doi:10.6041/j.issn.1000-1298.2023.12.026

基于融合对抗训练的农作物品种信息抽取方法

许　鑫　马文政　张　浩　马新明　乔红波
(河南农业大学信息与管理科学学院，郑州 450002)

摘要：针对我国作物品种种类多，资源信息规范性差，模型训练精度低等问题，本文以小麦、水稻、玉米、大豆、棉花、花生、油菜 7 种作物为对象，以品种、形态、产量和品质等参数为指标，构建了 83 个品种实体，采用人工标注方法，通过融合对抗训练技术，提出了农作物品种信息抽取 4 层网络模型(BERT－PGD－BiLSTM－CRF)。模型基于深层双向 Transformer 构建的 BERT(Bidirectional encoder representation from transformers)模型作为预训练模型获取字词语义表示，使用 PGD(Projected gradient descent)对抗训练方法为样本增加扰动，提高模型鲁棒性和泛化性，利用双向长短期记忆网络(Bidirectional long short-term memory，BiLSTM)学习长距离文本信息，结合条件随机场(Conditional random field，CRF)学习标签约束信息。对比 18 个不同信息抽取模型的训练效果，结果表明，本研究提出的 BERT－PGD－BiLSTM－CRF 模型精确率为 95.4%、召回率为 97.0%、F1 值为 96.2%，说明利用对抗训练技术的 BERT－PGD－BiLSTM－CRF 模型能够有效对作物品种信息进行抽取，同时也为农业信息抽取提供了技术参考。

关键词：农作物品种；信息抽取；BERT；对抗训练
中图分类号：TP391.1　　**文献标识码**：A　　**文章编号**：1000-1298(2023)12-0272-08　　**OSID**：

Crop Variety Information Extraction Method Based on Integrated Adversarial Training

XU Xin　MA Wenzheng　ZHANG Hao　MA Xinming　QIAO Hongbo
(*College of Information and Management Science*, *Henan Agricultural University*, *Zhengzhou* 450002, *China*)

Abstract: In response to the issues of a wide variety of crop types, poor resource information standardization, and low model training accuracy in China, focusing on seven crops: wheat, rice, maize, soybeans, cotton, peanuts, and rapeseed, using parameters like variety, morphology, yield, and quality as indicators, totally 83 crop variety entities were constructed. A manual annotation approach was adopted and an information extraction four-layer network model (BERT－PGD－BiLSTM－CRF) was introduced by incorporating adversarial training techniques. The model utilized the bidirectional encoder representation from transformers (BERT) model, based on a deep bidirectional transformer, as a pre-training model to acquire semantic representations of words and phrases. It employed projected gradient descent (PGD) adversarial training to introduce perturbations to the samples, thereby enhancing model robustness and generalization. Additionally, it leveraged a bidirectional long short-term memory (BiLSTM) network to capture long-distance text information and combined conditional random fields (CRF) to learn label constraint information. Comparing the training results with 18 different information extraction models, the research indicated that the proposed BERT－PGD－BiLSTM－CRF model achieved a precision of 95.4%, a recall of 97.0%, and an F1 score of 96.2%. This suggested that the BERT－PGD－BiLSTM－CRF model, utilizing adversarial training techniques, was effective in extracting crop variety information and also provided a technological reference for agricultural information extraction.

Key words: crop variety; information extraction; BERT; adversarial training

收稿日期：2023－08－25　修回日期：2023－09－26
基金项目：河南省科技创新杰出人才项目(184200510008)
作者简介：许鑫(1984—)，男，副教授，博士，主要从事智慧农业与大数据技术研究，E-mail：xuxin468@163.com
通信作者：马新明(1962—)，男，教授，博士，主要从事精准农业与信息技术研究，E-mail：wheatdoctor@163.com

0　引言

种子在植物学上是指由胚珠发育而成的繁殖器官。在农业生产上，种子是最基本的生产资料[1]。种子作为农业的核心要素，扮演着粮食生产的基础角色[2-3]。作物品种是在一定的生态条件和经济条件下，根据人类的需要所选育的某种作物的一定群体[4]。我国作物品种繁多，涉农网站数目众多，但信息异质、异构、分散、重复，形成一个个“信息孤岛”，难以发挥农业信息资源的集成效用[5]。利用现代信息抽取技术，可以快速准确地进行作物品种特性信息抽取，为品种知识图谱的构建以及品种推荐提供助力，对品种推广与利用有着重大意义。

信息抽取专注于从非结构化和半结构化信息中抽取出结构化信息[6]。传统的信息抽取通常由人工进行，通过收集各种纸质材料或书籍来获取有价值且可利用的信息。自 20 世纪 60 年代起，国外学者开始研究信息抽取技术。至 90 年代初期，中文信息抽取研究起步。随着时间的推移，信息抽取技术已成为一项不可或缺的信息技术。信息抽取在智能检索、内容推荐、客户分析、知识构建、辅助决策等多个领域中都扮演着重要角色。

