doi:10.6041/j.issn.1000-1298.2020.S2.039

基于部首嵌入和注意力机制的病虫害命名实体识别

郭旭超　唐　詹　刁　磊　周　晗　李　林
(中国农业大学信息与电气工程学院, 北京 100083)

摘要: 为了解决农业病虫害命名实体识别过程中存在的内在语义信息缺失、局部上下文特征易被忽略和捕获长距离依赖能力不足等问题,以农业病虫害文本为研究对象,提出一种基于部首嵌入和注意力机制的农业病虫害命名实体识别模型(Chinese agricultural diseases and pests named entity recognition with joint radical-embedding and self-attention, RS-ADP)。首先,该模型将部首嵌入集成到字符嵌入中作为输入,用以丰富语义信息。其中,针对部首嵌入设计了 3 种特征提取策略,即卷积神经网络(Convolutional neural network, CNN)、双向长短时记忆网络(Bidirectional long short-term memory network, BiLSTM)和 CNN-BiLSTM;其次,采用多层不同窗口尺寸的 CNNs 层提取不同尺度的局部上下文信息;然后,在 BiLSTM 提取全局序列特征的基础上,采用自注意力机制进一步增强模型提取更长距离依赖的能力;最后,采用条件随机场(Conditional random field, CRF)联合识别实体边界和划分实体类别。在包含 11 个类别和 24 715 条标注样本的农业病虫害自制语料上进行了实验。结果表明,本文模型 RS-ADP 在该数据集上精确率、召回率和 F_1 值分别为 94.16%、94.47% 和 94.32%;在具体实体类别上,RS-ADP 在作物、病害、虫害等易识别实体上 F_1 值高达 95.81%、97.76% 和 97.23%。同时,RS-ADP 在草害、病原等难以识别实体上 F_1 值仍保持 86% 以上。实验结果表明,本文所提模型能够有效识别农业病虫害命名实体,其识别精度优于其他模型,且具有一定的泛化性。

关键词: 农业病虫害; 命名实体识别; 部首嵌入; 自注意力机制; 双向长短时记忆网络; 卷积神经网络
中图分类号: TP391.1　　**文献标识码**: A　　**文章编号**: 1000-1298(2020)S2-0335-09

Recognition of Chinese Agricultural Diseases and Pests Named Entity with Joint Radical-embedding and Self-attention Mechanism

GUO Xuchao　TANG Zhan　DIAO Lei　ZHOU Han　LI Lin
(*College of Information and Electrical Engineering*, *China Agricultural University*, *Beijing* 100083, *China*)

Abstract: Chinese named entity recognition in agricultural diseases and pests domain (CNER-ADP) plays an important role in agricultural natural language processing such as relation extraction, agricultural knowledge graph construction, and agricultural knowledge question and answering, but it still presents some problems, i.e., the neglect of inherent semantic information and local contextual features and the insufficiency of capturing long-distance dependencies, which will lead to low accuracy and robustness. To solve the above problems and tackle the CNER-ADP task, a novel Chinese named entity recognition method for agricultural diseases and pests via jointly using radical-embedding and self-attention (RS-ADP) was proposed. Firstly, the model integrated radical embedding and character embedding as input to enrich semantic information. Among them, three different strategies, including CNN and BiLSTM were both designed to capture the radical-level embedding. Secondly, a CNNs layer with different kernel sizes was considered capturing multi-scale local contextual features. Thirdly, based on the BiLSTM layer, self-attention mechanism was used to further enhance the ability of the model to extract longer-distance dependencies. Finally, the conditional random field (CRF) was utilized to identify entity boundaries and category. The experiments were carried out on the corpus of agricultural diseases and pests, named AgCNER, which contained 11 categories and 24 715 samples. At macro-level, the RS-ADP model

收稿日期: 2020-08-10　修回日期: 2020-09-20
基金项目: 国家重点研发计划项目(2016YFD0300710)
作者简介: 郭旭超(1992—),男,博士生,主要从事自然语言处理与知识图谱研究,E-mail: gxc@cau.edu.cn
通信作者: 李林(1963—),女,教授,博士生导师,主要从事大数据管理与挖掘研究,E-mail: lilincau@126.com

achieved optimal precision, recall, and F_1 values of 94.16%, 94.47%, and 94.32%, respectively. In terms of specific categories, it achieved F_1 values as high as 95.81%, 97.76%, and 97.23% on easily identifiable entities such as crop, disease, and pest. Meanwhile, this model still maintained over 86% of F_1 value on some other difficultly recognized entities such as weed and pathogeny. The experimental results showed that the proposed model could effectively recognize the named entities of agricultural pests and diseases without feature engineering. Moreover, it had certain generalization and outperformed other models.

Key words: agricultural diseases and pests; named entity recognition; radical embedding; self-attention mechanism; bidirectional long short-term memory network; convolutional neural network

0 引言

面向农业病虫害领域的中文命名实体识别任务旨在从非结构化文本中识别出与农业病虫害相关的命名实体,它是进行农业病虫害关系抽取、知识图谱构建[1]和知识问答等下游任务的基础和关键。

目前,面向农业病虫害的命名实体识别方法可分为基于规则、基于字典匹配和传统机器学习等方法,但基于规则和字典匹配的方法存在通用性差、人工操作耗时长等问题。传统机器学习方法旨在将命名实体识别问题转换为序列标注问题,已在命名实体识别方面取得一定成果。其中,条件随机场因其能够联合预测实体类别而具有较高的准确率,进而成为命名实体识别的常用方法。如文献[2-5]均采用 CRF 识别农业病虫害领域的命名实体,如作物、农药和病虫害等。但传统机器学习方法严重依赖人工特征,需要很高的人力和时间成本[6]。

