之间的依赖关系。AttBi-LSTM 和 Bi-LSTM-CRF 的结果相当,注意力机制的引入能有效解决长文本间的长期依赖关系,条件随机场(CRF)通过求解最大概率得到最优序列,能很好地标注序列标签。本文提出的 AttBi-LSTM-CRF 模型在准确率、召回率、

F1 值上取得了最好的结果。模型训练时的 F1 值变化和召回率变化如图3~4所示。由图3和图4可知,本文提出的 AttBi-LSTM-CRF 模型的 F1 值、召回率和收敛速度均取得了最好的结果

表5 不同模型的准确率、召回率和 F1 值对比

Tab. 5 Comparison of accuracy, recall rate and F1 value among different models

模型	准确率	召回率	F1 值
Bi-LSTM	0.741	0.665	0.700
AttBi-LSTM	0.754	0.683	0.717
Bi-LSTM-CRF	0.752	0.691	0.720
AttBi-LSTM-CRF	0.786	0.756	0.771

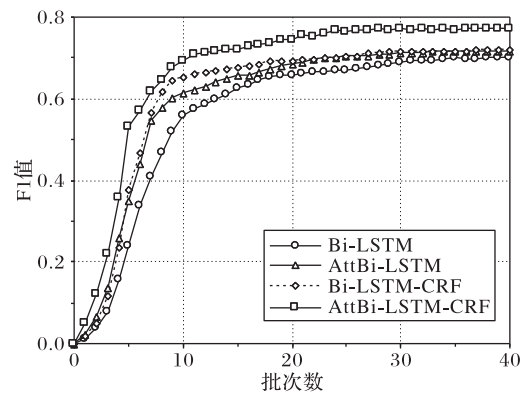


图3 模型训练时的 F1 值变化

Fig. 3 Change of F1 value during model training

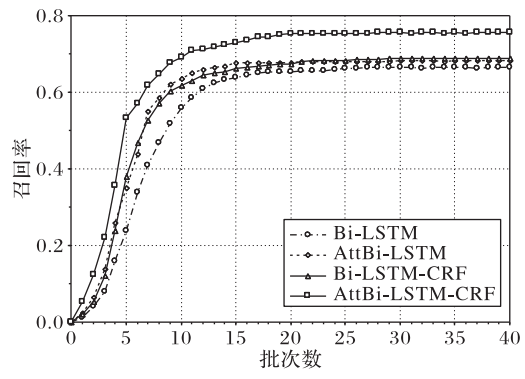


图4 模型训练时的召回率变化

Fig. 4 Change of recall rate during model training

3.4.2 Attention 模块

从表5中可以看出:AttBi-LSTM 模型的准确率为 0.754,召回率为 0.683,F1 值为 0.717,相较 Bi-LSTM 模型 F1 值高出 0.017,准确率和召回率也有相应的提升。比较 Bi-LSTM 模型和 AttBi-LSTM 模型的实验数据可知,AttBi-LSTM 模型在 coreEntityEmotion_train 语料的核心实体识别及其情感分析任务上的表现更好。这是由于采用的 Bi-LSTM 模型存在一个问题:不论输入长短都将其编码成一个固定长度的向量表示,这使模型对于长输入序列的学习效果很差。而 Attention 机制则克服了上述缺陷,原理是 Attention 模块能高效地分配有限的注意力资源,有选择性地注意力投放到高质量的信息上,在模型输出时会选择性地专注考虑输入中对应高度相关的信息。获取的信息价值越高,实验的结果越好。

3.4.3 线性 CRF 模块

表5给出了 Bi-LSTM 模型和 Bi-LSTM-CRF 模型的实验结

果对比。Bi-LSTM-CRF模型的实验结果数据是准确率为0.752,召回率为0.691,F1值为0.720,相较Bi-LSTM模型F1值高出0.020,准确率和召回率也有相应的提升,表明线性CRF模块可以获得更好的效果。线性CRF是一种序列标注模型,它和LSTM等分类器标注模型不同,它考虑的不是长远的上下文信息,它是计算某个序列中的最优联合概率,优化整个序列(最终目标)。

3.4.4 AttBi-LSTM-CRF模型

通过实验结果数据对比可以看到,本文使用的AttBi-LSTM-CRF模型相较于传统的Bi-LSTM,核心实体的识别及其情感分类的准确性有了大幅度的提升,相较于在Bi-LSTM基础上只增加Attention或者线性CRF任意一种模块的模型,也有更优的结果。

4 结语

针对搜狐coreEntityEmotion_train语料核心实体识别和核心实体情感分析的任务,本文提出了AttBi-LSTM-CRF模型。Bi-LSTM网络可以获取长远的上下文信息对文本标注产生影响,加入Attention模块可以从输入中获取与输出的标注有关的重要信息,在AttBi-LSTM层后加上线性CRF层获取整个序列最优标注。在搜狐coreEntityEmotion_train语料上进行的对比实验结果表明,本文使用的AttBi-LSTM-CRF模型在核心实体识别和核心实体的情感分析任务上取得了较高的准确值、召回率和F1值,相较Bi-LSTM、AttBi-LSTM、Bi-LSTM-CRF三种模型有一定的优越性。最近Devlin等^[16]提出了基于BERT的神经语音模型,该模型在上10种自然语言任务中都取得了最好的结果,因此,可以使用本文的数据对BERT进行迁移训练后改变输出层的网络结构进行核心实体的识别与情感分类。

