doi:10.6041/j.issn.1000-1298.2021.12.027

基于多源信息融合的中文农作物病虫害命名实体识别

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

摘要: 随着农作物病虫害研究文献的快速增长,对农作物病虫害领域文献进行文本挖掘变得越来越重要。开发有效、准确的农作物病虫害命名实体识别系统有助于在农作物病虫害相关研究报告中提取研究成果,为农作物病虫害的治理提供有效建议。本文针对中文农作物病虫害数据集缺失问题,提出了基于半远程监督的停等算法,利用该算法构建中文农作物病虫害领域语料库,大幅度减少标注过程的人工成本和时间成本;同时,提出了中文农作物病虫害命名实体识别模型(Agricultural information extraction, Agr-IE),该模型基于 BERT-BILSTM-CRF,辅以多源信息融合(多源分词信息和全局词汇嵌入信息)丰富字符向量,使其充分结合字符级与词汇级的信息,以提高模型捕捉上下文信息的能力。实验表明,该模型可以有效地识别病害、虫害、药剂、作物等实体,F1 值分别为 96.56%、95.12%、94.48%、95.54%,并对识别难度较大的病原实体具有较好的识别效果,F1 值为 81.48%,高于 BERT-BILSTM-CRF、BERT 等模型的相应值。本文所提模型在 MSRA 和 Weibo 等其他领域数据集上与 CAN-NER、Lattice-LSTM-CRF 等模型进行了对比实验,并取得最佳的识别效果,F1 值分别为 95.80%、94.57%,表明该算法具有一定的泛化能力。

关键词: 命名实体识别; 农作物病虫害; 农业自然语言处理; 中文分词; 停等算法
中图分类号: TP391.1　　**文献标识码**: A　　**文章编号**: 1000-1298(2021)12-0253-11　　**OSID**:

Named Entity Recognition of Diseases and Insect Pests Based on Multi Source Information Fusion

LI Lin　ZHOU Han　GUO Xuchao　LIU Chengqi　SU Jie　TANG Zhan
(*College of Information and Electrical Engineering*, *China Agricultural University*, *Beijing* 100083, *China*)

Abstract: Crop diseases and insect pest text mining is becoming increasingly important as the number of crop diseases and insect pest documents rapidly grows. The development of effective and highly accurate named entity recognition (NER) systems of crop diseases and insect pests can be beneficial to extract research results from related research reports and provide effective suggestions for the control of diseases and insect pests. Stop-wait algorithm based on semi-remote supervision was proposed to construct the corpus of Chinese crop diseases and insect pests to solve the problem of corpus missing. Moreover, an agricultural information extraction (Agr-IE) method was proposed. The method was based on BERT-BILSTM-CRF, and multi-source word segmentation information and global lexical embedding was used to enrich the information of character vector before character information integrated. Experiments performed by Agr-IE on the datasets of crop diseases and insect pests showed that the model can effectively distinguish four types of entities: the F1 score of diseases, pests, pharmaceuticals, and plant were 96.56%, 95.12%, 94.48% and 95.54%, respectively. And the model also performed well in identifying entities about pathogens (81.48% F1 score), which was higher than the corresponding values of BERT-BILSTM-CRF, BERT and other models. The recognition effect was higher than that of the compared models. In addition, the proposed model was compared with CAN-NER, Lattice-LSTM-CRF and other models on MSRA, Weibo datasets, and the best recognition results were obtained. The F1 scores were 95.80% and 94.57% respectively, which showed that the algorithm had good generalization ability and stability.

Key words: named entity recognition; crop diseases and insect pests; agricultural natural language processing; Chinese word segmentation; Stop-wait algorithm

收稿日期: 2020-12-05　修回日期: 2021-01-01
基金项目: 国家重点研发计划项目(2016YFD0300710)
作者简介: 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

0 引言

农作物病虫害知识图谱构建的基础是命名实体识别,即从未加工的农作物病虫害文本中识别出特定类别的专有名词实体。其准确率直接影响农作物病虫害领域多种自然语言处理技术的结果[1-3]。

国内外众多学者对命名实体识别进行了研究,并取得了较多研究成果[4-8]。但这些方法都基于机器学习,需要人工添加大量特征,无法适应数据模式的变化。近年来,研究人员结合不同的半监督学习和深度神经网络(Deep neural networks,DNN),以找到命名实体识别、关系抽取[9]和知识图谱构建等分块任务的最佳解决方案[10]。在英语命名实体识别上,文献[11]使用基于长短时记忆网络和条件随机场相结合(Long short-term memory - conditional random field,LSTM - CRF)的中文分词联合训练方法,精确度提高了 5%。文献[12]首次将双向长短时记忆网络和条件随机场相结合的方法(Bidirectional long short-term memory - conditional random field,BILSTM - CRF)应用于自然语言处理(Natural language processing,NLP)基准序列标记数据集,与之前的工作相比,取得了更好的结果。文献[13]使用迭代扩张卷积神经网络(Iterated dilated convolutional neural networks,IDCNN)方法,从整个文档中聚合上下文的 IDCNN 识别实体,准确性较高,且相对于长短时记忆网络(Long short-term memory,LSTM),IDCNN 可实现并行计算,提高了实体识别速度。

与英语命名实体识别相比,汉语命名实体识别研究起步较晚,难度更大[14-16]。嵌入字或者词汇向量能较好地捕捉中文语义信息。文献[17]提出了 Lattice - LSTM 模型,该模型利用字符级别和词汇级别的信息,在 OntoNotes、MSRA、Weibo 和 Resume 等 4 种中文数据集上取得了较好的效果。文献[18]提出了一种卷积注意力网络模型(Convolutional attention network for named entity recognition,CAN - NER),该模型使用字符级特征捕获单词级特征和上下文信息,在不同领域的中文数据集上取得了很好的性能。文献[19]提出了 BERT - IDCNN - CRF 模型,该模型使用 BERT 嵌入词向量,在中文命名实体任务上优于 Lattice - LSTM 模型,F1 值提高了 1.23%,与基于 BERT 微调的方法相比,该方法的 F1 值略低,但是训练时间大幅度缩短。

目前,中文农作物病虫害命名实体识别存在如下问题:已有的相关工作是基于传统机器学习方法,需要一种能自动学习文本特征的农作物病虫害命名实体识别方法;文献[20-21]虽然提出了深度学习的方法对农业命名实体进行识别,但该方法不能很好地捕捉词汇的边界信息,阻碍了模型对词汇信息的充分利用;中文农作物病虫害语料缺失,增加了对相关命名实体识别的难度。

为解决上述问题,本文提出一种基于半远程监督的停等算法,构建中文农作物病虫害语料库;提出一种基于多源分词信息和全局词汇嵌入信息的神经网络模型 Agr - IE,该模型以 BERT - BILSTM - CRF[22]为基本框架,采用实体词典与多种分词工具相结合的方式划分词汇边界;采用加权的方式重置词汇中各字符向量,使其充分结合字符级与词汇级的信息,以提高模型捕捉上下文信息的能力。