近年来,基于深度学习的命名实体识别已广泛应用于医学病历[7-8]、生命科学[9]和社交媒体[10]等领域。其中,基于双向长短期记忆神经网络使用最为广泛。如文献[11]以词为基本单元,首次联合采用 BiLSTM 和 CRF 进行序列标注任务,在 CoNLL 语料上 F_1 值为 90.10%,证明了该模型的有效性。文献[12-15]分别采用 CNN 和 BiLSTM 提取词内字符信息并结合词嵌入信息进行命名实体识别,解决了 OOV(Out-of-vocabulary)问题。文献[16]同时采用 CNN 和 BiLSTM 提取了词内高层次字符级特征表示信息,并结合词级嵌入作为 BiLSTM-Attention-CRF 模型输入,在 NCBI-Disease 和 JNLPBA 数据集上分别取得最高 F_1 值为 86.93% 和 75.31%。然而,由于中文不存在空格等明显分隔符,基于中文分词的中文词嵌入容易产生错误传播问题,因此采用词嵌入作为输入并不是最优选择。为了充分利用汉字内部丰富的语义信息,文献[7]采用 CNN 提取部首级嵌入信息,并结合字符级嵌入用于临床医学领域的命名实体识别,分别在 CCKS 2017 和 TP_CNER 上取得 F_1 值为 93.00% 和 86.34%。此外,文献[17]提出了面向命名实体识别的迭代扩张的卷积模型(Iterated dilated convolutional neural networks, IDCNN),实验结果证明了该模型能够充分利用 GPU 并行运算优势,加快训练速度。基于上述模型,文献[18]提出了 BERT-IDCNN-CRF 模型,该模型保持 BERT 参数不变[19],仅提供上下文特征,既保持了多样性又减少了训练参数[20]。文献[8]结合标准 CNN 和 IDCNN 提出了残差扩张卷积模型(Residual dilated convolutional neural network with conditional random field, RD-CNN-CRF),在 CCKS2017 数据集上的实验结果表明了 RD-CNN-CRF 模型在计算性能和训练时间方面都与基于 BiLSTM 模型相当。

农业病虫害领域的命名实体识别仍面临如下问题:①除文献[1]外,鲜有基于深度学习的农业病虫害命名实体识别模型,且农业病虫害领域语料和标注数据十分有限,这在一定程度上增加了识别农业病虫害实体的成本和难度。② 中文农业病虫害文本蕴含丰富的形态信息,如“稻瘟病”、“赤霉病”和“疫霉病”中的“病”字偏旁为“疒”,与病害密切相关,能够起到分隔符的作用,基于 BiLSTM-CRF 模型仅能捕获句子的全局上下文信息,而无法捕获这些形态信息。③全局上下文信息和局部上下文特征在农业病虫害文本识别中同等重要,但现有 BiLSTM 模型往往忽略了局部上下文信息。④虽然 BiLSTM 理论上可以捕获句子的长距离依赖信息,但由于梯度消失和文本距离过长等实际问题,导致捕获更长距离依赖信息的能力无法充分发挥。

因此,为了能够更好地完成农业病虫害领域的命名实体识别任务,本文提出一种基于部首嵌入和自注意力机制的神经网络模型 RS-ADP。该模型以 BiLSTM-CRF 为基本框架,设计 3 种部首特征提取策略,并结合字符嵌入以丰富语义信息;采用多个不同窗口尺寸的卷积神经网络提取不同尺度的局部上下文特征;采用自注意力机制弥补 BiLSTM 的不足,以提高模型捕获更长距离上下文依赖的能力。

1　材料与方法

1.1　语料库采集与标注

1.1.1　文本收集

为了保证语料品质，以具体病虫害名称为关键字，从中国知网（https://www.cnki.net/）、万方数据（http://www.wanfangdata.com.cn/）爬取了与农业病虫害相关的文本，经去重预处理后，最终获得了249 094 条文本作为预训练模型的语料。

1.1.2　类别划分

在参照中外早期研究成果[1]的基础上，将农业病虫害的中文命名实体类型进一步细分，例如将“病虫害”细分为“病害”和“虫害”，并添加了病原、部位、作物品种等多种新的实体类别；此外，为了确保完整性，本文还定义了“其他”实体类，用于描述一些概念性和潜在的不确定实体，以备后续扩充。最终确定了虫害、病害、作物、药剂、草害、作物品种、肥料、病原、周期、部位和其他等 11 种实体类别。

1.1.3　标注实体

选取部分语料，在农业病虫害领域专家指导下，采用自开发标注工具 ChineseNERAnno（https://github.com/guojson/ChineseNERAnno），历时 3 个月最终完成了包含 200 596 个实体，24 715 条样本的中文农业病虫害命名实体语料（Agricultural diseases and pests corpus for Chinese named entity recognition，AgCNER）。

1.1.4　特征分析

如表 1 所示，AgCNER 数据集包含与农业病虫害相关的专业领域命名实体，与通用语料相比，在实体类型和专业性等方面都有很大的不同。除了无明确边界特征外[1]，该数据集还有如下特点：

表 1　AgCNER 数据集样例

Tab.1　Some samples of AgCNER

序号	样例
1	应用 40% 毒死蜱乳油 500 mL/667 m^2，40% 甲基异柳磷乳油 150、200 mL/667 m^2 和 40% 氯·辛乳油250 mL/667 m^2 在插种期 1 次施药防治花生田蛴螬的田间效果分别为 50.4%、42.3%、52.3% 和 48.6%；在播种期和开花扎针期 2 次施药效果更好，校正防效提高 25.1% ~42.3%
2	小麦条纹叶枯病是由灰飞虱传播的病毒病，近年来逐年加重，通过对小麦条纹叶枯病发病规律和发生特点进行调查研究，掌握其综防技术

（1）实体类别多。该数据集除了包含药剂、病害、虫害和草害等 4 类常用实体外，还包含作物品种、肥料等其他 7 种实体类别。特别对于作物品种，常由字母、汉字、数字等多种符号构成，如“郑单958”、“山农 29 号”，容易导致识别不一致。

（2）专业性强。在 AgCNER 文本中存在大量与农业病虫害相关的专业词汇，如“多菌灵”、“蚜霉净”和“掖单 2 号”等。因此，将通用领域模型迁移到该特定领域进行命名实体识别的难度很大。此外，由于现有分词工具对专业名词不敏感，更容易导致分词错误，进而降低 CNER 的识别精度。

（3）存在嵌套实体。本数据集中存在一定比例嵌套实体，即一个实体可由多个子实体构成，如“烯肟菌胺·苯醚甲环唑·噻虫嗪悬浮种衣剂”，又可拆分为“烯肟菌胺”、“苯醚甲环唑”、“噻虫嗪”和“悬浮种衣剂”等子实体。容易干扰识别，造成模型误判。

