基于BERT的混合神经网络实体识别方法

王卫红,吕红燕,曹玉辉,霍 峥
(河北经贸大学 信息技术学院,河北 石家庄 050061)

摘 要:针对命名实体识别方法中语义分析不足及准确率较低的问题,提出一种基于BERT模型的混合神经网络实体识别方法。对命名实体识别研究现状进行了调查与分析,发现现有命名实体识别研究中存在数据分析与特征提取不充分导致准确率较低的问题。利用BERT预训练语言模型动态生成字的语义向量,丰富其文本特征。使用卷积神经网络(convolutional neural network,CNN)模型再次抽取语义特征,实现语义的自动抽取,二者联合作为下一步的输入向量。采用引入注意力机制的双向长短时记忆网络(bi-directional long short-term memory,BiLSTM)获取单个字在字符级别上前后两个方向上的信息。通过条件随机场(conditional random field,CRF)模型解码序列标签,得到全局最优标注序列。在《人民日报》和MSRA两个数据集上的实验结果表明,该方法相比于其他模型,能有效地获取语义信息,在准确率、召回率和F1值上均有所提升。

关键词:命名实体识别;BERT模型;卷积神经网络;双向长短期记忆网络;条件随机场
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2021)08-0100-06
doi:10.3969/j.issn.1673-629X.2021.08.017

A Hybrid Neural Network Entity Recognition Method Based on BERT Model

WANG Wei-hong,LYU Hong-yan,CAO Yu-hui,HUO Zheng
(School of Information Technology,Hebei University of Economics and Business,Shijiazhuang 050061,China)

Abstract:Aiming at the problem of insufficient semantic analysis and low accuracy in named entity recognition method,a hybrid neural network entity recognition method based on BERT model is proposed. The research status of named entity recognition was investigated and analyzed,and it was found that the problem of low accuracy resulted from insufficient data analysis and feature extraction existed in the research of named entity recognition. The semantic vector of the word is generated dynamically by using BERT pre-training language model to enrich its text features. The semantic features are extracted again using the convolutional neural network (CNN) model to realize the automatic semantic extraction,and the two are combined as the next step of the input vector. BiLSTM is used to obtain the information of a single word in two directions before and after the character level. The conditional random field (CRF) model was used to decode the sequence tags and obtain the global optimal labeling sequence. Experiments on two data sets of People's Daily and MSRA show that compared with other models,the proposed method can effectively obtain semantic information,and it is improved in accuracy,recall rate and F1 value.

Key words:named entity recognition;BERT model;convolutional neural network;bi-directional long short-term memory;conditional random field

0 引 言

随着社会信息化进程的飞速发展,信息呈爆炸式增长,各类数据海量存在,其中文本数据也不例外。而文本数据中常常包含了大量有价值的信息,尤其是文本中的实体是句子的主体,包含了丰富的语义信息,因此命名实体识别任务在文本数据的理解与处理过程中具有非常重要的意义。除此之外,命名实体识别是信息抽取中的基础任务,而信息抽取是知识图谱构建中的重要步骤。近几年来,知识图谱的发展使得命名实体识别工作更为重要[1]。

命名实体识别[2]旨在识别出文本中的专有名词并将其划分到相应的实体类型中。其中常见的命名实

收稿日期:2020-09-24 **修回日期**:2021-01-25
基金项目:国家自然基金项目(62002098);河北省自然科学基金(F2020207001);河北经贸大学科学研究与发展计划基金项目(2021ZD03)
作者简介:王卫红(1970-),女,教授,博士,CCF会员(34423M),研究方向为知识图谱、移动协同计算;吕红燕(1994-),女,硕士研究生,CCF会员(A3847G),研究方向为知识图谱、自然语言处理。

