基于 BERT 的农作物命名实体识别模型研究

沈子雷 杜永强*
(信阳农林学院信息工程学院 河南 信阳 464000)

摘 要 随着数字农业的快速发展,农作物命名实体识别作为农业领域知识图谱构建的基础,成为一种高效率的农作物研究领域识别方法。由于农作物实体识别呈现结构复杂、实体指称不一致、干扰因素多等特征,严重制约了农作物领域实体识别的性能,提出一种基于预训练语言模型的实体识别模型,使用 BERT 为文本中词进行编码、采用双向 LSTM(Long-Short Term Memory)获取句子中关键词的上下文,采用 CRFs(Conditional Random Fields)捕获词之间的依赖关系,并结合所构建的农作物命名实体识别数据集进行验证。实验证明该模型能够有效对农作物实体进行识别,且性能优于当前已有的实体识别模型。

关键词 命名实体识别 BERT 预训练语言模型 双向 LSTM 农作物

中图分类号 TP391.1 **文献标志码** A DOI:10.3969/j.issn.1000-386x.2024.06.033

BERT-BASED NAMED ENTITY RECOGNITION MODEL FOR AGRICULTURAL DOMAINS

Shen Zilei Du Yongqiang*
(*College of Information Engineering, Xinyang Agriculture and Forestry University, Xinyang 464000, Henan, China*)

Abstract With the rapid development of digital agriculture, crop named entity recognition, as the basis of knowledge graph construction in agriculture, is becoming an efficient crop recognition method. Since crop entity recognition presents complex structure, inconsistent entity designations, and multiple confounding factors, which seriously restrict the performance of entity recognition in crop domain, the paper proposes an entity recognition model based on pre-trained language models. BERT was used to encode words in text, bi-directional LSTM was used to obtain the context of each keyword in a sentence, and CRFs was used to capture the dependencies between words. The model was validated with the constructed crop named entity recognition dataset. The experiments demonstrate that the model can effectively recognize crop entities and outperforms the existing entity recognition models.

Keywords Named entity recognition BERT Pre-trained language models Bi-directional long and short-term memory Crops

0 引 言

随着数字农业的快速发展,农业相关领域的数字化成为当前研究的热点之一。命名实体识别技术[1-2]作为自然语言处理领域的一项核心技术,能够对农业领域文本数据进行自动处理和分析,从而为后续的数据分析、知识图谱的构建提供重要的支持。农业领域命名实体识别是指从包含关键农业信息的文本中识别具有特殊意义的词或短语,在农作物分类、农业技术研究、农业数据分析等领域都具有广泛的应用前景,例如对具体农作物实体进行识别和分类,可以为农业技术人员提供数字化的支持、帮助农业产业化人员获取竞争产品的关键信息、为农业生产人员提供农业播种的技术指导、帮助农产品销售人员提供农产品市场的最新趋势等。

近年来,在通用领域实体识别任务上,基于神经网络的方法受到了广泛关注。主流方法是基于 LSTM-CRF 框架[3]进行命名实体识别任务,其核心是使用长短期记忆(LSTM)来学习字符的隐含表示,条件随机场

收稿日期:2023-05-23。河南省科技攻关项目(222102110189)。**沈子雷**,讲师,主研领域:深度学习,模式识别,自然语言处理。**杜永强**,副教授。

(CRFs)进行联合标记解码。Wu 等[4]提出了基于 CNN-LSTM-CRF 的联合统一框架解决中文实体识别问题,利用现有标注数据自动生成伪标注样本。Lample 等[5]提出了一种新的命名实体识别架构,改进单向的 LSTM 为双向的 LSTM,即 BiLSTM 网络,从而提升学习文本的隐含表示,并使用 CRFs 进行标记解码。Zhu 等[6]提出了使用卷积注意力网络解决命名实体识别问题,其中包含一个字符级卷积神经(Character-level Convolutional Neural Network,CCNN)网络局部自注意力层,一个门控循环单元(Gated Recurrent Unit,GRU)和全局注意力(Global Attention,GA)层,从相邻字符获取句子上下文信息以增强模型对句子语义的理解。Luo 等[7]提出了一个由基于注意机制的 BiLSTM 层和 CRFs 层组成神经模型,用于解决化学领域的实体识别问题。以上方法致力于解决通用领域的实体识别,但在具体领域实体识别应用方面,特别是农作物领域的实体识别方面,相关的研究还不够深入和完善。