（4）句子序列很长。经统计分析，AgCNER 中样本平均长度为 180.08 个字符，远大于 Resume 和 MSRA 等语料的平均长度，分别为 32.15、46.17 个字符，这对模型提取长距离依赖的能力是一种挑战。

1.2　RS－ADP 模型

针对农业病虫害文本的特点和存在的问题，本文提出了一种适用于农业病虫害命名实体识别的 RS－ADP 模型，该模型架构如图 1 所示。由图 1 可知，本文以 BiLSTM－CRF 为基本框架，共包含 5 层：嵌入层、CNN 层、BiLSTM 层、自注意力层和 CRF 层，重点阐述嵌入层、CNN 层、自注意力层的实现细节，BiLSTM 层和 CRF 层参照文献[11]。

1.2.1　嵌入层

嵌入层主要用于将离散的文本序列转换为低维稠密的分布式嵌入表示。研究表明，这种嵌入表示能够为模型提供更好的初始化，使得模型具有较强的泛化能力和加速模型收敛性[20]。如图 2 所示，除字符嵌入外，该层还包含部首嵌入。

1.2.1.1　字符嵌入

由于中文分词存在错误传播问题，且已有研究成果表明了将字符嵌入作为输入要优于中文词嵌入，因此本文将字符级嵌入作为基本单元。由于本文收集的语料规模还远未达到 BERT 预训练所需的百万量级[21]，所以暂未采用 BERT 进行预训练，而 Word2Vec[22] 对语料规模要求稍低，且 Skip－Gram 能够有效处理低频字符，因此本文采用 Word2Vec 中的 Skip－Gram 训练字符嵌入，并在训练过程中进行微调。形式上，给定长度为 n 的输入序列 $\boldsymbol{X}_c=(w_1,w_2,\cdots,w_n)$，则其对应的字符嵌入序列为 $\boldsymbol{E}_c=(\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_n)$，其中第 i 个字符的嵌入向量表示为 $\boldsymbol{e}_i\in\mathbf{R}^{d_c}$，$d_c$ 为输出维度。

1.2.1.2　部首嵌入

部首嵌入是本文研究的重点。与英文不同，汉字属于象形文字，多数仍然保留着它们的原意，因此

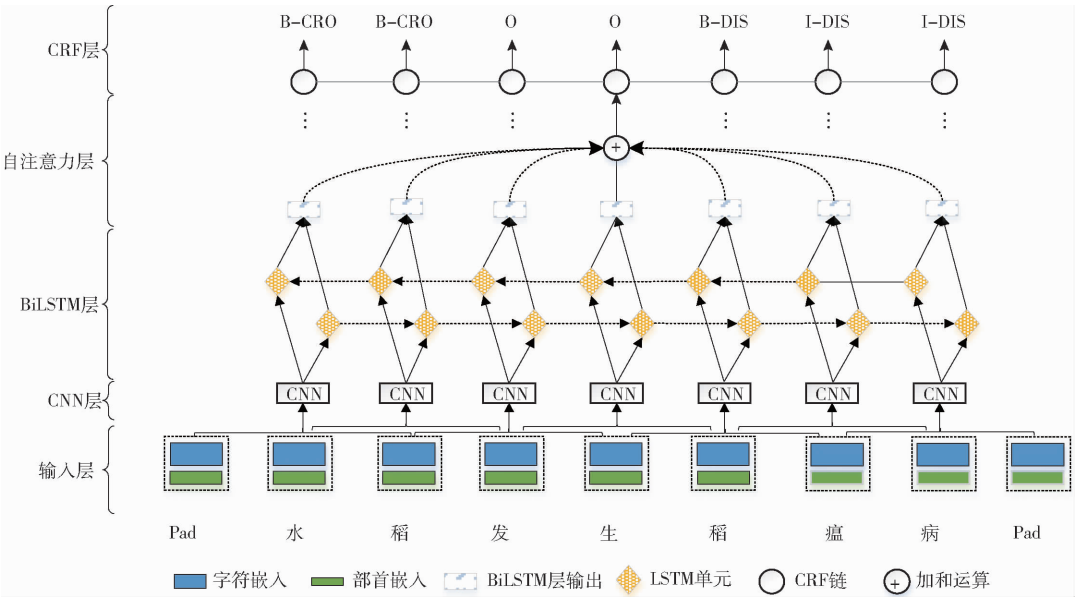

图 1 RS－ADP 模型架构

Fig. 1 Main framework of RS－ADP

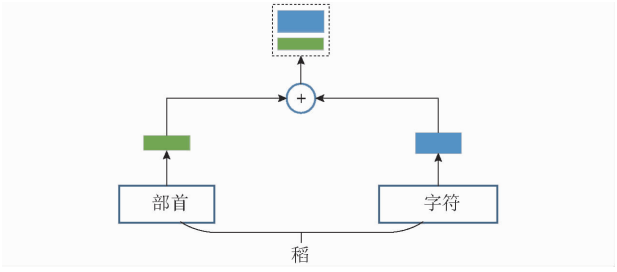

图 2 嵌入层的部首嵌入和字符嵌入特征

Fig. 2 Radical and character embedding in embedding layer

具有丰富的内在特征。这意味着相似的字可能包含相似的意义，而这种相似性主要体现在部首上。部首特征在临床医学命名实体识别中研究较多[7-8]，但在农业病虫害命名实体识别领域，其尚未得到充分利用。如图 3 所示，以“酯”和“病”为例，“酯”部首为“酉”，经常在化学药剂中出现，如“醚菊”、“噻霉酮”等；而“病”的部首“疒”代表疾病，多在病害实体中出现，如“稻瘟病”、“赤霉病”等。因此，部首在一定程度上能够衡量实体的相似性。此外，部首嵌入还能增强只出现在测试集中而不在训练集中的

图 3 部分汉字部首示例

Fig. 3 Radical samples for some Chinese words

字符语义信息，从而提高模型的泛化能力[23]。本文为了充分利用农业病虫害文本中潜在的内在特征，分别设计了基于 CNN、基于 BiLSTM 及其混合模型 CNN－BiLSTM 等 3 种部首提取架构，如图 4 所示，图中“疒”、“曰”和“皿”为“瘟”的偏旁部首。实验结果表明基于 CNN 的部首嵌入最有效。