体包括人名、地名、机构名等。命名实体技术从开始发展至今,可以将其分为三大阶段,基于词典和规则的方法、基于传统机器学习和基于深度学习的方法、现在热门的注意力机制和图神经网络等方法应用于命名实体识别中。命名实体识别技术发展得越来越成熟。早期的基于规则的命名实体识别方法主要是通过人工来构建规则库,再从文本中寻找匹配这些规则的字符串从而识别出文本中的命名实体。这种方法在特定的语料上可以获得较高的识别效果,但是不具有通用性,迁移能力较差,而且规则库的构建需要大量的人力,耗费时间长。

随着机器学习在自然语言处理领域的兴起,命名实体识别的研究逐渐转向基于统计机器学习的方法,主要分为两种思路,一种是先识别出命名实体的边界,然后将命名实体进行分类,另一种是序列化标注方法[3-4]。序列化标注方法是目前最为有效,也是最为普遍的一种命名实体识别方法。近年来,基于神经网络模型的深度学习技术不断发展,成为机器学习领域新的热潮。各类神经网络模型被用到命名实体识别的研究中。

文中提出的基于 BERT 模型的混合神经网络实体识别方法结合预训练语言模型的同时充分利用各类神经网络的优势,来获取句子、实体中更加丰富的语义信息,以提高命名实体识别的有效性和通用性。

1　相关研究

命名实体识别任务在 1991 年第一次被提出,之后在很多会议中将其作为评测任务,例如 MUC-6、MUC-7、CoNLL-2002、CoNLLC-2003 等会议。许多学者对命名实体识别任务进行研究。

近些年来,命名实体识别常常被看作是序列标注问题,在标注语料上进行监督学习。早期,经典机器学习分类模型被成功地用来进行命名实体的序列化标注,而且获得了较好的效果,如条件随机场 CRF[5]、最大熵 ME[6] 和最大熵马尔可夫模型 MEMM[7] 等。Collobert 等学者[8]在 2011 年首次将神经网络应用于命名实体识别任务中,提出了基于神经网络的命名实体识别方法。此后,随着深度神经网络的发展,越来越多的学者将神经网络模型运用到命名实体识别任务中。GUL Khan Safi Qamas 等[9]提出了一种基于深度神经网络、结合长短时记忆和注意力机制的命名实体识别方法,提高了命名实体识别的准确率。N. Bölücü 等[10]将双向 LSTM-CNN 模型进行了扩展,添加了句法和词级特征,并通过实验证明了在不进行特征工程的情况下,改进后的模型优于基线模型。Peng N 等[11]提出将 LSTM 与 CRF 相结合应用于命名实体识别任务中,并通过实验证明了该方法的有效性。X. Yang 等[12]利用 BiLSTM 结合 CRF 来获取单词表示,将其用于生物医学领域的命名实体识别,并通过实验证明了该方法在生物医学领域的有效性。BiLSTM-CRF 模型在很多领域的命名实体识别任务中都取得了不错的效果,因此,许多学者在该模型的基础上进行改进。例如,Q. Zhong 等[13]在该模型的基础上加入了注意力机制,提高了命名实体识别任务的准确率。谢腾等[14]利用 BERT 模型生成基于上下文的词向量作为 BiLSTM-CRF 的输入进行中文实体识别并取得了较好的效果。赵平等[15]将 BERT+BiLSTM+CRF(简称 BBC)深度学习实体识别模型应用于旅游领域的文本,提高了旅游领域中实体识别的准确率。刘宇鹏等[16]针对中文命名实体识别提出了一种基于 BiLSTM-CNN-CRF 的方法,真正意义上的端到端的结构,自动获取基于字符级别和词语级别的表示,并在人民日报和医疗文本数据上进行了验证。此外,还有一些学者在神经网络模型基础上引入部首嵌入[17]、顺序遗忘编码[18]或者是笔画 ELMo 和多任务学习[19]等,实体识别效果均略有提升。