具体到面向行业领域的命名实体识别,农业领域的研究有待深化,由于农作物实体分类数据集不足、命名实体结构复杂等,这就使得通用域实体识别方法在农作物领域上失去了优异的识别性能。因此如何结合农作物实体的特点,构建农作物命名实体识别模型,进而为后续相关的数据分析、知识图谱的构建提供支撑,是值得深入研究的问题。

本文的贡献如下:

(1) 构建一个农作物领域的实体识别数据集。针对农作物领域文本数据来源窄、领域实体类别广等问题,构建一个农作物领域实体识别数据集,以便评估该模型在农作物领域实体识别上的性能,该数据集有一共有 5 378 条句子,10 种类别。

(2) 提出一种基于 BERT 的农作物领域命名实体识别模型。该模型由编码层、特征融合层、推理解码层组成。编码层使用 BERT 作为编码器,充分利用预训练语言模型中的先验知识为句子中每个词生成准确的向量表示。特征融合层利用 BiLSTM 网络作为特征提取器编码句子中每个词的上下文语义信息。推理解码层采用 CRFs 刻画标签之间的依存关系,对句子中的词进行序列标注,以识别出文本中的农作物实体。

(3) 开展相关实验和对比分析,证明模型在农作物领域实体识别任务的性能,并开展相关的对比研究和对比分析。

1 相关工作

在通用领域实体识别任务上,有几种典型的分类方法,如基于字典、规则、机器学习、深度学习等。基于字典的方法[8]是通过将字典中的术语与目标文本序列中的单词匹配以实现实体的提取。虽然这种方法简单,但农作物领域实体的数量剧增和符号的多样化使实体识别变得困难。基于规则的方法[9]是对一个特定领域的实体进行识别,往往表现出高性能但不具有泛化性。基于机器学习的方法[10]是使用各种算法和统计模型高效地执行实体识别。深入分析可知,基于规则和基于机器学习的方法都高度依赖于特征工程,这不仅费时费力,而且需要大量的领域知识。基于深度学习的方法[11-13]使用神经网络能自动提取特征,不仅解决了人工构建特征的问题,并且通过不同的神经网络架构[14-15]构建特定的实体识别数据集,并在识别效果上达到最佳性能。

在深度学习模型和方法的研究实践中,卷积神经网络作为最常用的模型之一,广泛地用于捕捉相关文本的上下文局部信息识别中。随着深度学习模型的发展,混合模型也已经发展起来,通过将卷积神经网络与条件随机场(CRF)算法结合起来,提高预测准确性。在这些混合模型中,BiLSTM 通常用于处理时序数据。在 BiLSTM 模型的嵌入层中,不仅能使用单词级别的嵌入向量,还可使用字符级别的嵌入向量作为输入,以处理词汇表中不存在的未知单词(Out-Of-Vocabulary,OOV)。许多研究者也从不同应用领域提出了不同类型的基于 CNN 和 BiLSTM 的模型,用于提取有意义的字符级别的嵌入。

作为一种用于语言表征的预训练模型,BERT 能够生成深度的双向语言表征,成为当前命名实体识别的研究热点[17],其能将不同类型实体的数据集被训练在同一个预训练语言模型上,可以利用从相关任务中获得的信息改进性能。在生物医药领域,研究者在多任务学习框架中动态更新预训练语言模型的权重,解决生物医药医疗领域跨类别实体识别[18-20]。在农业领域应用中,王春雨[21]采用 CRFs 识别农业领域实体但未能充分利用预训练语言模型中的先验知识。刘晓俊[22]采用 BiLSTM 和 CRFs 的架构作为基准模型,对其进行优化改进,提出了一种基于稠密连接的深层 BiLSTM 模型从农业文本中识别实体。

2 数据集构建

2.1 数据来源及标注

与通用领域数据相比,农作物领域数据存在独有特征,比如类型多、结构复杂等,这要求模型的构建能

充分反映农作物领域实体特征的中文命名实体识别数据集。为保证数据集的质量和可靠性,需要通过以下步骤完成农作物领域实体识别数据集构建:

Step1　收集农作物领域内的文本语料库。从农业相关网站收集了大量的农作物文本,以尽可能涵盖农作物领域数据的特征。

Step2　筛选具有代表性的文本。在收集的文本中,参照"内容清晰、表述准确、结构完整"的准则过滤信息不完整的文本。此外,还利用 jieba 对所爬取的数据进行去空格、空行、停用词及特殊符号处理以保证标注实体的完整性和可靠性。