(1)基于 CNN 的部首嵌入

研究表明，CNN 是一种善于捕获字符形态信息的有效方法[7]。因此，采用 CNN 提取字符内部隐含的部首信息，并在训练过程中进行微调。如图 4a 所示，该模型共包含 3 部分：部首嵌入层、卷积层和最大池化层。本文所用所有部首均从新华字典(http://tool.httpcn.com/Zi/)中获得。形式上，给定含有

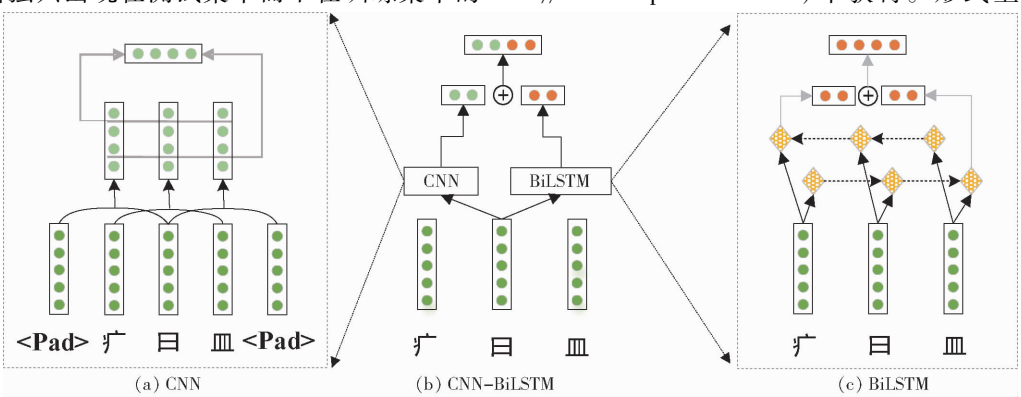

图 4 部首级嵌入表示提取架构

Fig. 4 Framework of radical embedding

m 个部首偏旁的部首序列 $\boldsymbol{X}_r=(r_1,r_2,\cdots,r_m)$，其对应的输出维度为 d_r。部首级嵌入 E_r 计算公式为

$$\boldsymbol{E}_r=\max\_\text{pool}(\text{Conv}(\boldsymbol{X}_r)) \tag{1}$$

采用 ReLU[24] 作为激活函数。从范围为[-1，1]的均匀分布中随机初始化部首嵌入。

(2)基于 BiLSTM 部首嵌入

由于每个汉字的部首都有其特定的位置，因此部首的顺序特征对实体识别也具有一定辅助作用[25]。采用适合处理文本序列的 BiLSTM 捕获上述部首顺序特征，其作用过程如图 4c 所示，其中⊕表示级联。给定部首序列 $\boldsymbol{X}_r=(r_1,r_2,\cdots,r_m)$，经隐藏层维度为 d_l 的 BiLSTM 模型处理后得到的部首级嵌入为 $\boldsymbol{E}_l=[\overrightarrow{h};\overleftarrow{h}]$，式中 $\overrightarrow{h}$ 和 $\overleftarrow{h}$ 分别为前、后向嵌入表示，$\boldsymbol{E}_l\in\mathbf{R}^{2d_l}$。

(3)混合部首嵌入

CNN 模型善于获取部首的空间特征，而 BiLSTM 则能够提取序列特征。基于此，综合上述 2 种模型，设计了一种同时捕获部首序列的空间和时间特征的混合模型，如图 4b 所示，则其最终的输出为

$$\boldsymbol{E}_j=\boldsymbol{E}_r\oplus\boldsymbol{E}_l \tag{2}$$

则嵌入层整体输出为部首嵌入和字符嵌入的级联，即

$$\boldsymbol{E}=\boldsymbol{E}_c\oplus\boldsymbol{E}_j \tag{3}$$

1.2.2　CNN 层

局部上下文特征能够反映相邻字符间的上下文关系，有助于提高模型识别命名实体的能力。而这在农业病虫害命名实体识别中同样重要，如“药剂功夫”和“中国功夫”，“药剂”对判断“功夫”是一种药剂名称而不是一项体育运动的名字起到决定性作用，因此，很有必要捕捉文本序列的局部上下文信息。

卷积神经网络已经被广泛应用于计算机视觉领域[26]，它可以提取图像的局部特征[27]。基于此，本文将卷积神经网络应用在农业病虫害领域的命名实体识别上，以提取局部上下文信息。为了充分发掘相邻字符间的上下文关系，采用多层不同窗口尺寸的卷积神经网络提取不同尺度的局部上下文信息。本节与 1.2.1 节 CNN 不同点在于前者作用于整个样本序列，而后者仅作用于字内部的部首偏旁序列。给定输入序列 $\boldsymbol{E}=(\boldsymbol{e}_1,\boldsymbol{e}_2,\cdots,\boldsymbol{e}_n)$，第 i 个字符经窗口尺寸为 k 的卷积神经网络得到的输出维度为 o 的局部上下文信息为 $c_i^k\in\mathbf{R}^o$，$\boldsymbol{w}\in\mathbf{R}^{kd_e}$ 为训练权重，其中 d_e 为输出维度。则该层总输出为 $\boldsymbol{C}=(c_1,c_2,\cdots,c_n)$。

$$c_i^k=\text{ReLU}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{E}_{\lfloor i-\frac{k-1}{2}\rfloor:\lfloor i+\frac{k-1}{2}\rfloor}+\boldsymbol{b}) \tag{4}$$

式中　$\boldsymbol{b}$——偏置

1.2.3　自注意力机制层

BiLSTM 层作用于整个句子，用于提取文本序列的全局上下文特征。经该层训练得到的嵌入序列可表示为 $\boldsymbol{H}=(h_1,h_2,\cdots,h_n)$，式中 $\boldsymbol{H}\in\mathbf{R}^{n\times 2l}$，$l$ 为隐藏层数量。

本文需训练的 AgCNER 数据集样本长度稍长。而研究表明，虽然 BiLSTM 理论上能够捕获长距离依赖信息，但实际上由于梯度消失问题，其获取长距离依赖的能力会随着本文序列的增加而降低。自注意力机制(Self-attention)能够捕获字符之间的上下文关系，根据字符的重要性赋予不同的权值，即对起重要作用的特定字符赋予较大权值，而对其他无用字符赋予较小权重。此外，自注意力机制能够充分利用 GPU 的优势进行并行运算[28]。因此，为了解决 BiLSTM 存在的问题，采用自注意力机制来进一步增强模型捕获更长距离上下文依赖的能力。其形式化定义为

$$\boldsymbol{A}=\text{softmax}\left(\frac{\boldsymbol{H}\boldsymbol{W}^Q(\boldsymbol{W}^K)^{\mathrm{T}}\boldsymbol{H}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{H}\boldsymbol{W}^V=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d}}\right)\boldsymbol{V} \tag{5}$$