随着预训练语言模型的发展,越来越多的研究者将其用于命名实体识别的工作中,目前 BERT 模型[20]在各类自然语言处理任务中相较与其他预训练语言模型效果相对较好,而且应用较为广泛。M. Zhang 等[21]在 BiLSTM-CRF 模型中加入了 BERT 模型用于中文临床文本中,取得了良好的效果。Fábio Akhtyamova L[22]将 BERT 应用到西班牙生物医学领域中的命名实体识别任务,并且取得了不错的效果。王子牛等[23]针对传统机器学习算法对中文实体识别准确率低等问题,提出了将 BERT 模型和神经网络方法结合进行命名实体识别,并通过实验证明了该方法提升了实体识别的准确率、召回率和 F1 值。李妮等[24]利用 BERT 模型获取句子中丰富的句法和语法信息,并针对其训练参数过多,训练时间过长的问题,提出了一种基于 BERT-IDCNN-DRF 的中文命名实体识别的方法,并在 MSRA 语料上证明了该方法优于 Lattice-LSTM 模型,且训练时间大幅度缩短。

综上所述,命名实体识别的现有研究中缺乏充分利用各类神经网络及预训练语言模型的优势来进行实体识别任务。

文中的组织结构:第 2 节介绍了基于 BERT 模型的混合神经网络实体识别方法的模型架构并对各层原理或者结构进行说明解释;第 3 节在两个数据集上进行实验,比较不同方法的准确率、召回率和 F1 值,证明文中方法在命名实体识别任务中的有效性和通用性;第 4 节对全文进行总结并提出下一步工作方向。

2 基于 BERT 模型的混合神经网络实体识别方法

2.1 模型架构

文中提出了基于 BERT 模型的混合神经网络实体识别方法，其模型架构为 BERT + CNN + BiLSTM + Attention+CRF，如图 1 所示。

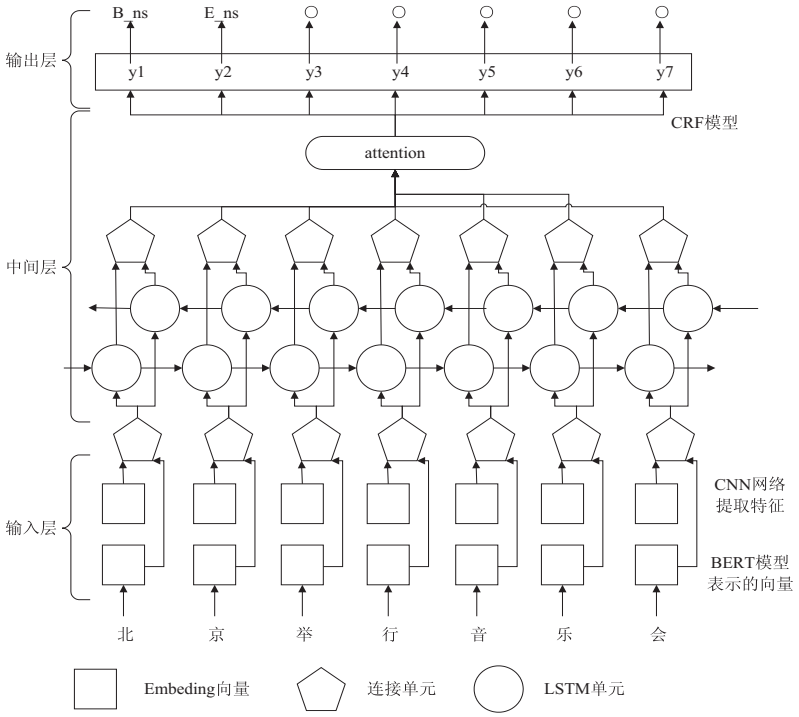

图 1 模型架构

首先是输入层，由 BERT 模型和 CNN 神经网络模型构成，BERT 模型训练基于字级别的字向量表示，CNN 神经网络模型提取文本语义特征，将两者结合作为下一层的输入向量。然后是由带有注意力机制的 BiLSTM 模型组成的中间层。最后是输出层，使用的是 CRF 模型来解码序列标签，从而得到全局最优标注序列。