Step3　标注农作物领域命名实体。首先预定义十种实体类别,分别是 CER、VEG、FRU、OIL、SUG、LEG、FIB、TIM、IND、TEA。其中:CER 表示粮食作物;VEG 表示蔬菜作物;FRU 表示水果作物;OIL 表示油料作物;SUG 表示糖料作物;LEG 表示豆类作物;FIB 表示纤维作物;TIM 表示树材作物;IND 表示产业作物;TEA 表示茶品作物。采用"BIE0"的标注方式进行实体标注,"B"标注实体的开始位置,"I"标注实体的内部位置,"E"标注实体的结束位置,"0"标注文本中非实体部分。各种实体类别的标注策略如表 1 所示,具体文本标注样例如表 2 所示。

表 1　农作物领域实体标注策略

实体类型	实体开始标注	实体中间标注	实体结束标注
粮食	B-CER	I-CER	E-CER
蔬菜	B-VEG	I-VEG	E-VEG
水果	B-FRU	I-FRU	E-FRU
油料	B-OIL	I-OIL	E-OIL
糖料	B-SUG	I-SUG	E-SUG
豆类	B-LEG	I-LEG	E-LEG
纤维	B-FIB	I-FIB	E-FIB
树材	B-TIM	I-TIM	E-TIM
产业	B-IND	I-IND	E-IND
茶品	B-TEA	I-TEA	E-TEA
非实体类别	O	O	O

表 2　文本序列标注样例

文本序列	转	基	因	马	铃	薯
实体标注	B-VEG	I-VEG	I-VEG	I-VEG	I-VEG	E-VEG
实体类型	VEG					

Step4　进行数据清洗和验证。为提升数据集的质量,需要进行数据清洗和验证工作,删除了重复、错误、不规范等标注错误的实体,以保证数据集的准确性和可靠性。

2.2　数据特点分析

与通用实体识别数据集相比,农作物领域的实体名称在文本结构和领域术语等方面有明显自身特点,主要体现在以下三个方面:

(1) 实体结构复杂。农作物领域的实体除了由纯文本组成外,还存在由特殊符号、数字、特殊名称等元素组成,比如"High-affinity K + transporter 1 基因"。

(2) 实体名称长。农作物领域实体常出现长序列实体,如"苏云金杆菌基因的土豆品种",这容易导致模型难以捕获实体的始末位置,加剧了领域实体识别的难度。

(3) 实体中存在干扰名称。在农作物领域实体中常出现由干扰名称组成的实体,比如"淀粉支链酶 IIb 基因"中存在"淀粉",这容易使模型陷入局部语义,降低模型的识别性能。

3　模型构建

针对农作物实体识别的具体要求,提出基于预训练语言模型的实体识别模型,该模型由三部分组成,分别是编码层、特征融合层和推理解码层。编码层采用 BERT 增强文本上下文的语义表征。特征融合层使用 BiLSTM 对上下文信息进行有效编码词,为句子中每个词产生一个融合上下文信息的表示。推理解码层利用 CRFs 可以刻画句子中每个词之间的依存关系,以便能更准确地定位实体的边界。模型的架构如图 1 所示。

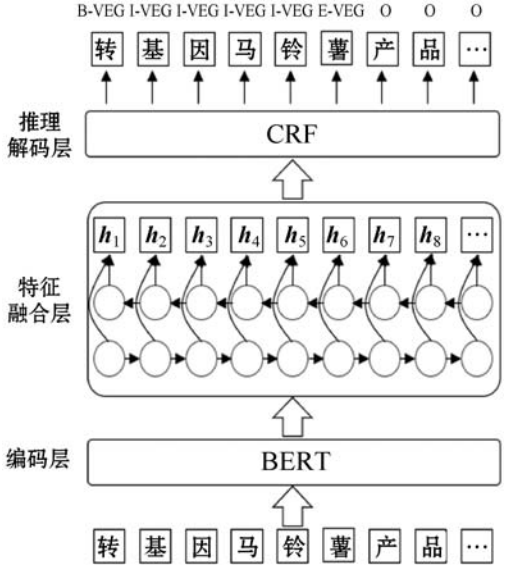


图 1　模型架构

基于该模型的农作物实体识别具体过程如下:

(1) 通过 BERT 预训练语言模型获得输入句子的语义表达,即句子中的每个词映射到一个向量空间。

(2) 利用 BiLSTM 在文本序列的向量空间中进行特征融合,以获得词级特征,即提取句子中每个词的上下文本表示。

(3) 采用 CRF 刻画标签间的依赖关系,将标签序列中输出最大概率的标签作为预测结果,即通过标签序列间依存推断句子中实体。

(4) 重复步骤(1)至步骤(3),直至所有数据都被预测。

3.1 编码层

编码层主要目标是将文本中关键词表示成向量。为解决传统静态编码模型存在的多义词无法辨识等问题,在该模型中使用 BERT 进行编码工作,使用通用的词向量策略,处理多项自然语言处理任务,同时考虑文本中上下文的左右两侧信息,能够有效地提升自然语言的语义理解能力。

BERT 以 Transformer 的编码器结构作为基础,通过在大量未标注数据上进行预训练,来生成深层次的双向语言表征。BERT 的预训练过程包括两个阶段:第一阶段是基于掩码语言模型(Masked Language Model,MLM)的预训练,第二阶段是基于下一句预测(Next Sentence Prediction,NSP)的预训练。在预训练完成之后,BERT 可以通过添加一个额外的输出层进行微调,以适应各种下游自然语言处理任务,而无须对其进行任何特定任务的结构修改。

在潜在语义理解方面,利用 BERT 编码文本序列,即给定由 n 个词组成的句子 $W=(\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_n)$,$\boldsymbol{w}_t$ 表示句子中第 t 个词。利用 BERT 将句子中词 $\boldsymbol{w}_t$ 转化相应的向量 $\boldsymbol{x}_t$,如式(1)所示。

$$\boldsymbol{x}_t=\mathrm{BERT}(\boldsymbol{w}_t)\quad t\in[1,n] \tag{1}$$

3.2 特征融合

在特征融合方面,采用 BiLSTM 来进行文本序列的处理,BiLSTM 神经网络架构由两个单向 LSTM 组成,一个正向输入序列,另一个反向输入序列。BiLSTM 能够有效地增加网络的信息量,提高算法的上下文感知能力,在识别实体时,BiLSTM 能够捕捉到一个单词前后文本的信息,提高识别实体的准确性。

模型的输入是文本序列的词嵌入,模型可以捕捉到词语在句子中前后顺序的依赖关系,并且能够学习到“从前向后”和“从后向前”的信息,从而对于实体边界的识别具有较好的效果。

通过 BiLSTM 获取全面的上下文信息并学习上下文之间的依赖关系,双向 LSTM 将前向和后向 LSTM 网络应用于每个训练序列,并将这个双向 LSTM 网络连接到输出层,形成特定的网络结构。

LSTM 计算单元包含输入门、遗忘门和输出门,计算单位的具体计算过程如式(2)到式(7)所示。

$$\boldsymbol{i}_t=\sigma(\boldsymbol{w}_i[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_i) \tag{2}$$

$$\boldsymbol{f}_t=\sigma(\boldsymbol{w}_f[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_f) \tag{3}$$

$$\boldsymbol{o}_t=\sigma(\boldsymbol{w}_o[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_o) \tag{4}$$

$$\tilde{\boldsymbol{c}}_t=\tanh\boldsymbol{w}_c[\boldsymbol{h}_{t-1},\boldsymbol{x}_t]+\boldsymbol{b}_c \tag{5}$$

$$\boldsymbol{c}_t=\boldsymbol{f}_t\times\boldsymbol{c}_{t-1}+\boldsymbol{i}_t\times\tilde{\boldsymbol{c}}_t \tag{6}$$

$$\boldsymbol{h}_t=\boldsymbol{o}_t\times\tanh(c^t) \tag{7}$$

式中:$\boldsymbol{w}_i$、$\boldsymbol{w}_f$、$\boldsymbol{w}_o$ 为 LSTM 的权重矩阵;$\boldsymbol{b}_i$、$\boldsymbol{b}_f$、$\boldsymbol{b}_o$ 为偏置项;$\boldsymbol{i}_t$ 为输入门、$\boldsymbol{f}_t$ 为遗忘门、$\boldsymbol{o}_t$ 为输出门;$\boldsymbol{x}_t$ 表示时间 t 的输入向量;$\boldsymbol{h}_t$ 表示时间 t 的隐藏状态;$\tilde{\boldsymbol{c}}_t$ 表示细胞状态的候选值,$\boldsymbol{c}_t$ 表示细胞状态。