1 材料与方法

1.1 数据收集与标注

1.1.1 中文农作物病虫害数据集构建

对比农药百科、农业问答等文本数据,论文摘要蕴含了大量农作物病虫害知识,且更具有权威性和可靠性。通过挖掘论文摘要中“农作物-病(虫)害-农药”和“农作物-病(虫)害-病原”等关联信息构建农作物病虫害知识图谱,能准确收集最新的农作物病虫害防治方法。本文基于爬虫技术,采用深度优先搜索的策略获取 CNKI 网站中农作物病虫害类期刊论文摘要作为原始文本数据。本文研究目标主要是识别病害、虫害、病原、药剂和作物等实体,实体定义和样例如表 1 所示。

表 1 实体定义和样例

Tab. 1 Definition of entity and corresponding examples and notes

实体	定义	样例
病害	农作物病害名称	水稻纹枯病、赤霉病
虫害	农作物虫害名称	稻飞虱、白粉病、玉米螟
病原	引起农作物病害的病原微生物	水稻纹枯病菌、藤仓镰刀菌
药剂	防治农作物病虫害的药剂学名、俗名、生物防治药剂名	瑞劲特、稻丰收
作物	农作物名称	水稻、小麦、马铃薯

在摘要数据中,包含了英文大小写、简写、连字符和实体嵌套等多种情况,如“水稻条叶纹枯病”为“水稻”和“条叶纹枯病”两种实体组成,摘要数据示例如图 1 所示。

1.1.2 数据标注

本文自主开发了一款标注软件“LableText”来半自动标注摘要数据中的实体,减少了标注数据所需要的人力成本以及时间成本,如图 2 所示。

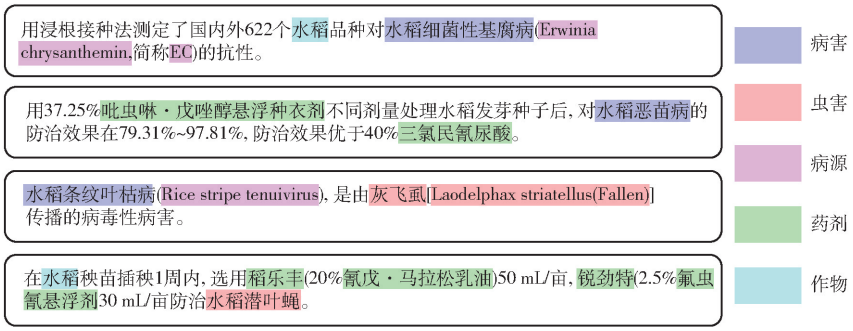

图 1　摘要数据示例

Fig. 1　Abstract data example

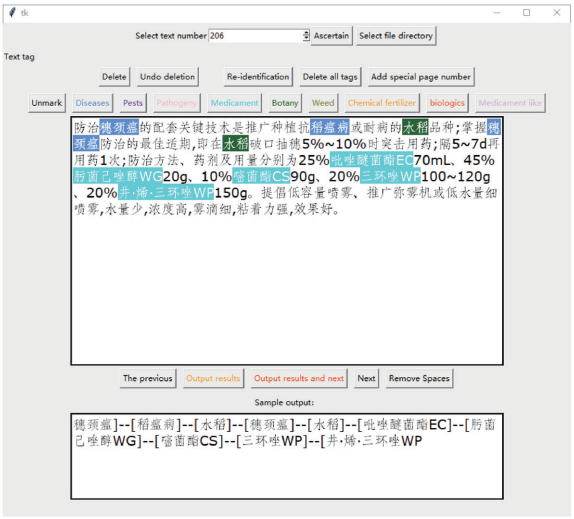

图 2　基于 LableText 的标注示意图

Fig. 2　Schematic of sample annotation based on LabelText

“LableText”使用停等算法实现了原始数据分析处理。

本文随机选取 20 000 条摘要进行标注。命名实体识别的常用标签模式包括“BIO”、“BIEO”和“BIEOS”。由于本文实体识别算法模型 Agr－IE 的需求,故采用“BIEOS”表达策略,其中“B”表示实体的开头,“I”表示实体的中间部分,“E”表示实体的结尾,“S”表示单个字符表示的实体,“O”表示不是实体。表 2 给出了 BIEOS 标记的样例。

1.1.3　数据集划分

在预处理 20 000 条摘要文本数据后,剩下 16 014 条可用数据。将剩下的文本数据划分为训练集、测试集和验证集。其中训练集 10 014 条,验证集和测试集都为 3 000 条。为了适应训练中训练集

表 2　文本以及相应的 BIEOS 格式标签样例

Tab. 2　Examples of text and corresponding BIEOS format tags

句子/字符	BIEOS 标签
不同制剂的锐劲特单用或与常用农药 20% 三唑磷乳油混合使用,施用 1 次即可对二化螟均有较好的防治效果,且持效时间长,可在生产中推广使用。	(O)不(O)同(O)制(O)剂(O)的(O)锐(B－DRUG)劲(I－DRUG)特(E－DRUG)单(O)用(O)或(O)与(O)常(O)用(O)农(O)药(O)2(O)0(O)%(O)三(B－DRUG)唑(I－DRUG)磷(I－DRUG)乳(I－DRUG)油(E－DRUG)混(O)合(O)使(O)用(O),(O)施(O)用(O)1(O)次(O)即(O)可(O)对(O)二(B－PEST)化(I－PEST)螟(E－PEST)均(O)有(O)较(O)好(O)的(O)防(O)治(O)效(O)果(O),(O)且(O)持(O)效(O)时(O)间(O)长(O),(O)可(O)在(O)生(O)产(O)中(O)推(O)广(O)使(O)用(O)。(O)
稻。	稻(S)。(O)

的规模,设置训练集与测试集的比例高于传统的样本划分。训练集、验证集、测试集中句子、实体的数量以及未登录词(Out of vocabulary,OOV)比例如表 3 所示。

1.2　基于半远程监督的停等算法

为减少数据标注过程中产生的人力成本与时间成本,本文提出停等算法半自动标注已获取的摘要文本。该算法通过词典与人工相结合的方式标注部分数据,在进行一部分标注后,以已标注的数据作为训练集,未标注的数据作为测试集,执行一次算法,将最后获取的词人工检查后存入词典,并对剩下数据进行标注,随着词汇增加,能自动识

表 3　农作物病虫害数据集划分统计

Tab. 3　Statistics of crop diseases and insect pests dataset division

实体	训练集		验证集		测试集		OOV
	句子	标签	句子	标签	句子	标签	比例/%
病害	5 452	11 907	1 478	4 266	2 129	6 184	22. 16
虫害	3 641	10 297	831	2 998	771	2 019	21. 36
病原	959	2 229	234	327	265	268	24. 14
药剂	2 707	7 629	843	3 873	973	3 374	22. 23
作物	9 721	32 941	2 375	9 424	2 523	8 243	23. 54