信息抽取包括实体抽取、关系抽取、事件抽取等。目前，信息抽取技术主要基于两种方式：基于规则表达式和基于机器学习或深度学习。基于规则表达式的抽取模型针对性较强，模型优劣和规则的制定有着很大的联系。YAMASHITA 等[7]开发了胰腺囊性病变 PCL 识别模型，模型结果和医生观察结果一致性达到 0.951。HAN 等[8]创建了一系列地震事件信息提取规则，实现了从网络文本中提取地震信息。利用规则进行信息抽取，无法考虑到字词的语义信息，对于复杂表述的内容抽取效果较差。基于机器学习或深度学习的抽取模型可以对语义信息进行分析，从而获得良好的抽取效果。YUAN 等[9]在 BiGRU－Dual Attention 模型中加入单词级注意力机制和句子级注意力机制，从而更好地分配权重，消除噪声，提高实体关系提取的识别精度，在测试集上的 F1 值达到 88.05%。XIE 等[10]开发了微调管道，通过掩蔽语言建模(MLM)进行无监督的领域适配，并测试了 BERT、RoBERTa 和 Bio_ClinicalBERT 3 个预训练的神经网络模型，F1 值超过 80%。TIFTIKCI 等[11]使用 BIOHD 标注方法，集成 BiLSTM 和 SciMiner 规则方法实现了不规则实体识别。PENG 等[12]使用多级神经网络实现了迁移网页信息抽取。MIWA 等[13]提出一种端到端神经网络模型，通过堆叠 BiLSTM 获得更好的效果。然而品种信息繁多，实体种类多且相似，存在描述和单位不统一等表述不规范问题，需要对实体数据进行增强，同时实现实体清洗、对齐等操作，这也是农业信息抽取中常见的问题。

对抗训练技术可以实现数据增强，帮助模型识别字词边界信息，提高模型的鲁棒性和泛化性。很多学者进行了一系列有益探索。ZHENG 等[14]提出一种对抗性门控双线性注意力神经网络 AGBAN，有效利用异构数据源的共性，解决了社交媒体帖子因文本短且缺乏上下文而造成命名实体识别任务困难的问题。WANG 等[15]提出了一种对抗训练的 LSTM－CNN (ASTRAL) 系统，在 3 个基准上进行评估并取得了当时最先进的结果。LAI 等[16]采用对抗训练提高了 ERIGAT 模型的性能和鲁棒性。朱红等[17]提出了 CW－ADV 模型，利用字词融合和对抗训练来解决特色行业人物实体识别中的语义和边界不准确问题。HUANG 等[18]实验发现 BiLSTM 联合 CRF 可以取得比 LSTM、BiLSTM 和 CRF 更好的效果。BERT 出现后，LI 等[19]在 BiLSTM－CRF 模型的基础上加入 BERT 以获得具有丰富语义特征的序列向量。JIANG 等[20]使用 BERT－BiLSTM－CRF 进行中国医学命名实体识别并发现对抗训练的加入有助于提高模型的鲁棒性。说明对抗训练技术可以提升自然语言处理模型性能。

品种信息实体种类多，且存在不同的语言表述方式以及不同实体距离近、形式类似等问题。因此，为了研究农业信息抽取方法，提高农业信息的共享和利用率，本文以作物品种信息为研究对象，利用爬虫技术获取品种数据，设计包含预训练层、对抗训练层、语义特征提取层及条件约束层 4 层网络的神经网络模型，进行作物品种信息抽取，以期为品种推荐以及农业信息抽取提供技术支持。

1　研究方法与技术

1.1　技术路线

本研究的技术路线如图 1 所示，选取网页为数据来源通过爬虫技术批量获取品种数据，数据在经过标注后划分数据集。以数据集作为序列输入，词嵌入编码后使用预训练的 BERT 模型获取语义表示，通过 PGD 对抗训练计算样本扰动，使用正确的样本和扰动样本训练 BiLSTM 模型，并将模型输出作为 CRF 的输入，CRF 输出最终预测结果。

1.2　数据获取

本文通过爬虫、XPath 以及正则表达式等技术，以“第 1 种业”网(http://www.a-seed.cn/)作为数据来源，以小麦、水稻、大豆、玉米、棉花、花生、油菜

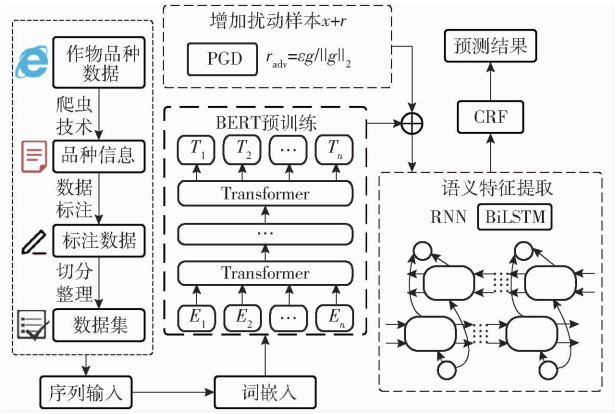

图 1 技术路线图

Fig. 1 Technology roadmap

7 种作物为对象,选取小麦、水稻、玉米各 200 篇文档,大豆、棉花、花生、油菜各 100 篇文档,构建农作物品种数据集。

1.3 实体设计

该研究致力于信息抽取工作,目的是将文本转换为结构化表格,因此只定义实体,将其作为属性值进行抽取。参照中国作物种质信息网的农作物品种资源信息处理规范,定义 83 个实体,并按照实体的大致含义将其分为 4 类,分类及名称见表 1。