LSTM 前向计算表示为式(8)。

$$\overrightarrow{\boldsymbol{h}}_t=\mathrm{LSTM}(\boldsymbol{x}_t,\overrightarrow{\boldsymbol{h}}_{t-1}) \tag{8}$$

LSTM 反向计算表示为式(9)。

$$\overleftarrow{\boldsymbol{h}}_t=\mathrm{LSTM}(\boldsymbol{x}_t,\overleftarrow{\boldsymbol{h}}_{t-1}) \tag{9}$$

将 LSTM 前向计算和 LSTM 反向计算连接起来得到 BiLSTM 在时间 t 的输出,即式(10)。

$$\boldsymbol{h}_t=[\overrightarrow{\boldsymbol{h}}_t;\overleftarrow{\boldsymbol{h}}_t] \tag{10}$$

式中:[* ; *]表示拼接操作。

3.3 推理解码

LSTM 表达了文本的上下文信息,但其并没有将标记之间的依赖关系表达清楚,因此,通过引入 CRFs 来学习标记与标记的相邻关系,从而确保文本序列标记的合理性。基于概率的无向图模型,CRFs 提供了自然语言处理中的标注功能,针对农作物领域实体识别建模成序列标注问题,利用 CRFs 对文本序列进行标注。对于文本序列中的每个位置,CRFs 都可以根据该位置的上下文信息来预测该位置所属的实体类型。

具体而言,在 BiLSTM 输出层后添加全连接层以获得含有标签信息的句子特征,即 $p=(p_1,p_2,\cdots,p_n)$。最后,通过 CRFs 预测文本序列相对的标注序列,CRF 层的参数由转移矩阵 $\boldsymbol{S}$ 表示,S_{ij} 表示第 i 个标签转向第 j 个标签的分数。该推理解码模型对于文本序列 W 的标注序列 $Y=(y_1,y_2,\cdots,y_n)$ 的概率为:

$$score(W,Y)=\sum_{i=1}^{n}P_{i,y_i}+\sum_{i=1}^{n+1}S_{y_i,y_{i+1}} \tag{11}$$

式中:$S_{y_i,y_{i+1}}$ 定义为从第 y_i 个标记移动到第 y_{i+1} 个标记的概率;P_{i,y_i} 定义为文本序列中第 i 个词的标签是 y_i 的概率。