别的实体数量逐渐增大,极大地减少了数据标注过程中产生的人力成本与时间成本。为创建初始领域词典,本文选取具有公信力的政府平台、大型

企业和科研院所的官方网站作为词典语料的主要来源，例如中国农药信息网（http：// www. chinapesticide. org. cn/hysj/index. jhtml）、中国农业信息网（http：// www. agri. cn/）、山东省农业和农村厅植保技术网（http：// www. sdny. gov. cn/yysjk/zbjs/）和世纪农药网（https：// www. nongyao001. com/sell/）等。从这些网站上共获取 41 583 条结构化信息，因药剂名称具有多样性，本文将药剂名称与剂型合并后作为新的词汇，并将新词汇与原来的药剂名称一起存入词典，例如，将“丙草胺”与“乳油”合并，得到词汇“丙草胺乳油”，并将“丙草胺乳油”与“丙草胺”一起存入药剂词典。将所有实体存在的别名和英语名称，采用重复性检查后再存储到药剂词典中，最终，本文得到词汇数量为 113 915 的领域词典。

本文使用 $f(x;\theta_{(0)})$ 表示由 θ 参数化的模型 LSTM－CRF，$\{Y_n\}_{n=1}^{N}$ 表示已标记部分实体的句子，N 为句子长度，$\{Z_m\}_{m=1}^{M}$ 表示模型停止时识别出的实体，M 为实体的数量。停等算法如下：

Input：未标记的句子 $\{X_n\}_{n=1}^{N}$；已构建的词典 K_{ws}；模型 LSTM－CRF，$f(x;\theta_{(0)})$；迭代次数 T

// 词典匹配

$$\{Y_n\}_{n=1}^{N}=\text{Matching}(\{X_n\}_{n=1}^{N},K_{ws})$$

//模型训练

For $t=1,2,\cdots,T$ do

$$f(\{Y_n\}_{n=1}^{N};\theta_{(0)})$$

更新模型 $f(x;\theta_{(0)})$：

$$\theta_{(t)}=V(\theta_{(t-1)})$$

//得到循环 T 次后的模型以及每次识别出的实体：

$$f(x;\theta_{(T)}),\{Z_m\}_{m=1}^{M}$$

//统计每次迭代过程中都识别出来的实体：

For $t=1,2,\cdots,T$ do

$$D_a=\{\{Z_m\}_{m=1}^{M}\}_{t=1}^{T}$$

// 将获取到的新词存入词典 K_{ws}

$$K_{ws}=\{K_{ws},D_a\}$$

Output：新词典 K_{ws}

停等算法通过词典与人工先验知识相结合的方式扩充了领域词典，降低了标注软件自动识别实体过程中所产生的噪声，提高了软件自动识别实体的精度。同时，为了防止模型出现过拟合的情况，本文将 LSTM－CRF 的迭代次数设置为 7，使模型提前停止迭代，进而提升模型的召回率。因此，该算法可以更好地识别未知实体并将其分类。图 3 展示了各情况下模型对实体识别效果的拟合图。

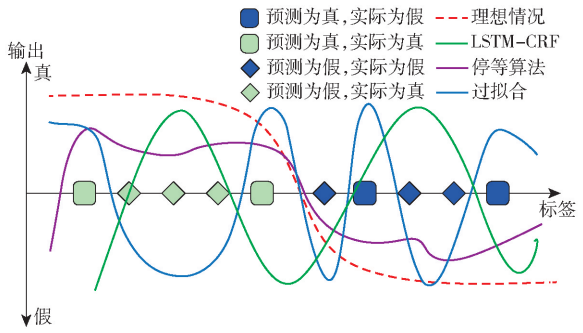

图 3 不同情况下模型与理想情况的拟合图

Fig. 3 Fitting diagram of model annotation effect in different case

2 研究方法

近年来，中文命名实体识别工作通常采用字符级嵌入或使用预训练模型处理数据，这种方式虽然能取得不错的效果，但不能充分利用词汇级别信息。本文在模型中引入分词信息与全局词汇嵌入信息优化表示字符向量，以获取丰富的语义知识。本文提出的 Agr－IE 模型结构如图 4 所示，模型主要包括字符嵌入层、优化表示层和 BILSTM－CRF 层。其中，e_1～e_7 表示 BILSTM 对句子的正向建模信息，e'_1～e'_7 表示 BILSTM 对句子的反向建模信息。

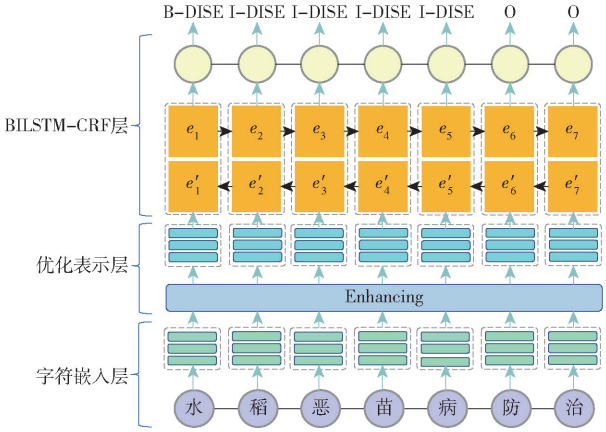

图 4 Agr－IE 结构图

Fig. 4 Structure diagram of Agr－IE

2.1 字符嵌入层

本文使用 BERT 中文语言模型嵌入离散的文本。BERT 在预训练时，使用 WordPiece 模型创建一个包含所有字符以及所训练语料库中常见词汇和子词汇的 tokenizer。为了将词汇表示为 tokenizer 中单个字符的集合，BERT 在新的语料训练时对词汇表中的词汇分解为尽可能大的子词，最后将词汇分解为子字符。例如：不给“水稻恶苗病”和词汇表之外的词汇分配重载的未知词汇表标记，而是将该词汇拆分为子单词标记[水、##稻、##恶、##苗、##病]，使组合后的词汇能够保留原始词汇的上下文含义。

2.2　优化表示层

使用 BERT 预训练模型能在一定程度上利用词汇级别的信息，但仍存在词汇级别信息利用不充分问题，对此，本文提出了优化表示层解决该问题。优化表示层主要分为词边界信息融合和分词信息整合两部分，结构如图 5 所示。