参考文献 (References)

- [1] 冯贵兰,李正楠,周文刚. 大数据分析技术在网络领域中的研究综述[J]. 计算机科学, 2019, 46(6): 1-20. (FENG G L, LI Z N, ZHOU W G. Research on application of big data analytics in network[J]. Computer Science, 2019, 46(6): 1-20.)
- [2] HABIBI M, WEBER L, NEVES M, et al. Deep learning with word embeddings improves biomedical named entity recognition[J]. Bioinformatics, 2017, 33(14): i37-i48.
- [3] 刘璟. 中文命名实体识别方法研究[J]. 电脑知识与技术, 2019, 15(9): 179-180. (LIU J. Research on Chinese named entity recognition[J]. Computer Knowledge and Technology, 2019, 15(9): 179-180.)
- [4] LONG S, YUAN R, YI L, et al. A method of Chinese named entity recognition based on CNN-BiLSTM-CRF model[C]// Proceedings of the 2018 International Conference of Pioneering Computer Scientists, Engineers and Educators, CCIS 902. Singapore: Springer, 2018: 161-175.
- [5] 祖木然提古丽·库尔班,艾山·吾买尔. 中文命名实体识别模型对比分析[J]. 现代计算机, 2019(14): 3-7. (KUERBAN Z, WUMAIER A. Analysis and comparison of Chinese named entity recognition model[J]. Modern Computer, 2019(14): 3-7.)
- [6] 郑远攀,李广阳,李晔. 深度学习在图像识别中的应用研究综述[J]. 计算机工程与应用, 2019, 55(12): 20-36. (ZHENG Y P, LI G Y, LI Y. Survey of application of deep learning in image recognition[J]. Computer Engineering and Applications, 2019, 55(12): 20-36.)
- [7] 王蔚,胡婷婷,冯亚琴. 基于深度学习的自然与表演语音情感识别[J]. 南京大学学报(自然科学版), 2019, 55(4): 660-666. (WANG W, HU T T, FENG Y Q. Speech emotion recognition in nature and scripted state based on deep learning[J]. Journal of Nanjing University (Natural Science), 2019, 55(4): 660-666.)
- [8] XING F Z, CAMBRIA E, WELSCH R E. Natural language based financial forecasting: a survey[J]. Artificial Intelligence Review, 2017, 50(1): 49-73.
- [9] HUANG Z, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging [EB/OL]. [2019-05-05]. <https://hxlwlf.github.io/assets/Bidirectional%20LSTM-CRF%20Models%20for%20Sequence%20Tagging.pdf>.
- [10] MA X, HOVY E. End-to-end sequence labeling via bi-directional LSTM-CNNs-CRF [EB/OL]. [2019-05-09]. <https://arxiv.org/pdf/1603.01354.pdf>.
- [11] LUO L, YANG H, CHIN F Y L. EmotionX-DLC: self-attentive BiLSTM for detecting sequential emotions in dialogue [C]// Proceedings of the 6th International Workshop on Natural Language Processing for Social Media. Stroudsburg: Association for Computational Linguistics, 2018: 32-36.
- [12] ZHANG H, GOODFELLOW I, METAXAS D, et al. Self-attention generative adversarial networks [EB/OL]. [2019-05-21]. <http://www.ivs-n-group.com/seminar/2019/PDF/LYF20190530.pdf>.
- [13] ZENG Y, YANG H, FENG Y, et al. A convolution BiLSTM neural network model for Chinese event extraction [C]// Proceedings of the 2016 International Conference on Computer Processing of Oriental Languages, LNCS 10102. Cham: Springer, 2016: 275-287.
- [14] SHAKUROVA L, NYARI B, LI C, et al. Best practices for learning domain-specific cross-lingual embeddings [C]// Proceedings of the 4th Workshop on Representation Learning for NLP. Stroudsburg: Association for Computational Linguistics, 2019: 230-234.
- [15] 杨朔,陈丽芳,石瑀,等. 基于深度生成式对抗网络的蓝藻语义分割[J]. 计算机应用, 2018, 38(6): 1554-1561. (YANG S, CHEN L F, SHI Y, et al. Semantic segmentation of blue-green algae based on deep generative adversarial net[J]. Journal of Computer Applications, 2018, 38(6): 1554-1561.)
- [16] DEVLIN J, CHANG M-W, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2019-05-21]. <https://arxiv.org/pdf/1810.04805.pdf>.

This work is partially supported by the Major Research Program Cultivation Project of the National Natural Science Foundation of China (91746116).

HU Tiantian, born in 1993, M. S. candidate. Her research interests include natural language processing, big data mining.

DAN Yabo, born in 1993, M. S. candidate. His research interests include material informatics, big data mining.

HU Jie, born in 1990, Ph. D., associate professor. His research interests include natural language processing.

LI Xiang, born in 1994, M. S. candidate. His research interests include material informatics.

LI Shaobo, born in 1973, Ph. D., professor. His research interests include intelligent manufacturing, big data mining.