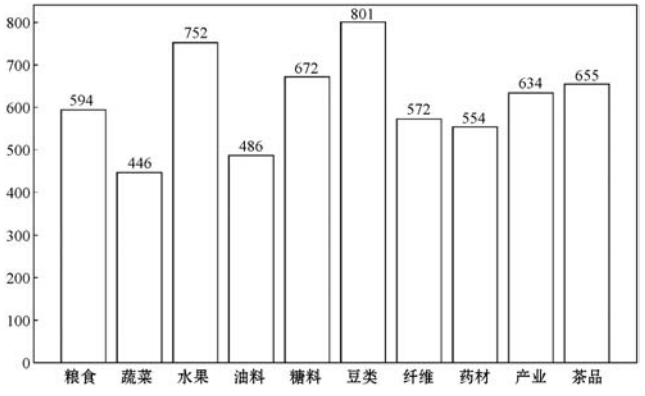
利用 softmax 函数将 score 归一化,如式(12)所示。

$$P(Y\mid W)=\frac{\exp(score(w,y))}{\sum_{y'}\exp(score(w,y'))} \tag{12}$$

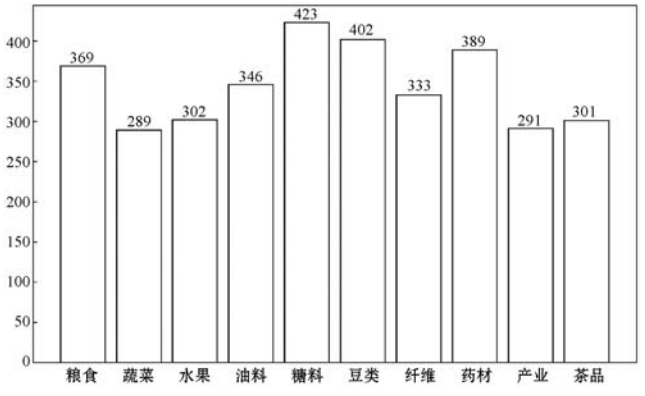
4　实　验

4.1　实验数据集

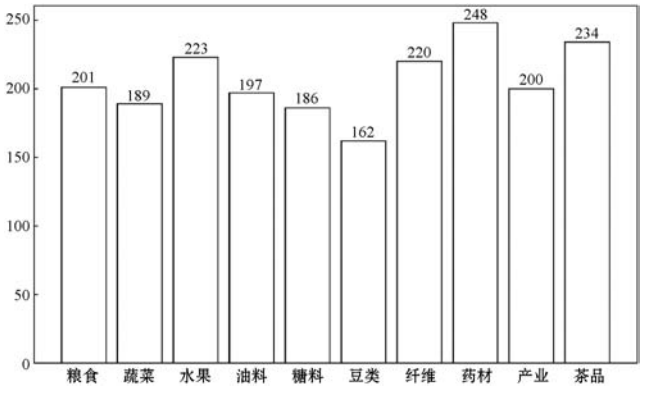
在面向领域的实体识别数据集构建方法研究的基础上,本文结合农业领域相关的文本数据的爬取,构建一个具有精标注的农作物命名实体识别数据集。该数据集包含约 5 378 条句子、11 671 个实体、10 种实体类别。为了验证所提模型的有效性和实用性,本文按照 8∶1∶1的比例将句子划分为训练集、验证集和测试集。为避免数据集存在类别不均衡的问题,本文微调数据集的类别分布,每种实体在划分的数据集上的分布如图 2 所示。



(a) 训练集



(b) 验证集



(c) 测试集

图 2　各数据子集实体类别分布

4.2　实验环境与超参数

实验在 Ubuntu 20.04、显卡型号 NVIDIA GeForce RTX 3090、Python 3.9 版本、PyTorch 1.10.0 版本下进行。在实验过程中,采用的超参数通过在训练集上训练,验证集微调所得的最优参数组合:训练次数为 40,学习率为 10^{-5},丢弃率为 0.45。

超参数微调过程中,发现合适的学习率对模型的收敛很重要,选择较大的学习率,会造成网络收敛慢或不能收敛,过小的学习率容易出现局部最优解,导致模型的性能变差;另一方面,合理的批次大小对模型的训练时间和收敛也很重要,过小的批次会导致训练时间长且会出现收敛慢的问题,过大的批次会导致训练时间短但会因为局部最优解而导致模型性能差。此外,为防止模型过拟合,采用随机失活率来随机剪枝神经元之间的复杂关系,以提供模型鲁棒性,为此实验选用了 adam 作为优化器来优化网络参数。

4.3　评价指标

农作物领域中存在大量的专业术语和复杂结构形式,这就需要合理的评价指标能够反映出算法准确性和覆盖率。

精确度(Precision,P)计算的是模型预测出全部实体中正确实体的比例,反映了模型预测出实体的精准程度;召回率(Recall,R)计算的是所有实体中模型正确预测实体的比例,反映了模型对于正确实体的召回性能。模型召回能力越强,则会有更多的错误实体被预测导致精准率下降:模型的预测精准程度越高,则会有更多的实体没有被预测出来;F1 值(F1-score)平衡了精确率与召回率,能够综合评估模型的实体识别性能,F1 分数越高说明模型的实体识别综合能力越优异。

$$P = \frac{T_P}{T_P + F_P} \tag{13}$$

$$R = \frac{T_P}{T_P + F_N} \tag{14}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{15}$$

5　实验结果与分析

为了充分验证基于 BERT 的农作物领域命名实体识别模型的性能,进一步开展了农作物领域实体识别数据集上的全面实验评估。

5.1　不同模型的实验结果对比

在不同模型的实验结果对比中,选择 7 个实体识

别任务的基线方案进行对比,相关实验的对比结果如表3所示。

表3 不同模型的对比实验(%)

模型	P	R	F1
CNN	70.21	68.87	69.53
CNN-CRF	72.03	69.97	70.98
LSTM	71.39	70.01	70.69
LSTM-CRF	72.89	70.65	71.75
CNN-LSTM	71.78	71.81	71.79
CNN-LSTM-CRF	73.42	72.67	73.04
CNN-BiLSTM-CRF	74.78	73.21	73.99
BERT-BiLSTM-CRF	76.78	77.56	77.17