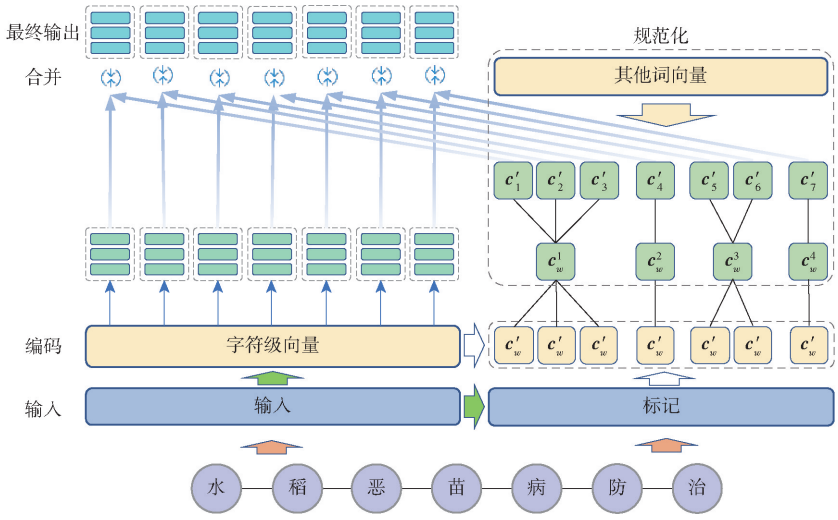

图 5　优化表示层结构图

Fig. 5　Structural model diagram of optimized presentation layer

在分词边界信息融合中，给定句长为 N 的句子 $\{S_n\}_{n=1}^{N}$，利用分词工具对输入的句子进行处理后得到 M 个词汇，记为 $\{W_m\}_{m=1}^{M}$，$M \leqslant N$。模型利用所有输入字符进行自我注意力运算，得到字符级注意得分矩阵 $\boldsymbol{A}_w$，$\boldsymbol{A}_w$ 的计算公式为

$$\boldsymbol{A}_w = \text{softmax}\left(\frac{(\boldsymbol{K}\boldsymbol{W}_1)(\boldsymbol{K}^{\mathrm{T}}\boldsymbol{W}_2)^{\mathrm{T}}}{\sqrt{d}}\right) \tag{1}$$

式中　$\boldsymbol{W}_1$、$\boldsymbol{W}_2$——可训练权重矩阵

$\boldsymbol{K}$——预训练向量的编码表示

softmax——归一化指数函数

同时，本文将矩阵 $\boldsymbol{A}_w$ 写成以字符串为序列的集合

$$\boldsymbol{A}_w = \{\boldsymbol{a}_w^1, \boldsymbol{a}_w^2, \cdots, \boldsymbol{a}_w^i, \cdots, \boldsymbol{a}_w^n\} \tag{2}$$

式中　$\boldsymbol{a}_w^i$——矩阵 $\boldsymbol{A}_w$ 的第 i 行向量

设 $\{W_m\}_{m=1}^{M} = [\{s_1, s_2, s_3\}, \{s_4\}, \cdots, \{s_n\}]$，则 $\{W_m\}_{m=1}^{M}$ 的向量表示 $\boldsymbol{B}_m$ 为

$$\boldsymbol{B}_m = (\{\boldsymbol{a}_w^1, \boldsymbol{a}_w^2, \boldsymbol{a}_w^3\}, \{\boldsymbol{a}_w^4\}, \cdots, \{\boldsymbol{a}_w^n\}) = (\boldsymbol{b}_1, \boldsymbol{b}_2, \cdots, \boldsymbol{b}_m) \tag{3}$$

随后，优化表示层模型分别归一化表示每组向量，使每个词汇得到一个行向量编码 $\boldsymbol{c}_w^i$。例如“水稻”一词包含“水”、“稻”两个字符，因此，将“水”、“稻”两个字符的表示向量划分为一组，将其归一化后，得到“水稻”的表示向量。该模型使用平均池化与最大池化相结合的方式归一化计算词向量，并利用可学习参数 γ 对两种池化方式进行平衡。$\boldsymbol{c}_w^i$ 的计算公式为

$$\boldsymbol{c}_w^i = \gamma\text{Maxpooling}(\boldsymbol{b}_i) + \text{Avepooling}(\boldsymbol{b}_i) \tag{4}$$

式中　Maxpooling——最大池化函数

Avepooling——平均池化函数

但这种方法只考虑了当前词汇信息的表示，而没有考虑到词汇中字符的位置信息，为此，匹配结果无法复原原始的词汇信息，从而导致信息损失。例如“稻瘟病”和“水稻”两个词汇中“稻”的表示向量相同，但位置信息不同。本文使用静态加权的方式对词汇集合进行编码，用来解决位置信息缺失问题。例如，在“稻瘟病”和“水稻”两个词汇中，“稻”在前者的标签为“B”，在后者中的标签为“I”，则将“稻瘟病”和“水稻”分配到不同的集合中，并加权计算所有集合 1 与集合 2 的词汇编码，得到“稻”的表示向量 $\boldsymbol{c}_i'$ 为

$$\boldsymbol{c}_i' = \frac{1}{D+E}\sum\left(d(\boldsymbol{c}_w^i)\boldsymbol{c}_w^i + e(\boldsymbol{c}_w^j)\boldsymbol{c}_w^j\right) \tag{5}$$

式中　D——集合 1 中的词汇数量

E——集合 2 中的词汇数量

$\boldsymbol{c}_w^j$——集合 2 中的词汇向量

$d(\boldsymbol{c}_w^i)$——集合 1 中的相应词汇的词频

$e(\boldsymbol{c}_w^j)$——集合 2 中的相应词汇的词频

为使向量维度与初始向量一致，本文模型采用上采样的方法整合两种编码：将该词包含的所有字符在原始字符串中的位置都赋予 $\boldsymbol{c}_i'$，本文使用 $\boldsymbol{A}_{nw}$ 表示对齐后的注意力矩阵

$$\boldsymbol{A}_{nw} = [\{\boldsymbol{a}_{nw}^1, \boldsymbol{a}_{nw}^1, \boldsymbol{a}_{nw}^1\}\quad \{\boldsymbol{a}_{nw}^2\}\quad \cdots\quad \{\boldsymbol{a}_{nw}^m\}] \tag{6}$$

为了融合上下文信息，利用本文模型计算 attention 矩阵与原始预训练向量编码矩阵 $\boldsymbol{P}$，计算公式为

$$\boldsymbol{S} = \boldsymbol{A}_{nw}\boldsymbol{P}\boldsymbol{W}_n \tag{7}$$

式中 $\boldsymbol{W}_n$——可学习的权重参数

$\boldsymbol{S}$——上下文信息

同时,模型使用多头注意力机制,获取不同上下文信息 $\boldsymbol{S}_i$。并将不同上下文信息 $\boldsymbol{S}_i$ 与分词工具的输出结合,本文将每个分词工具所得到的词汇信息与字符编码的拼接信息 $\overline{\boldsymbol{H}}$ 表示为

$$\overline{\boldsymbol{H}} = \text{Concat}(\boldsymbol{S}_1, \boldsymbol{S}_2, \cdots, \boldsymbol{S}_i, \cdots, \boldsymbol{S}_k)\boldsymbol{W}_0 \tag{8}$$

式中 k——注意力矩阵个数

$\boldsymbol{W}_0$——可学习的权重参数

Concat——矩阵合并函数

由于分词工具的分词粒度、分割误差和不同知识,所得结果会有一定的误差。因此本文采用多种分词工具获取上下文信息后,再整合所有信息,则对多源分词工具整合后的最终编码输出为

$$\widetilde{\boldsymbol{H}} = \sum_{m=1}^{M} \tanh(\overline{\boldsymbol{H}}_m \boldsymbol{W}_g) \tag{9}$$