具体地，对于上层输入 $\boldsymbol{H}$，首先，通过可训练权重 $\boldsymbol{W}^Q$、$\boldsymbol{W}^K$、$\boldsymbol{W}^V\in\mathbf{R}^{2l\times d}$ 将 $\boldsymbol{H}$ 分别映射为查询向量 $\boldsymbol{Q}$、关键字向量 $\boldsymbol{K}$ 和值向量 $\boldsymbol{V}$；然后采用缩放点积计算各字符向量 $\boldsymbol{q}_i\in\boldsymbol{Q}$ 对 $\boldsymbol{K}$ 中每个关键字向量 $\boldsymbol{k}_i$ 的相似性，并采用 softmax 进行归一化处理，得到权重系数；最后，根据权重系数对每个 $v_j\subseteq\boldsymbol{V}$ 进行加权求和，得到最终的输出 $\boldsymbol{A}=(a_1,a_2,\cdots,a_n)$，$\boldsymbol{A}\in\mathbf{R}^{n\times d}$。

在解码层，由于相邻字符间具有很强的关联性，如 I-DIS 通常在 B-DIS 或 I-DIS 后出现，但其不会紧随 B-CRO 或 I-CRO 出现。因此，本文选用条件随机场 CRF 作为联合解码器，采用 Veterbi 算法搜索标签序列。CRF 详情参照文献[29]。

2　实验结果及分析

2.1　实验设计

实验选用农业病虫害数据集 AgCNER 作为训练样本，训练集与验证集划分比例为 8∶2，统计信息如表 2 所示。采用 Tensorflow 1.13.1 框架，运行环境为 GTX 1080Ti GPU 11GB。

2.1.1　参数设置

以字符作为基本输入单元，采用 BIO 标注模式。为了避免语义缺失，本文将样本最大长度视为训练最大长度，采用 dropout 防止过拟合。其他参数详见表 3。

表 2 AgCNER 统计信息

Tab. 2 Statistic information for AgCNER

类别	训练集	占比/%	验证集	占比/%
虫害	37 186	18.54	8 840	18.581
其他	38 557	19.22	8 903	18.714
作物	31 688	15.80	7 613	16.002
病害	22 521	11.23	5 354	11.254
药剂	20 507	10.22	4 780	10.047
周期	20 115	10.03	4 907	10.314
部位	14 557	7.26	3 445	7.241
作物品种	9 471	4.72	2 336	4.910
病原	3 739	1.86	876	1.841
草害	1 050	0.52	223	0.469
肥料	1 205	0.60	298	0.626
总计	200 596	100	47 575	100

表 3 实验超参数设置

Tab. 3 Settings of hyper-parameters

超参数	数值	超参数	数值
字符嵌入维度	300	BiLSTM 层数	1
部首嵌入 CNN 维度	50	注意力输出维度	600
部首嵌入 CNN 窗口尺寸	3	Dropout	0.5
部首嵌入 BiLSTM 维度	50	Batch_size	64
CNN 层窗口尺寸	1、3、5	最大迭代次数	100
过滤器数量	300	学习率	0.001
步长	1	优化算法	Adam
BiLSTM 层维度	300		

2.1.2 评估指标

采用精确率(Precision, P)、召回率(Recall, R)和 F_1 值作为评价指标,当且仅当实体的边界和类别均被正确识别时该实体才被视为识别正确。各评价指标计算式为

$$P = \frac{T_P}{T_P + F_P} \times 100\% \tag{6}$$

$$R = \frac{T_P}{T_P + F_N} \times 100\% \tag{7}$$

$$F_1 = \frac{2PR}{P + R} \tag{8}$$

式中 T_P——预测结果和实际结果均为正例的样例个数

F_P——实际为反例,预测结果为正例的样例个数

F_N——实际为正例,预测结果为负例的样例个数

2.2 实验结果及分析

2.2.1 预训练字符嵌入

分别采用随机生成和 Word2Vec 字符嵌入作为对比实验,以说明预训练字符嵌入有助于 RS-ADP 模型识别农业病虫害领域命名实体。由图 5 可知,基于预训练字符嵌入的 RS-ADP 模型在 AgCNER 数据集上的 P、R 和 F_1 值明显优于基于随机字符嵌入取得的相应值。图 6 给出了 RS-ADP 模型在 AgCNER 上的 F_1 值随迭代次数的变化曲线。由图可知,基于 Word2Vec 的字符嵌入有助于加速模型收敛,提高模型的识别准确度。因此,上述实验结果说明了基于 Word2Vec 的预训练字符嵌入模型有助于提高模型的能力,加快模型收敛。

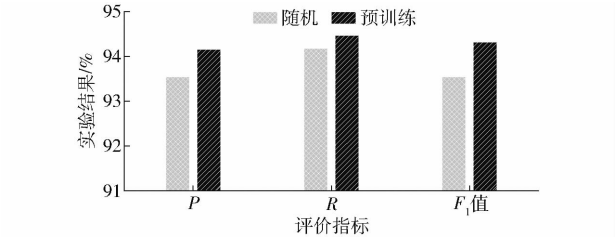

图 5 采用 2 种嵌入的 RS-ADP 在 AgCNER 上的识别结果

Fig. 5 Results of RS-ADP with random and pre-trained embedding

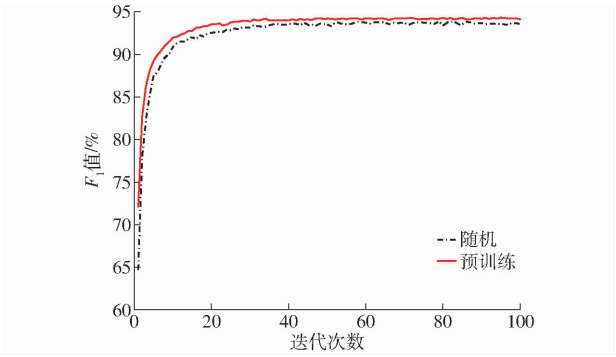

图 6 采用两种嵌入的 RS-ADP 的 F_1 值随迭代次数的变化曲线

Fig. 6 Trends of F_1 value for RS-ADP over epoch

2.2.2 部首嵌入

由表 4 实验组 1~4 结果可知,融合部首嵌入的 BiLSTM-CRF 模型针对农业病虫害命名实体的识别结果明显优于只有字符嵌入作为输入的基准模型,其 F_1 值分别提高了 0.35、0.25、0.29 个百分点。这说明了 CNN 与 BiLSTM 均能够有效提取部首特征,且有助于模型识别农业病虫害相关命名实体。其中,CNN 能够最大程度上提升模型识别性