可以看出,CNN 模型在精确度、召回率、F1 上均取得最差性能,LSTM 模型取得了比 CNN 模型稍好的性能,可能的原因是 CNN 不能记忆上下文语义而 LSTM 能够捕获句子的上下文语义。BERT-BiLSTM-CRF 在三项指标上均达到了最佳性能,这表明该模型具备识别农作物命名实体的能力。进一步分析可知,CRF 与 CNN、LSTM、BERT 组合均有利于实体的识别。

5.2 不同实体类别的对比实验

在不同实体类别的对比实验中,验证了该模型在各类农作物领域实体上的识别性能,各类实体识别的对比分析如表4所示。

表4 不同实体类别的对比实验(%)

模型	实体类别	P	R	F1
BERT-BiLSTM-CRF	粮食	76.02	78.43	77.21
	蔬菜	73.56	72.78	73.17
	水果	79.78	81.02	80.40
	油料	74.12	73.47	73.79
	糖料	77.88	78.32	78.10
	豆类	80.54	82.05	81.29
	纤维	75.43	76.02	75.72
	树材	75.78	74.86	74.81
	产业	77.67	79.49	78.05
	茶品	77.43	78.89	77.64

可以看出,该模型比较擅长识别“豆类作物”和“水果作物”相关的实体,但对“蔬菜作物”和“油料作物”相关实体的识别较差,出现这个问题的主要原因是模型在训练数据中未能充分学习到“蔬菜作物”和“油料作物”相关实体的特征,在训练集中增加相应类别的实体可以弥补这一缺陷。进一步深入的实验分析发现,该模型在训练集中得到充分训练,且在验证集上恰当优化,都能在测试集上取得良好的性能。

5.3 消融实验

为了验证模型的各个模块对农作物领域实体识别性能的影响,对模型进行了消融实验,实验结果如表5所示。

表5 模型的消融实验(%)

模型	P	R	F1
BERT-BiLSTM-CRF	76.78	77.56	77.17
BERT-BiLSTM	75.21	75.68	75.44
BERT-CRF	76.23	77.04	76.63
BERT	75.89	76.21	76.05
BiLSTM-CRF	73.93	75.66	74.78
BiLSTM	73.02	74.88	73.94

可以看出,删除 CRF 模块时,F1 下降了1.73百分点。删除 BiLSTM 模型时,F1 下降了0.54百分点。删除 BERT 模块时,F1 下降了2.39百分点。由此,证明了该模型中的三个模块均有助于农作物领域实体的识别。进一步分析可知,在该模型中,BERT 对农作物领域实体识别的性能提升最大,CRF 次之,BiLSTM 最少,出现这种情况的原因是 BERT 在训练时已经包含 BiLSTM 所学的知识。

6 结 语

数字农业的快速发展带来了农业相关领域文本识别需要的快速增加,命名实体识别技术能够有效解决农业领域相关文本识别中的基础问题,该模型结合编码、特征融合、推理解码等各个阶段需要解决的问题,针对性地提出相关的解决方案,在构建的农作物领域实体识别数据集支持下,能够有效识别文本中的农作物实体。实验结果和分析表明,该模型能够有效识别农作物领域中的实体,并且性能优于当前现有的实体识别模型,为农作物领域命名实体识别提供有效的手段,对促进农业领域文本信息处理、知识图谱的构建等具有较大的价值和意义。

参 考 文 献

[1] Wang Y, Tong H, Zhu Z, et al. Nested named entity recognition: A survey[J]. ACM Transactions on Knowledge Discovery from Data (TKDD),2022,16(6):1-29.

[2] 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,2020,34(1):50-70.