式中 $\boldsymbol{W}_g$——可学习的权重参数

$\overline{\boldsymbol{H}}_m$——第 m 个分词工具所得词汇编码信息

$\widetilde{\boldsymbol{H}}$——字符的最终编码输出向量

由于不同的分词工具存在不同的识别粒度,会使分词结果产生不同的语义,本文使用了 6 种常用的中文分词工具分别对“咪鲜·杀螟丹可湿性粉剂”、“水稻纹曲病”和“稻瘟病”进行分词,分词结果如表 4 所示。

表 4 6 种常用的中文分词工具分词效果对比

Tab. 4 Comparison of word segmentation effects of six commonly used Chinese word segmentation tools

工具	咪鲜·杀螟丹可湿性粉剂	水稻纹曲病	稻瘟病
jieba	咪鲜/·/杀/螟丹/可湿性/粉剂	水稻/纹曲病	稻瘟病
thulac	咪鲜/·/杀螟丹可湿性/粉剂	水稻/纹曲病	稻瘟/病
LAC	咪鲜·杀螟丹/可湿性粉剂	水稻纹曲病	稻瘟病
HanLP	咪/鲜/·/杀/螟/丹/可湿性/粉剂	水稻/纹/曲/病	稻瘟病
pkuseg	咪鲜·杀螟丹/可湿性/粉剂	水稻/纹曲病	稻瘟病
SnowNLP	咪/鲜/·/杀/螟丹/可湿性/粉/剂	水稻/纹/曲/病	稻/瘟/病

由表4可知,所有的分词工具对存在实体名称嵌套的情况,都不能较好地识别实体名词边界,会导致多源分词信息整合后的词汇向量仍然存在误差。本文提出了一种判断机制优化该模型,在数据文本输入后,将已有的实体标签与分词结果对比,如果分词结果破坏了嵌套子实体的完整性,则舍弃该分词结果。例如,对于包含病害名词“水稻粒瘟病”的文本,如果被划分为“水稻粒”、“瘟病”两部分,则舍弃该结果;如果被划分为“水稻”、“粒瘟病”,则将这两部分组合起来,形成最终结果:“水稻粒瘟病”。

2.3 BILSTM - CRF 层

农作物病虫害文本中常出现长距离依赖关系,仅使用词汇级别的信息识别实体存在一定困难,如“用枯草芽孢杆菌拌棉花种子,对黄萎病有一定的防治效果”中,“枯草芽孢杆菌”是一种生物防菌剂,“黄萎病”为病害名,但“枯草芽孢杆菌”单独出现时,存在特征“菌”,易被判断为病原菌,因此,需要捕捉文本中存在的上下文信息。BILSTM 基于 LSTM 的网络结构,能捕捉句子级别的信息,它主要包含输入门 i_t、遗忘门 f_t、细胞状态 c_t 和输出门 o_t 共 4 部分。LSTM 的计算过程为

$$i_t = \text{sigmoid}(\boldsymbol{w}_i[h_{t-1}, x_t] + \boldsymbol{b}_i) \tag{10}$$

$$f_t = \text{sigmoid}(\boldsymbol{w}_f[h_{t-1}, x_t] + \boldsymbol{b}_f) \tag{11}$$

$$\tilde{c}_t = \tanh(\boldsymbol{w}_c[h_{t-1}, x_t] + \boldsymbol{b}_c) \tag{12}$$

$$c_t = f_t \otimes c_{t-1} + i_t \otimes \tilde{c}_t \tag{13}$$

$$o_t = \text{sigmoid}(\boldsymbol{w}_o[h_{t-1}, x_t] + \boldsymbol{b}_o) \tag{14}$$

$$h_t = o_t \otimes \tanh(c_t) \tag{15}$$

式中 $\boldsymbol{w}$——运算权重 $\boldsymbol{b}$——运算偏差

x_t——观测预料序列

$\tilde{c}_t$——输入的中间状态

$\boldsymbol{w}_i$——输入门计算对应的权重矩阵

$\boldsymbol{w}_f$——遗忘门计算对应的权重矩阵

$\boldsymbol{w}_c$——中间状态计算对应的权重矩阵

h_t——t 时刻的输出结果

但 BILSTM 的输出是独立的,这会导致非法标注结果的出现,例如标签{O,B,O,I,I,O}中,标记“I”只能出现在“B”或者“I”之后,因此在 BILSTM 网络层后接入一个 CRF 层,作为模型的最终解码器。CRF 将序列标注看成是一个 k^n 分类的问题。即计算条件概率。

$$P(y_1, y_2, \cdots, y_n | x_1, x_2, \cdots, x_n) = P(y_1, y_2, \cdots, y_n | X) \tag{16}$$

式中 y——需要的语料标注序列

P——条件概率

X——输入的文本序列,$X = \{x_1, x_2, \cdots, x_n\}$

为了计算条件概率,CRF 假设:该分布为指数族分布;输出之间的关联仅发生在相邻的位置,且关联具有指数相加性。由此可得

$$P(Y|X) = \frac{1}{Z(x)} \exp\left(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} u_l s_l(y_i, x, i) \right) \tag{17}$$

其中 $$Z(x) = \sum_{y} \exp\left(\sum_{i,k} \lambda_k t_k(y_{i-1}, y_i, x, i) + \sum_{i,l} u_l s_l(y_i, x, i) \right)$$

式中　t_k——对应 x 标记位于第 k 时刻的特征函数

s_l——对应 x 标记位于 l 的特征函数

λ_k——特征函数 t_k 对应的权重

u_l——特征函数 s_l 对应的权重

在解码过程中，CRF 计算所有可能标签序列的总得分，并采用 Viterbi 算法搜索所有输出的标签序列，将具有最高预测总得分的序列作为最终输出，得到最佳标记结果。

3　实验

3.1　实验过程

实验过程分为数据采集、模型构建和测试 3 个阶段，具体流程如图 6 所示。在数据采集阶段，本文收集了大量农作物病虫害相关的摘要文本，并进行了文本处理（包括数据去重、空格删除和编码转换），利用基于半远程监督的停等算法构建了农作物病虫害命名实体识别数据集。模型构建阶段分为 3 部分：网络结构分析，参数调整，模型训练。测试评估阶段，本文分别在已有的公开数据集（MSRA，Weibo）上进行了实验，以验证该方法的泛化能力和稳定性。

所有实验基于 Tensorflow 框架进行，服务器平台配置为：Intel（R）Core（TM）i3 - 3220 CPU，3.30 GHz 处理器，16 GB 运行内存，硬盘容量为 128 GB + 1 TB，使用 GeForce GTX 1080 Ti GPU，内存 11 GB。本文采用 3 种在自然语言处理领域通用的评价指标对模型的性能进行评估，即准确率（Precision，P）、召回率（Recall，R）、F1 值（F1 score）[23]。