2.2 基于 BERT 模型的向量表示

基于 BERT 模型的向量表示能够表达句子丰富的句法和语法信息，在自然语言处理领域中有着十分广泛的应用。BERT 模型是近几年来刚刚被提出与应用的，是预训练语言模型中表现较为突出的一个。BERT 模型是综合 GPT 和 ELOM 两个模型各自的优势构造出来的，采用了双向 Transformer 进行编码，充分利用字两侧的文本信息，能够动态生成字级别和词级别的语义向量，具有很强的语义表征优势。BERT 模型的本质是通过在海量的语料基础上运行自监督学习方法为单词学习一个好的特征表示，可以根据任务微调或者固定之后作为特征提取器。此外，BERT 的源码和模型已经开源。BERT 模型的网络结构如图 2 所示。由已有研究可知，BERT 模型在命名实体识别任务中具有良好的表现。文中方法利用 BERT 预训练语言模型将文本训练为句子向量作为输入层的一部分。

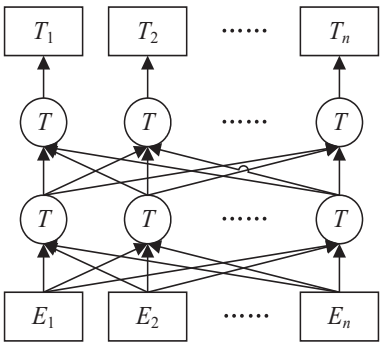

图 2 BERT 模型的网络结构

2.3 基于 CNN 网络的特征提取

CNN 网络的主要特点是它强大的卷积层能够获取足够丰富的特征。经典的 CNN 最开始主要应用于图像分类中，并且在图像分类领域取得了较好的成果。如今，经过学者们的不断研究与探索，慢慢地将 CNN 应用于自然语言处理中，例如命名实体识别、文本分类和自动摘要等工作。CNN 网络中的卷积层和池化层具有强大的特征提取和选择能力，能够防止过拟合，对特征进行降维。文中在卷积层中通过不同数量的过滤器和不同大小的卷积窗口进行卷积运算。池化层使用的是 Max Pooling 操作抽取出卷积层中最具有明显特征表征，从而得到基于 CNN 网络的文本特征向量，同样作为输入层的一部分。

2.4　基于注意力机制的 BiLSTM 网络

随着自然语言处理领域的不断进步和发展,LSTM 神经网络模型应用于自然语言处理领域有较好的表现。与传统的 RNN 网络结构相比,LSTM 增加了输入门、遗忘门和输出门三个门结构,能够更好地提取有用的信息。LSTM 单元结构如图 3 所示。

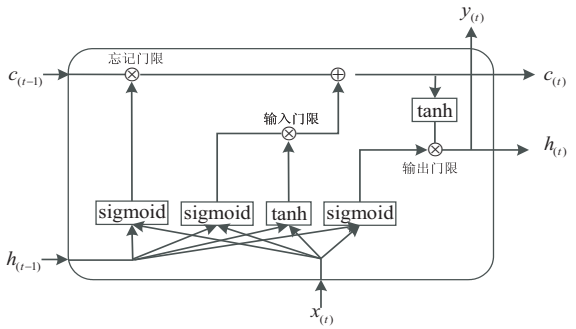

图 3　LSTM 单元结构

单向的 LSTM 只能获取一个方向的信息,但是在自然语言处理中充分利用上下文信息十分重要,双向 LSTM 网络,即 BiLSTM 应运而生。在命名实体识别任务中文本的上下文信息同样重要,因此,文中提出的基于 BERT 的混合神经网络实体识别方法中使用的便是 BiLSTM 网络模型结构。

注意力机制最开始被提出是应用于机器翻译问题中的,现在已经成为神经网络研究中的一个十分重要的研究领域。在神经网络结构中引入注意力机制能够自动学习权重用来捕捉编码器隐藏状态和解码器隐藏状态的相关性,从而提高神经网络模型的效果。注意力机制被广泛应用于各种不同类型的深度学习任务中,如自然语言处理、图像识别以及语音识别等任务。当然,在自然语言处理的子任务命名实体识别中,注意力机制的引入也同样会起到一定的效果。