[3] 王红,李浩飞,邸帅. 民航突发事件实体识别方法研究

[J]. 计算机应用与软件,2020,37(3):166 – 172.
[4] Wu F, Liu J, Wu C, et al. Neural Chinese named entity recognition via CNN-LSTM-CRF and joint training with word segmentation[C]//The World Wide Web Conference,2019: 3342 – 3348.
[5] Lample G, Ballesteros M, Subramanian S, et al. Neural architectures for named entity recognition[C]//2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2016:260 – 270.
[6] Zhu Y, Wang G. CAN-NER: Convolutional attention network for Chinese named entity recognition[C]//2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019:3384 – 3393.
[7] Luo L, Yang Z, Yang P, et al. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition[J]. Bioinformatics,2018,34(8):1381 – 1388.
[8] Quimbaya A P, Múnera A S, Rivera R A G, et al. Named entity recognition over electronic health records through a combined dictionary-based approach[J]. Procedia Computer Science, 2016,100:55 – 61.
[9] Fresko M, Rosenfeld B, Feldman R. A hybrid approach to NER by MEMM and manual rules[C]//14th ACM International Conference on Information and Knowledge Management,2005: 361 – 362.
[10] Ekbal A, Bandyopadhyay S. Named entity recognition using support vector machine: A language independent approach [J]. International Journal of Electrical and Computer Engineering, 2010,4(3):589 – 604.
[11] Zhu Q, Li X, Conesa A, et al. GRAM-CNN: A deep learning approach with local context for named entity recognition in biomedical text[J]. Bioinformatics,2018,34(9):1547 – 1554.
[12] Li D, Yan L, Yang J, et al. Dependency syntax guided BERT-BiLSTM-GAM-CRF for Chinese NER[J]. Expert Systems with Applications,2022,196:116682.
[13] Liu J, Chen Y, Xu J. Low-resource NER by data augmentation with prompting[C]//31st International Joint Conference on Artificial Intelligence,2022:4252 – 4258.
[14] Haque M Z, Zaman S, Saurav J R, et al. B-NER: A novel Bangla named entity recognition dataset with largest entities and its baseline evaluation[J]. IEEE Access, 2023, 11: 45194 – 45205.
[15] Frei J, Kramer F. GERNERMED: An open German medical NER model[J]. Software Impacts,2022,11:100212.
[16] Peng D L, Wang Y R, Liu C, et al. TL-NER: A transfer learning model for Chinese named entity recognition[J]. Information Systems Frontiers,2020,22:1291 – 1304.
[17] Giorgi J M, Bader G D. Towards reliable named entity recognition in the biomedical domain[J]. Bioinformatics,2020, 36(1):280 – 286.
[18] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding [C]//2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,2019:4171 – 4186.
[19] Wang X, Zhang Y, Ren X, et al. Cross-type biomedical named entity recognition with deep multi-task learning[J]. Bioinformatics,2019,35(10):1745 – 1752.
[20] Zhao S, Liu T, Zhao S, et al. A neural multi-task learning framework to jointly model medical named entity recognition and normalization[C]//AAAI Conference on Artificial Intelligence,2019,33(1):817 – 824.
[21] 王春雨. 基于 CRF 的农业命名实体识别研究[D]. 保定:河北农业大学,2014.
[22] 刘晓俊. 面向农业领域的命名实体识别研究[D]. 合肥:安徽农业大学,2019.

(上接第 174 页)

[11] 齐榕,贾瑞生,徐志峰,等. 基于 YOLOv3 的轻量级目标检测网络[J]. 计算机应用与软件,2020,37(10):208 – 213.
[12] Sandler M, Howard A, Zhu M L, et al. Mobilenetv2: Inverted residuals and linear bottlenecks[C]//IEEE Conference on Computer Vision and Pattern Recognition, 2018: 4510 – 4520.
[13] Ioffe S, Szegedy C. Batch normalization: Accelerating deep network training by reducing internal covariate shift[C]// 32nd International Conference on Machine Learning, 2015: 448 – 456.
[14] Glorot X, Bordes A, Bengio Y. Deep sparse rectifier neural networks[C]//14th International Conference on Artificial Intelligence and Statistics,2011:315 – 323.
[15] Chieng H, Wahid N, Pauline O, et al. Flatten-T Swish: A Thresholded ReLU-Swish-like activation function for deep learning[J]. International Journal of Advances in Intelligent Informatics,2018,4(2):76 – 86.
[16] Rezatofighi H, Tsoi N, Gwak J Y, et al. Generalized intersection over union: A metric and a loss for bounding box regression[C]//IEEE Conference on Computer Vision and Pattern Recognition,2019:658 – 666.
[17] 邹承明,薛榕刚. 融合 GIoU 和 Focal loss 的 YOLOv3 目标检测算法[J]. 计算机工程与应用,2020,56(24):214 – 222.
[18] Shao Z F, Wu W J, Wang Z Y, et al. SeaShips: A large-scale precisely-annotated dataset for ship detection[J]. IEEE Transactions on Multimedia,2018,20(10):2593 – 2604.
[19] 章永来,周耀鉴. 聚类算法综述[J]. 计算机应用,2019,39(7):1869 – 1882.