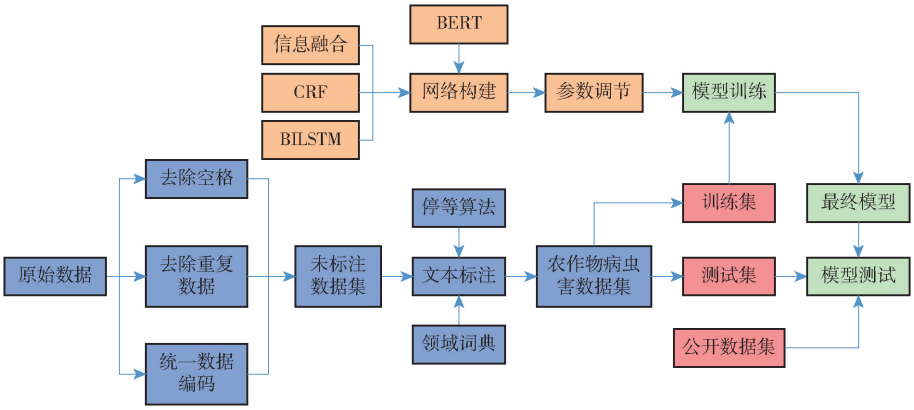

图 6　试验过程

Fig. 6　Experimental process

MSRA 是一个命名实体识别研究人员广泛使用的大型数据集合，是 SIGHAN Bakeoff 2006 使用的中文命名实体识别数据集，MSRA 与 Weibo 数据集均包括 PER、ORG、LOC 共 3 个类别的实体。MSRA 和 Weibo 的统计信息如表 5 所示。

表 5　MSRA 和 Weibo 数据集划分统计

Tab. 5　Statistics of MSRA and Weibo datasets　　条

数据集	训练集		验证集		测试集	
	句子	实体	句子	实体	句子	实体
MSRA	50 700	74 700			4 400	6 200
Weibo	11 700	29 500	4 500	6 000	4 600	7 300

3.2　实验参数

本文对以句号结尾的最长句子计算其长度，最长句子长度为 217 个汉字，由中文语言学提出的“句号表示一句话的结束，新一句话的开始”的观点，设置最大句长为 256。Adaptive moment estimation（Adam）[24]是一种结合了 Adagrad[25]善于处理稀疏梯度和 RMSprop[26]善于处理非平稳目标的优点的自适应算法，能够加快网络收敛速度，本文采用 BERT 模型自带的 Adam Weight Decay Optimizer 为优化器，将 β_1 设置为 0.9，β_2 设置为 0.999 进行模型训练[24]，分词工具采用 jieba、thulac、pkuseg。同时，在经过参数调节实验后，本文将学习率设置为 0.5，dropout 设置为 0.000 05。

3.3　实验结果

3.3.1　Agr - IE 模型与常用模型的比较

本文首先在自建的农作物病虫害数据集上进行实验，以验证 Agr - IE 模型的合理性。将 Agr - IE 与目前常用模型 BILSTM - CRF、BILSTM - Softmax、IDCNN - CRF、BERT - BILSTM - CRF、BERT 对比实验，实验结果如表 6 所示，参与实验的模型均可有效地识别病害、虫害、药剂和作物 4 类实体。召回率和准确率几乎都在 90% 以上，但所有对比模型对病原实体的识别均存在一定困难，准确率和召回率多数在 70% ~80% 之间，不符合应用要求。Agr - IE 对病原的识别效果明显高于其他模型，F1 值为 81.48%，因此，Agr - IE 对农作物病虫害的适用性优于其他模型。

3.3.2　Agr - IE 模型的泛化性和稳定性实验

为了验证 Agr - IE 的泛化性和稳定性，分别对

表 6 农作物病虫害识别效果对比

Tab. 6 Comparison of identification effect of crop diseases and insect pests %

模型	评价指标	病害	虫害	病原	药剂	作物
BILSTM - CRF	P	94.87	94.72	75.50	91.71	91.61
	R	95.38	92.57	70.15	91.14	91.57
	F1 值	95.12	93.64	72.73	91.42	91.59
BILSTM - Softmax	P	94.44	90.82	57.28	83.45	91.29
	R	95.36	90.14	69.03	88.77	92.67
	F1 值	94.90	90.48	62.61	86.03	91.98
IDCNN - CRF	P	94.12	92.70	75.41	73.01	91.28
	R	94.40	93.71	68.66	73.50	88.90
	F1 值	94.26	93.20	71.88	73.25	90.07
BERT	P	94.13	91.32	72.15	91.08	90.20
	R	94.31	91.23	80.22	92.89	91.12
	F1 值	94.22	91.28	75.97	91.97	90.66
BERT - BILSTM - CRF	P	96.14	92.62	77.39	92.62	93.80
	R	96.78	93.81	81.72	94.78	95.14
	F1 值	96.46	93.21	79.49	93.69	94.46
Agr - IE	P	96.45	95.23	80.88	94.33	95.27
	R	96.67	95.00	82.09	94.64	95.81
	F1 值	96.56	95.12	81.48	94.48	95.54

Agr - IE 与 BILSTM - CRF、BERT - IDCNN - CRF、Lattice - LSTM - CRF、CAN - NER、BERT - BILSTM - CRF 等主流模型在 MSRA、Weibo 两个公开数据集上进行命名实体识别实验,实验结果如表 7 所示。

表 7 各模型在公开数据集上识别效果对比

Tab. 7 Comparison of recognition effect of each model on public datasets %

模型	MSRA			Weibo		
	P	R	F1 值	P	R	F1 值
BILSTM - CRF	91.37	90.62	90.99	80.25	84.43	82.29
BERT - IDCNN - CRF	94.57	93.88	94.23	94.79	89.95	92.30
Lattice - LSTM - CRF	93.82	92.67	93.24	83.56	86.09	84.81
CAN - NER	93.42	92.35	92.88	82.86	84.15	83.50
BERT - BILSTM - CRF	95.43	95.19	95.31	92.07	93.11	92.59
Agr - IE	96.20	95.41	95.80	95.14	94.01	94.57

4 讨论

4.1 农作物病虫害数据集上各类模型实体识别效果分析

在农作物病虫害数据集上,本文分别对 Agr - IE 与 BILSTM - CRF、BILSTM - Softmax、IDCNN - CRF、BERT - BILSTM - CRF、BERT 等模型进行了实体识别能力判断实验,用以说明本文所提模型的合理性。实验结果如表 6 所示。

相比较于 IDCNN - CRF,BILSTM - CRF 在农作物病虫害数据集上的识别效果更好。这是因为 IDCNN 虽然能够通过在卷积核之间增加 0,使得 IDCNN 的卷积核在不需要进行 pooling 运算的情况下扩大信息的捕捉范围,但是 IDCNN 也会因此丢失局部信息,同时,IDCNN 有时捕获到的远距离信息没有相关性。

BILSTM - CRF 对各类实体的识别效果比 BILSTM - Softmax 优。这是因为 CRF 能较好地捕捉标签之间的依赖性,例如,本文在对部分错误标注的文本分析后发现,经由 Softmax 层标注的文本有时会出现“I”标签或者“E”标签单独出现的情况,但在 BIEOS 体系里,“I”标签只能跟在“B”标签之后,“E”标签只能出现在“I”标签或者“B”标签之后。