表 4 不同架构的 RS-ADP 在 AgCNER 上的识别结果

Tab. 4 Results of RS-ADP with different layers on AgCNER %

实验组	模型	P	R	F_1 值
1	BiLSTM-CRF(baseline)	93.09	94.34	93.55
2	+radical(CNN)	93.90	93.89	93.90
3	+radical(BiLSTM)	93.87	93.73	93.80
4	+radical(BiLSTM-CNN)	93.46	94.23	93.84
5	+CNNs+radical(CNN)	94.57	93.38	93.97
6	+Attention+radical(CNN)[7]	94.05	94.54	94.29
7	+CNNs+Attention(RS-ADP)	94.16	94.47	94.32

能。而 CNN 和 BiLSTM 联合提取策略性能稍低。理论上，该机制能够同时捕获部首更加丰富的语义特征（空间特征和序列特征），而实际应用中其效果并不理想，可能原因是仅采用级联方法无法有效融合 2 种特征信息，下一步将继续探究更有效的特征融合方法。综上，本文采用 CNN 提取部首嵌入特征。

2.2.3　CNN 层

采用 BiLSTM－CRF 为基准模型，分别以表 4 实验组 2、5 和实验组 6、7 为对照组，以验证 CNN 层的有效性。由实验结果可知，本文所提基于 CNN 层的模型的 F_1 值为 93.97%，且其识别准确性明显提升。在实验组 6、7 中，采用基于 CNN 层的模型同样取得最优值。这得益于本文所提不同窗口尺寸的多层 CNN 能够充分提取不同尺度的局部上下文特征，进而丰富语义信息。实验结果验证了本文所提 CNN 层的有效性。

2.2.4　Self－attention 机制

如表 4 中实验组 2、6 结果所示，在采用 Self－attention 机制情况下，模型对 AgCNER 数据集的 P、R 和 F_1 值分别提高了 0.15、0.65、0.39 个百分点；对照组 5、7 也有类似结果，不同点在于由于 CNN 层的加入，其差值进一步扩大，Self－attention 机制的性能得到进一步发挥。此外，本文以“水稻易患由稻瘟菌引起的稻瘟病”为例，可视化了权重矩阵 $\mathrm{softmax}(\boldsymbol{QK}^{\mathrm{T}}/\sqrt{d})$，以便直观理解 Self－attention 的作用机制。如图 7 所示，Self－attention 机制能够捕获不同距离的上下文依赖信息，并分配不同权重。Self－attetion 机制学习到“水稻”与“易患”和“引起”之间的依赖关系。通过颜色可知，与“引起”相比，“水稻”与“易患”具有更强的连接，这是因为“水稻”与“易患”的主谓关系要比从句关系更紧密。此外，Self－attention 还学到了“稻瘟菌”与“稻瘟病”的依赖，这除了能够说明“稻瘟菌”与“稻瘟病”本身既有的因果关系外，还说明了 Self－attention 机制具备

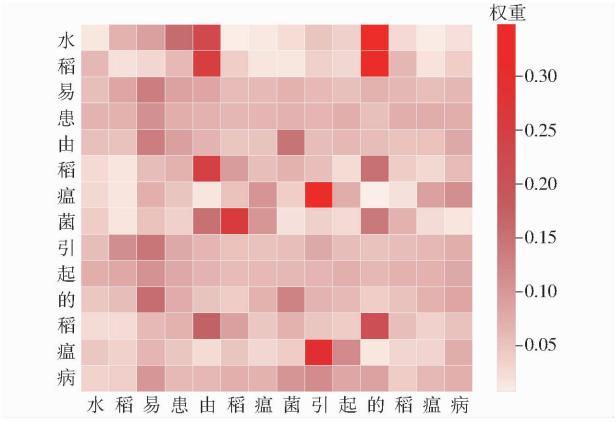

图 7　Self－attention 相似性可视化权值

Fig. 7　Visual weights of Self－attention on AgCNER

捕获更长距离依赖的能力。

2.2.5　模型性能比较

将 P、R 和 F_1 值作为目标参数，将 RS－ADP 与 BERT－IDCNN－CRF[18]、BERT－BiLSTM－CRF、TENER[30]、RD－CNN－CRF、GATEDCNN－CRF[31]、IDCNN[17]、BiLSTM－Attention－CRF[7] 等主流模型进行了比较，实验结果如表 5 所示。

表 5　各模型在 AgCNER 数据集上的识别结果

Tab. 5　Results of each model on AgCNER　%

模型	P	R	F_1 值
BERT－IDCNN－CRF	91.56	91.77	91.60
BERT－BiLSTM－CRF	93.69	91.41	92.54
TENER	94.14	93.52	93.83
RD－CNN－CRF	92.46	93.43	92.95
GATEDCNN－CRF	93.28	93.74	93.50
IDCNN	93.46	93.99	93.72
BiLSTM－attention－CRF	94.05	94.54	94.29
RS－ADP	94.16	94.47	94.32

BERT－IDCNN－CRF、BERT－BiLSTM－CRF 是基于 BERT 预训练嵌入的模型，虽然在 AgCNER 上取得一定效果，但由于未在农业病虫害文本上进行预训练以及句子长度过长导致语义缺失等原因导致了其性能未能充分发挥。此外，RD－CNN－CRF、GATEDCNN－CRF 和 IDCNN 等基于 CNN 的命名实体识别模型的识别结果略低，原因在于虽然这些模型在提取局部上下文特征方面优势明显，但提取全局上下文特征能力有待提升，因此不是识别农业病虫害实体的最佳模型。RS－ADP 模型在 AgCNER 数据集上取得了最大 F_1 值，一方面是因为该模型充分考虑了部首特征和局部上下文特征丰富的语义信息，另一方面，Self－attention 机制的使用进一步增强了模型提取长距离上下文特征的能力。因此，上述实验结果充分说明了 RS－ADP 模型能够有效识别农业病虫害命名实体。