文中实体识别方法的中间层使用的就是基于注意力机制的 BiLSTM 网络,将上述基于 BERT 模型的字符级向量和基于 CNN 网络提取的特征连接作为基于注意力机制的 BiLSTM 网络的输入向量。

2.5　基于 CRF 的输出层

基于 CRF 的输出层可以在最终的预测标签中添加一些约束,弥补 BiLSTM 无法处理相邻标签之间依赖关系的缺点,以确保最终的预测标签是有效的。这些约束可以由输出层的 CRF 在训练过程中从训练数据集自动学习。给定观察序列 X 时,某个特定标记序列 Y 的概率可定义为:

$$P(y\mid x)=\frac{1}{Z}\exp\Big(\sum_{j}\sum_{i=1}^{n-1}\lambda_j t_j(y_{i+1},y_i,X,i)+\sum_{k}\sum_{i=1}^{n}\mu_k S_k(y_i,X,i)\Big)$$

其中,$t_j(y_{i+1},y_i,X,i)$ 是定义在观测序列的两个相邻标记位置上的转移特征函数,刻画相邻标记变量之间的相关关系以及观测序列对它们的影响;$S_k(y_i,X,i)$ 是定义在观测序列的标记位置 i 上的状态特征函数,刻画观测序列对标记变量的影响,λ_j 和 μ_k 为参数,Z 为规范化因子。

3　实验与结果分析

3.1　数据集和标注方法

文中实验数据使用的是 1998 年《人民日报》语料数据集和 MSRA 语料数据集两个公开数据集,《人民日报》语料数据集中共有 19 484 个句子、52 735 个实体。MSRA 语料数据集中共有 28 100 个句子、80 884 个实体。对两个数据集中的人名(PER)、地名(LOC)和机构名(ORG)实体进行识别,其中训练集与测试集之比为 8∶2。两个数据集信息如表 1 所示。

表 1　数据集信息

数据集	句子数/个	实体数/个
《人民日报》	19 484	52 735
MSRA	28 100	80 884

常见的序列标注方法有很多种,例如 Markup 标注法、BIO 标注法和 BIEO 标注法等。文中使用的标注方法是 BIEO 标注法,其标注字母代表含义如表 2 所示。

表 2　BIEO 标注法含义

标注字母	代表含义
B	实体的开始
I	实体的中间字符
E	实体中的最后一个字符
O	一个非实体字符

3.2　参数设置和评价指标

文中使用的 BERT 预训练语言模型采用的是 BERT-Base,相关参数设置如表 3 所示。

表 3　相关参数设置

参数名	具体值
BERT-Base 层数	12 层
BERT-Base 隐层维度	768 维
BERT-Base 模式	12 头模式
字向量长度	128
CNN 卷积核数	50
CNN 窗口大小	5
BiLSTM 前向神经元个数	128
BiLSTM 后向神经元个数	128
learning_rate	0.001
batch_size	8
丢失率	0.5

采用准确率(Precision, P)、召回率(Recall, R)和 F1 值(F1-score)三个指标来衡量实体识别模型的效果。三个评价指标的计算公式如下:

$$P = \frac{RER}{AER}$$

$$R = \frac{RER}{AE}$$

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

其中,RER 表示正确识别出的实体数,AER 表示实际识别出的实体数,AE 表示实际实体总数。

3.3 实验结果分析

为了验证文中提出的基于 BERT 模型的混合神经网络实体识别方法的有效性,将该方法与 BiLSTM-CRF、LSTM-CNNs 和 CNN-BiLSTM-CRF 三种命名实体识别的方法在《人民日报》和 MSRA 两个数据集上进行了对比实验,比较四种命名实体识别方法的准确率、召回率和 F1 值。

首先为了确定合适的迭代次数,采用四种方法分别在两个数据集上进行了 50 次迭代,四种方法的 F1 值与迭代次数的关系如图 4 和图 5 所示。四种方法在两个数据集上均在 20 次迭代前后出现最高的 F1 值。此外,可以看出文中提出的基于 BERT 模型的混合神经网络实体识别方法在这两个数据集上的 F1 值均高于其他三种方法,具有良好的表现。