带有 BILSTM - CRF 层的模型能捕捉更多的信息。相对于 BERT 单独使用,BERT - BILSTM - CRF 在病害、虫害、病原、药剂、作物 5 类实体上的 F1 值分别提升了 2.24、1.93、3.52、1.72、3.8 个百分点。这是因为 BERT 在加入 BILSTM - CRF 层后,对上下文信息的捕捉能力有所增强。

相对于 BERT - BILSTM - CRF,利用随机初始化词向量参数的 BILSTM - CRF 对实体的识别效果较差,这是因为 BERT 在预训练阶段能通过自带的 Transformer 使训练结果携带丰富的语义信息,在微调训练过程中,BERT 结合了当前数据中存在的上下文信息,能更好地对词向量进行表示。

综合表 6 结果来看,本文提出的 Agr - IE 优于所有目前主流模型,例如,相对于其他 5 个模型中识别效果最好的 BERT - BILSTM - CRF,Agr - IE 在病害、虫害、病原、药剂、作物 5 类实体上的 F1 值分别提升了 0.1、1.91、1.99、0.79、1.08 个百分点。这是因为本文提出的全局词汇嵌入方式能使字符向量包含部分当前词汇语境下的意思,同时,分词信息能强化模型对于词汇边界的捕捉能力,使模型对实体名词识别的效果更好。

最后,为了验证优化表示层使用的两种策略的合理性,本文在 BERT - BILSTM - CRF 上对所使用的两种嵌入方式进行了实验,实验结果如图 7 所示。PAR(Participle)表示利用分词信息,EMB(Embedding)表示利用全局词汇嵌入信息。利用分词信息以及利用全局词汇嵌入信息进行词嵌入在 5 类实体的识别效果上基本持平,在部分实体的识别效果上略高于 BERT - BILSTM - CRF,但都低于 Agr - IE,这是因为单独利用分词信息,增强了模型对实体边界的识别能力,但未增强向量对词汇信息的携带能力;而单独利用全局词汇嵌入信息,虽然使模型对词汇信息的使用能力增强,但对实体边界的识别能力没有提升,易对词汇边界识别出错,因此将

两种方式结合的 Agr－IE 对 5 类实体的识别效果更好。

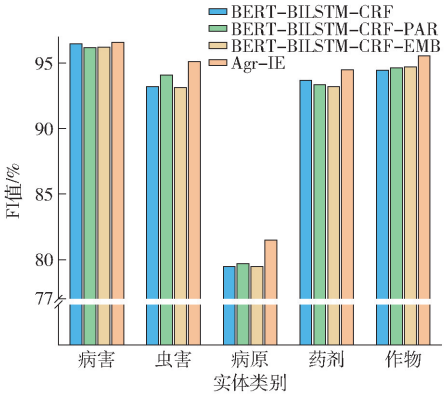

图 7　不同嵌入方式的性能

Fig. 7　Performance of different embedding methods

同时，文献[8]使用 CRF 方法对农作物、病虫害、农药等实体识别的准确率分别达 97.72%、87.63% 和 98.05%，然而这样的方式需要对每个输入的样本分析其复杂的特征，这就使得该方法很难实现农作物病虫害命名实体的自动识别。因此，与传统方法相比，Agr－IE 更适用于对具有不同特征的农作物病虫害相关文本数据识别命名实体。综上所述，本文所提出的 Agr－IE 作为农作物病虫害领域的识别模型具有较好的表现。

4.2　对比模型在中文农作物病虫害数据集上存在的问题

如表 6 所示，不同模型在病害、虫害、药剂和作物等 4 类实体识别上，都显示出了较好的效果，但对病原的识别效果较差。识别效果最差的模型为 BILSTM－Softmax，F1 值为 62.61%。这是因为含有病原类实体的文本较少，且存在一些生物防菌、实验菌株的干扰，例如：从实验生成的标注文件里，本文发现部分模型将“立枯丝核菌”等部分菌类识别为生防菌剂，如表 8 所示。在一些关于生物菌剂防治的论文摘要里，会出现大量的实验代号，例如在一篇文献中，“菌株 QYZ1”为一种未被确认的病原菌菌株代号，但是由于上下文的关系，被直接归类为病原菌。同时，测试集中存在部分训练集中未出现过的实体名称和未出现过的上下文信息模式，在训练时，模型对这些词汇和上下文信息的识别能力较差，导致了实体识别效果差。在中文命名实体识别中，每个字符都不是单独存在的，字符只有在特定的语境下才有具体的含义。Agr－IE 通过识别词汇边界，能强化识别实体所处语境的能力，因此，Agr－IE 在数量较少的实体类的识别上具有显著的优势。

表 8　各模型对“立枯丝核菌”的识别结果

Tab. 8　Identification results of different models for *Rhizoctonia solani*

模型	立	枯	丝	核	菌
真实值	B－PATH	I－PATH	I－PATH	I－PATH	I－PATH
BILSTM－CRF	B－DRUG	I－DRUG	I－DRUG	I－DRUG	I－DRUG
BILSTM－Softmax	B－DRUG	I－DRUG	I－DRUG	I－DRUG	I－DRUG
IDCNN－CRF	B－DRUG	I－DRUG	I－DRUG	I－DRUG	I－DRUG
BERT	B－DISE	I－DISE	I－DISE	I－DISE	O
BERT－BILSTM－CRF	B－DRUG	I－DRUG	I－DRUG	I－DRUG	I－DRUG
Agr－IE	B－PATH	I－PATH	I－PATH	I－PATH	I－PATH

4.3　Agr－IE 模型泛化性和稳定性分析

为了验证 Agr－IE 模型泛化性和稳定性，本文分别在 MSRA 和 Weibo 两个公开数据集上开展了实验，实验结果如表 7 所示。本文所提出的 Agr－IE 模型在 2 个公开数据集上表现良好，尤其是在 MSRA 上，F1 值达到了 95.80%，显著高于 BILSTM－CRF，略高于 BERT－IDCNN－CRF、Lattice－LSTM－CRF、CAN－NER，与 BERT－BILSTM－CRF 基本持平，这一趋势与在 Weibo 数据集上的识别结果基本一致。因此，本文认为 Agr－IE 模型具有良好的泛化性和稳定性。

5　结论

(1)提出了一种基于半远程监督的停等算法，解决了中文农作物病虫害语料库缺失和人工标注命名实体工作量大的问题。该算法通过在标注软件的自动识别模型过拟合前停止算法运行的方式，显著提高了软件对未知实体的召回率，使用该算法对中文农作物病虫害命名实体识别数据集进行构建，大幅度减少了标注过程中的人工成本和时间成本。

(2)提出了基于多源分词信息以及全局词汇嵌入信息的 Agr－IE 模型，实验结果表明，Agr－IE 的优化表示层能较好地解决字符嵌入层无法充分利用词汇级别信息的问题，在本文自主构建的中文农作物病虫害命名实体识别语料库上的实验表明，该模型能有效区分病害、虫害、药剂和作物等实体，在识别较为困难的病原实体上，Agr－IE 也表现出了较好的识别效果，准确率达到 80.88%，召回率达到 82.09%，F1 值达到 81.48%。

(3)Agr－IE 模型具有较好的泛化性和稳定性。Agr－IE 模型在 MSRA 和 Weibo 数据集上取得了最佳的识别效果，F1 值分别为 95.80%、94.57%。

参 考 文 献

[1] GUPTA N, SINGH S, ROTH D. Entity linking via joint encoding of types, descriptions, and context[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processin, 2017.