2.2.6　AgCNER 上的识别详情分析

表 6 列出了 RS－ADP 对 AgCNER 数据集各类实体的识别结果。由实验结果可知，RS－ADP 对病害、虫害、作物、药剂、周期及其他的识别效果要明显优于其他类别，尤其针对作物、病害和虫害的识别的最大 F_1 值分别达到 95.81%、97.76% 和 97.23%，这是因为上述实体往往存在如“病”、“虫”、和“期”等明显的边界特征。另外，“疒”、“酉”和“月”等部首特征也发挥了一定作用。RS－ADP 模型充分利用了部首特征和局部上下文特征，同时具有较强的捕获长距离依赖的能力，因此对作物品种、草害和病原等难识别实体仍保持 86% 以上的 F_1 值。

表 6 RS－ADP 在各类实体上的识别结果

Tab. 6 Results of RS－ADP for each category on AgCNER %

类别	P	R	F_1 值
其他	94.60	95.80	95.20
作物	95.85	95.78	95.81
病害	97.49	98.04	97.76
药剂	91.92	91.67	91.80
肥料	81.31	78.86	80.07
部位	88.94	89.67	89.30
周期	93.17	93.38	93.27
虫害	97.27	97.18	97.23
病原	88.34	89.04	88.69
作物品种	86.20	86.09	86.14
草害	90.35	92.38	91.35

当然，AgCNER 数据集中仍然存在部分识别错误的实体，除了部分完全识别错误的实体外，较典型的错误如表 7 所示：①存在部分未标注实体。如“病原”未标注，但识别正确。②存在部分未完整识别实体。如病害“水稻烂秧”、“综合防治技术”，这是因为存在实体嵌套，导致了识别的不一致性。③相邻实体间界限模糊。如“雌花穗花丝”本应识别为“雌花穗”和“花丝”，但由于其实体间界限模糊，干扰了模型的准确性。因此，在以后工作中，会进一步修正数据集，将人为噪声降到最低。

表 7 部分识别错误实体类型

Tab. 7 Some incorrect recognized examples

实体	标注标签	预测标签	实体	标注标签	预测标签
水	B－DIS	B－CRO	病	O	B－OTH
稻	I－DIS	I－CRO	原	O	I－OTH
烂	I－DIS	O	成	O	B－PER
秧	I－DIS	O	灾	O	I－PER
综	B－OTH	B－OTH	期	O	I－PER
合	I－OTH	I－OTH	雌	B－PART	B－PART
防	I－OTH	I－OTH	花	I－PART	I－PART
治	I－OTH	I－OTH	穗	B－PART	I－PART
技	O	I－OTH	花	I－PART	B－PART
术	O	I－OTH	丝	O	I－PART

注：DIS 为病害，CRO 为作物，OTH 为其他，PER 为周期，PART 为部位。

2.2.7 其他领域数据集识别详情分析

为了验证 RS－ADP 模型的泛化性，选用 2017 年全国知识图谱与语义计算大会（China Conference on Knowledge Graph and Semantic Computing，CCKS2017）公布的临床医学实体标注语料作为对照数据集，该数据集包含 5 种类别，即疾病、检查、症状、治疗和身体部位。由表 8 可知，RS－ADP 在该数据集上取得最大 F_1 值，明显优于 RD－CNN－CRF 和 GATEDCNN－CRF 等基于 CNN 的模型。并且相对于基准模型 BiLSTM－CRF，其 F_1 值提高了 1.11 个百分点。虽然 IDCNN 的识别结果略低于基准模型，但在实验过程中，该模型可利用 GPU 并行性，训练速度明显加快。TENER、BERT－BiLSTM－CRF 和 FT－BERT＋BiLSTM ＋CRF 等基于 Transformer 模型在 CCKS2017 上的识别结果低于 RS－ADP，原因可能是由于 CCKS2017 数据量较小影响了上述模型识别性能。综上，上述结果不仅说明了 RS－ADP 模型的有效性，而且还说明该模型具有一定的泛化能力。

表 8 各模型在 CCKS2017 上的识别结果

Tab. 8 Results for each model on CCKS2017 %

模型	P	R	F_1 值
BiLSTM－CRF	88.58	93.71	91.07
RD－CNN－CRF	90.63	92.02	91.32
GATEDCNN－CRF	90.02	92.44	91.21
TENER	91.24	93.08	92.15
IDCNN	91.85	89.32	90.53
BERT－IDCNN－CRF	91.44	92.14	91.79
BERT－BiLSTM－CRF	93.30	86.79	89.74
FT－BERT＋BiLSTM＋CRF[32]	92.06	91.15	91.60
RS－ADP	90.81	93.60	92.18

3 结束语

针对农业病虫害命名实体识别任务，提出了 RS－ADP 模型，该模型结合部首特征和字符嵌入在一定程度上解决了内在语义信息缺失的问题，丰富了语义信息。根据文献检索这是首次将部首特征用于农业病虫害领域的命名实体识别任务。其中，本文基于 CNN 和 BiLSTM 分别设计了 3 种部首级特征提取策略，实验结果表明 CNN 策略能够最大程度提取部首特征。此外，采用多层不同窗口尺寸的 CNN 层用于提取文本局部上下文信息。最后，采用 Self－attention 机制进一步解决了 BiLSTM 模型无法有效捕获长距离依赖的问题。在自标注语料 AgCNER 上的多方面的实验结果表明，本文采用的预训练模型、部首嵌入、CNN 层和 Self－attention 机制有助于提升模型性能，且与其他典型模型相比，RS－ADP 在 AgCNER 上取得了最优 F_1 值，充分说明了其在农业病虫害命名实体识别方面的优越性。同时在 CCKS2017 语料上的实验结果表明该模型还具有一定的泛化能力。