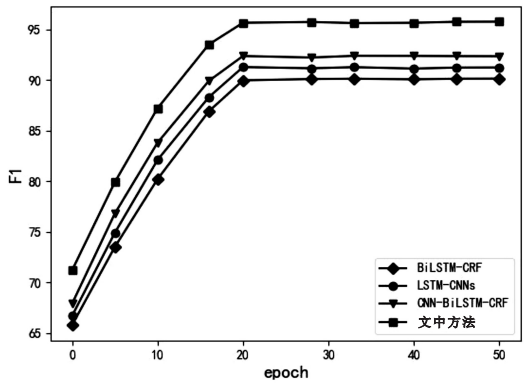

图 4 《人民日报》语料数据集上 F1 与迭代次数关系

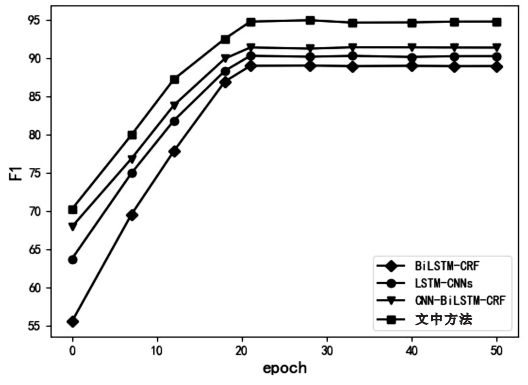

图 5 MSRA 语料数据集上 F1 与迭代次数关系

通过上述实验对比,将迭代次数设为 23 次,将四种方法在两个数据集上进行实验,对比其准确率、召回率和 F1 值,实验结果如表 4 和表 5 所示。

表 4 《人民日报》语料数据集

实体识别方法	准确率 P/%	召回率 R/%	F1/%
BiLSTM-CRF	90.86	89.63	90.24
LSTM-CNNs	91.67	90.56	91.11
CNN-BiLSTM-CRF	92.72	91.36	92.03
文中方法	96.13	95.32	95.72

表 5 MSRA 语料数据集

实体识别方法	准确率 P/%	召回率 R/%	F1/%
BiLSTM-CRF	90.46	88.52	89.48
LSTM-CNNs	91.52	90.41	90.96
CNN-BiLSTM-CRF	91.79	91.36	91.57
文中方法	95.23	94.37	94.80

从表 4 和表 5 可以看出,文中提出的基于 BERT 模型的混合神经网络实体识别方法在准确率、召回率和 F1 值上均优于其他三种方法。在《人民日报》语料数据集上,文中方法的 F1 值比 BiLSTM-CRF 方法高出大约 5.5%,比 LSTM-CNNs 方法高出大约 4.6%,比 CNN-BiLSTM-CRF 高出大约 3.7%。在 MSRA 语料数据集上,文中方法的 F1 值比 BiLSTM-CRF 方法高出大约 5.4%,比 LSTM-CNNs 方法高出大约 4%,比 CNN-BiLSTM-CRF 高出大约 3.2%。由此可见,文中提出的基于 BERT 模型的混合神经网络实体识别方法具有一定的有效性和通用性。

4 结束语

为了更好地解决命名实体识别方法中语义分析不足及准确率较低的问题,结合预训练语言模型和各类神经网络的优势及特点,提出一种基于 BERT 模型的混合神经网络实体识别方法。充分运用了 BERT 模型、CNN 网络、注意力机制以及 BiLSTM-CRF 模型的优势,更加充分地提取文本的语义信息,丰富其文本特征,进行命名实体识别任务。最后分别在两个数据集上证明了提出方法的有效性和通用性。后续将进一步针对如何获取更多文本特征方面进行研究。

参考文献:

[1] 王 鑫,邹 磊,王朝坤,等. 知识图谱数据管理研究综述[J]. 软件学报,2019,30(7):2139-2174.