[2] GENG Zhiqiang, CHEN Guofei, HAN Yongming, et al. Semantic relation extraction using sequential and tree-structured LSTM with attention[J]. Information Sciences, 2019, 509: 183 - 192.

[3] JI Bin, LIU R, LI S S. et al. A hybrid approach for named entity recognition in Chinese electronic medical record[J]. BMC Medical Informatics & Decision Making, 2019, 19(Supp.2): 64.

[4] KAUSHIK N, CHATTERJEE N. RENT: regular expression and NLP-based term extraction scheme for agricultural domain[C]// Proceedings of the International Conference on Data Engineering and Communication Technology, 2017.

[5] KAUSHIK N, CHATTERJEE N. Automatic relationship extraction from agricultural text for ontology construction[J]. Information Processing in Agriculture, 2017, 5: 60 - 73.

[6] GANGADHARAN V, GUPTA D. Recognizing named entities in agriculture documents using LDA based topic modelling techniques[J]. Procedia Computer Science, 2020, 171:1337 - 1345.

[7] 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)

[8] 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)

[9] 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)

[10] XIN Y W, JEAN-DAVID R, ETHAN J H. Deep hybrid neural network for named entity recognition: 692392.15[P]. 2019 - 02 - 28.

[11] PENG N Y, MARK D. Improving named entity recognition for Chinese social media with word segmentation representation learning[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL), 2016: 149 - 155.

[12] HUANG Z H, XU W, YU K. Bidirectional LSTM - CRF models for sequence tagging[J]. arXiv preprint arXiv: 2015, 1508.01991.

[13] EMMA S, PATRICK V, DAVID B, 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.

[14] DONG C H, WU H J, ZHANG J J, et al. Multichannel LSTM - CRF for named entity recognition in Chinese social media[C]// China National Conference on Chinese Computational Linguistics International Symposium on Natural Language Processing Based on Naturally Annotated Big Data, 2017: 197 - 208.

[15] PENG N Y, DREDZE M. Named entity recognition for Chinese social media with jointly trained embeddings[C]//Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, 2015: 548 - 554.

[16] HU Y, ZHENG C W. A double adversarial network model for multi-domain and multi-task Chinese named entity recognition[J]. IEICE Transactions on Information and Systems, 2020, 103(7): 1744 - 1752.

[17] ZHANG Y, YANG J. Chinese NER using lattice LSTM[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, 2018: 1554 - 1564.

[18] ZHU Y Y, WANG G X, KARLSSON B F. CAN - NER: convolutional attention network for Chinese named entity recognition[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 3384 - 3393.

[19] 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)

[20] 赵鹏飞, 赵春江, 吴华瑞, 等. 基于注意力机制的农业文本命名实体识别[J/OL]. 农业机械学报, 2021, 52(1): 185 - 192.
ZHAO Pengfei, ZHAO Chunjiang, WU Huarui, et al. Named entity recognition of Chinese agricultural text based on attention

mechanism[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(1): 185 - 192. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? flag = 1&file_no = 20210121&journal_id = jcsam. DOI:10.6041/j.issn.1000-1298.2021.01.021. (in Chinese)

[21] 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/OL]. 农业机械学报, 2020, 51(增刊2): 335 - 343.
GUO Xuchao, TANG Zhan, DIAO Lei, et al. Recognition of Chinese agricultural diseases and pests named entity with joint radical embedding and self-attention mechanism[J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(Supp. 2): 335 - 343. http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx? flag = 1&file_no = 2020s239&journal_id = jcsam. DOI:10.6041/j.issn.1000-1298.2020.S2.039. (in Chinese)

[22] YANG Y J, SHEN X J, WANG Y J. BERT - BiLSTM - CRF for Chinese sensitive vocabulary recognition[J]. Artificial Intelligence Algorithms and Applications, 2020, 1205: 257 - 268.

[23] JOHNSON S, SHEN S, LIU Y. CWPC_BiAtt: character-word-position combined BiLSTM - Attention for Chinese named entity recognition[J]. Information, 2020, 11(1): 45.

[24] ZHOU J, WANG H D, WEI J L, et al. Adaptive moment estimation for polynomial nonlinear equalizer in PAM8-based optical interconnects[J]. Optics Express, 2019, 27(22):32210.

[25] WARD R, WU X X, LEON B. AdaGrad stepsizes: sharp convergence over nonconvex landscapes, from any initialization[J]. arXiV preprint arXiv: 2018, 1806.01811.

[26] ZOU Fangyu, SHEN Li, ZHANG Weizhong, et al. A sufficient condition for convergences of adam and rmsprop[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. IEEE, 2019: 11127 - 11135.

[27] JACOB D, CHANG M W, KENTON L. BERT: pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv: 2018, 1810.04805.

(上接第 252 页)

[30] LIN T, PIOTR D, ROSS G, et al. Feature pyramid networks for object detection[C]// The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017: 2117 - 2125.

[31] HUANG Z, WANG J, FU X, et al. Dense connection and spatial pyramid pooling based YOLO for object detection[J]. Information Sciences, 2020, 522: 241 - 258.

[32] PENG C, MA J. Semantic segmentation using stride spatial pyramid pooling and dual attention decoder[J]. Pattern Recognition, 2020, 107: 107498.

[33] LIU X, XIA T, WANG J, et al. Fully convolutional attention localization networks: efficient attention localization for fine-grained recognition[J]. arXiv: 1603.06765.

[34] SUN M, YUAN Y, DING E. Multi-attention multi-class constraint for fine-grained image recognition[C]//Proceedings of the European Conference on Computer Vision (ECCV), 2018: 805 - 821.

[35] ZHAO B, WU X, FENG J, et al. Diversified visual attention networks for fine-grained object classification[J]. IEEE Transactions on Multimedia, 2017, 19(6): 1245 - 1256.

[36] ZHANG S W, SHANG Y J, WANG L. Plant disease recognition based on plant leaf image[J]. The Journal of Animal and Plant Sciences, 2015, 25(3): 42 - 45.

[37] SAMAJPATI B J, DEGADWALA S D. Hybrid approach for apple fruit diseases detection and classification using random forest classifier[C]//2016 International Conference on Communication and Signal Processing (ICCSP), 2016: 1015 - 1019.

[38] KAUR S, PANDEY S, GOEL S. Semi-automatic leaf disease detection and classification system for soybean culture[J]. IET Image Processing, 2018, 12(6): 1038 - 1048.