参 考 文 献

[1] 张善文，王振，王祖良．结合知识图谱与双向长短时记忆网络的小麦条锈病预测[J]．农业工程学报，2020，36(12)：172－178.
ZHANG Shanwen, WANG Zhen, WANG Zuliang. Prediction of wheat stripe rust disease by combining knowledge graph and

bidirectional long short term memory network[J]. Transactions of the CSAE, 2020, 36(12): 172 - 178. (in Chinese)
[2] 沈利言,姜海燕,胡滨,等.水稻病虫草害与药剂实体关系联合抽取算法[J].南京农业大学学报, 2020,43(6):1151 - 1161.
SHEN Liyan, JIANG Haiyan, HU Bin, et al. A study on joint entity recognition and relation extraction for rice diseases pests weeds and drugs[J]. Journal of Nanjing Agricultural University, 2020,43(6):1151 - 1161. (in Chinese)
[3] 李想,魏小红,贾璐,等.基于条件随机场的农作物病虫害及农药命名实体识别[J/OL].农业机械学报,2017,48(增刊):178 - 185.
LI Xiang, WEI Xiaohong, JIA Lu, et al. Recognition of crops, diseases and pesticides named entities in Chinese based on conditional random fields[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2017, 48(Supp.): 178 - 185. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? flag=1&file_no=2017s029&journal_id=jcsam. DOI:10.6041/j.issn.1000-1298.2017.S0.029. (in Chinese)
[4] MALARKODI C S, ELISABETH L, SOBHA L D. Named entity recognition for the agricultural fomain[J]. Research in Computing Science, 2016,117: 121 - 132.
[5] 王春雨, 王芳. 基于条件随机场的农业命名实体识别研究[J]. 河北农业大学学报, 2014,37(1):132 - 135.
WANG Chunyu, WANG Fang. Study on recognition of Chinese agricultural named entity with conditional random fields[J]. Journal of Agricultural University of Hebei, 2014,37(1):132 - 135. (in Chinese)
[6] 郑丽敏,齐珊珊,田立军,等.面向食品安全事件新闻文本的实体关系抽取研究[J/OL].农业机械学报,2020,51(7):244 - 253.
ZHENG Limin, QI Shanshan, TIAN Lijun, et al. Entity relation extraction of news texts for food safety events[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(7): 244 - 253. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? flag=1&file_no=20200728&journal_id=jcsam. DOI:10.6041/j.issn.1000-1298.2020.07.028. (in Chinese)
[7] YIN M, MOU C, XIONG K, et al. Chinese clinical named entity recognition with radical-level feature and self-attention mechanism[J]. Journal of Biomedical Informatics, 2019, 98: 103289.
[8] QIU J, WANG Q, 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). IEEE, 2018: 935 - 942.
[9] GIORGI J M, BADER G D. Towards reliable named entity recognition in the biomedical domain[J]. Bioinformatics, 2020, 36(1):280 - 286.
[10] AGUILAR G, MAHARJAN S, LOPEZ-MONROY A P, et al. A multi-task approach for named entity recognition in social media data[C]//Proceedings of the 3rd Workshop on Noisy User-generated Text, 2017: 148 - 153.
[11] HUANG Z, XU W, YU K. Bidirectional LSTM - CRF models for sequence tagging[J]. arXiv Preprint arXiv:1508.01991, 2015.
[12] MISAWA S, TANIGUCHI M, MIURA Y, et al. Character-based bidirectional LSTM - CRF with words and characters for Japanese named entity recognition[C]//Proceedings of the First Workshop on Subword and Character Level Models in NLP, 2017: 97 - 102.
[13] MA X, HOVY E. End-to-end sequence labeling via bi-directional LSTM - CNNs - CRF[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2016: 1064 - 1074.
[14] GRIDACH M. Character-level neural network for biomedical named entity recognition[J]. Journal of Biomedical Informatics, 2017,70:85 - 91.
[15] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural architectures for named entity recognition[C]//Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016: 260 - 270.
[16] CHO M, HA J, PARK C, et al. Combinatorial feature embedding based on CNN and LSTM for biomedical named entity recognition[J]. Journal of Biomedical Informatics, 2020, 103: 103381.
[17] STRUBELL E, VERGA 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.
[18] 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 Ni, GUAN Huanmei, YANG Piao, et al. BERT - IDCNN - CRF for named entity recognition in Chinese[J]. Journal of Shandong University (Natural Science), 2020, 55(1): 102 - 109. (in Chinese)
[19] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//NAACL - HLT (1), 2019.
[20] QIU X, SUN T, XU Y, et al. Pre-trained models for natural language processing: a survey[J]. arXiv Preprint arXiv:2003.08271, 2020.
[21] LI J, SUN A, HAN J, et al. A survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2020,99:1 - 1.
[22] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[J]. arXiv Preprint arXiv:1301.3781,2013.
[23] XU C, WANG F, HAN J, et al. Exploiting multiple embeddings for Chinese named entity recognition[C]//Proceedings of the 28th ACM International Conference on Information and Knowledge Management, 2019: 2269 - 2272.
[24] KRIZHEVSKY A, SUTSKEVER I, HINTON G E. Imagenet classification with deep convolutional neural networks[C]//Advances in Neural Information Processing Systems, 2012: 1097 - 1105.
[25] DONG C, ZHANG J, ZONG C, et al. Character-based LSTM - CRF with radical-level features for Chinese named entity recognition[M]//Natural Language Understanding and Intelligent Applications. Springer, Cham, 2016: 239 - 250.
[26] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. Nature, 2015, 521(7553): 436 - 444.
[27] HAO X, JIA J, KHATTAK A M, et al. Growing period classification of gynura bicolor DC using GL - CNN[J]. Computers and Electronics in Agriculture, 2020, 174: 105497.
[28] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems, 2017: 5998 - 6008.
[29] WALLACH H M. Conditional random fields: an introduction[J]. Technical Reports (CIS), 2004,53(2): 22.
[30] YAN H, DENG B, LI X, et al. Tener: adapting transformer encoder for name entity recognition[J]. arXiv Preprint arXiv: 1911.04474, 2019.
[31] WANG C, CHEN W, XU B. Named entity recognition with gated convolutional neural networks[M]//Chinese Computational Linguistics and Natural Language Processing Based on Naturally Annotated Big Data. Springer, Cham, 2017: 110 - 121.
[32] LI X, ZHANG H, ZHOU X H. Chinese clinical named entity recognition with variant neural structures based on BERT methods[J]. Journal of Biomedical Informatics, 2020,107: 103422.