[2] 孙 镇,王惠临. 命名实体识别研究进展综述[J]. 现代图书情报技术,2010(6):42-47.

[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.

[4] 熊回香,杨梦婷,李玉媛. 基于深度学习的信息组织与检索

研究综述[J]. 情报科学,2020,38(3):3-9.

[5] LAFFERTY J,MCCALLUM A,PEREIRA F C N. Conditional random fields:probabilistic models for segmenting and labeling sequence data[C]//18th international conference on machine learning. San Francisco,CA,USA:[s. n.],2001:282-289.

[6] BERGER A L,PIETRA V J D,PIETRA S A D. A maximum entropy approach to natural language processing[J]. Computational Linguistics,1996,20(1):39-71.

[7] MCCALLUM A,FREITAG D,PEREIRA F C N. Maximum entropy Markov models for information extraction and segmentation[C]//Proceedings of the seventeenth international conference on machine learning. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. ,2000:591-598.

[8] COLLOBERT R,WESTON J,BOTTOU L,et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research,2011,12(1):2493-2537.

[9] GUL Khan Safi Qamas,尹继泽,潘丽敏,等. 基于深度神经网络的命名实体识别方法研究[J]. 信息网络安全,2017(10):29-35.

[10] BLÜCÜ N,AKGL D,TU S. Bidirectional LSTM-CNNs with extended features for named entity recognition[C]//Scientific meeting on electrical-electronics and biomedical engineering and computer science. Ankara,Turkey:[s. n.],2019.

[11] PENG Nanyun,DREDZE M. Improving named entity recognition for Chinese social media with word segmentation representation learning[C]//54th annual meeting of the association for computational linguistics. Commonwealth of Pennsylvania: Association for Computational Linguistics, 2016: 149 - 155.

[12] YANG X,GAO Z,LI Y,et al. Bidirectional LSTM-CRF for biomedical named entity recognition[C]//2018 14th international conference on natural computation, fuzzy systems and knowledge discovery (ICNC-FSKD). Huangshan, China:[s. n.],2018.

[13] ZHONG Q,TANG Y. An attention-based BILSTM-CRF for Chinese named entity recognition[C]//2020 IEEE 5th international conference on cloud computing and big data analytics (ICCCBDA). Chengdu,China:IEEE,2020.

[14] 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.

[15] 赵　平,孙连英,涂　帅,等. 改进的知识迁移景点实体识别算法研究及应用[J]. 数据分析与知识发现,2020,4(5):118-125.

[16] 刘宇鹏,栗冬冬. 基于 BLSTM-CNN-CRF 的中文命名实体识别方法[J]. 哈尔滨理工大学学报,2020,25(1):115-120.

[17] 郭旭超,唐　詹,刁　磊,等. 基于部首嵌入和注意力机制的病虫害命名实体识别[J]. 农业机械学报,2020,51(S2):335-343.

[18] 杨贺羽,杜洪波,朱立军. 基于顺序遗忘编码和 Bi-LSTM 的命名实体识别算法[J]. 计算机应用与软件,2020,37(2):213-217.

[19] 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.

[20] DEVLIN J, CHANG M, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding [C]//The 18th annual conference of the north American chapter of the association for computational linguistics: human language technologies. New Orleans, Louisiana: ACL, 2019: 4171-4186.

[21] ZHANG M, WANG J, ZHANG X. Using a pre-trained language model for medical named entity extraction in Chinese clinic text[C]//2020 IEEE 10th international conference on electronics information and emergency communication (ICEIEC). Beijing,China:IEEE,2020.

[22] AKHTYAMOVA L. Named entity recognition in spanish biomedical literature: short review and bert model [C]//2020 26th conference of open innovations association (FRUCT). Yaroslavl,Russia:[s. n.],2020:1-7.

[23] 王子牛,姜　猛,高建瓴,等. 基于 BERT 的中文命名实体识别方法[J]. 计算机科学,2019,46(S2):138-142.

[24] 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.