1.4 数据标注

实体标注采取人工 BIO 标注方式,B 代表实体的起始位置,I 代表实体的中部及结束位置,O 代表非实体文本,如“豫 B-name/农 I-name/8 I-name/1 I-name/0 I-name/是 O/半 B-type/冬 I-type/性 I-type/品 O/种 O”。对于重复的实体,按照统一标准选取其中一个实体标注,最终共标注 22 332 个实体。标注完成后,将 1 000 篇文档进行切分,得到 7 524 条语料。

表 1 实体分类及名称

Tab. 1 Entity classification and name

实体分类	实体名称	实体数量
品种信息	作物种类、品种名称、品种类型、选育单位、品种来源、编号、年份、地域	8
形态特征	生育期、株高、穗长、粒长、长宽比、穗位高、成株叶片数、穗行数、底荚高度、主茎节数、分枝数、第一果枝节位、侧枝长、分枝部位高	14
产量指标	亩产、千粒重、亩穗数、穗粒数、单株荚数、单株粒数、百果重、百仁重、单株有效角果数、单株果枝、单株结铃、角粒数、百粒重、皮棉亩产、单铃重、结果枝、单株结果	17
参数指标	直链淀粉含量、粗蛋白含量、含油量、茎蔓粗蛋白含量、面粉湿面筋含量、硫苷含量、吸水率、最大拉伸阻力、籽粒容重、硬度指数、粗淀粉含量、品质、拉伸面积、沉淀值、胶稠度、碱消值、赖氨酸含量、透明度、断裂比强度、粗脂肪含量、形成时间、稳定时间、子指、衣分、霜前花率、纤维上半部长度、油酸含量、芥酸含量、马克隆值、结实率、出糙率、纺纱均匀性指数、籽仁亚油酸含量、精米率、整精米率、整齐度指数、伸长率、出籽率、饱果率、出仁率、反射率、垩白度、垩白粒率、黄度	44

1.5 品种信息抽取模型设计

作物品种信息抽取任务中,存在实体位置相近或实体值相似的问题,并且实体描述存在多种形式,对模型获取上下文语义的效果提出了要求。对此,本研究采用 BERT - PGD - BiLSTM - CRF 模型来实现作物品种信息抽取任务。如图 2 所示,该模型由输入层、BERT 预训练层、PGD 对抗训练层、BiLSTM 语义特征提取层和 CRF 条件约束层组成。输入层接收序列数据,BERT 层作为预训练模型,用于提取文本的上下文表示。PGD 层通过引入微小扰动来增加模型的鲁棒性。BiLSTM 层则用于学习文本的长序列语义特征,能够有效捕捉上下文信息。最后,CRF 层用于学习标签之间的约束规则,确保生成的标签序列具有合理的转移概率。

1.5.1 预训练层

BERT 由 Google AI 研究院提出,采用深层双向

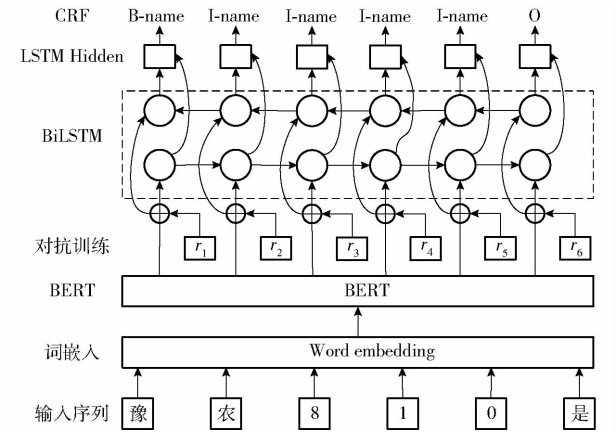

图 2 BERT - PGD - BiLSTM - CRF 模型结构

Fig. 2 BERT - PGD - BiLSTM - CRF model structure

Transformer 构建模型并利用 MLM(Masked language model)预训练。BERT 是一个非常庞大的预训练语言表征模型,其中的 Attention 机制可以获得更丰富的字级语义表示。本文使用 BERT 作为预训练模型,同时对 Fine - tuning 效果进行测试。

1.5.2 对抗训练层

对抗训练(Adversarial training)是一种在 NLP 领域中常用的防御机制,旨在提高模型鲁棒性和泛化性,可以获得比单独使用 dropout 更好的效果[21]。

其基本原理是通过向样本中添加微小但可能导致误分类的扰动来生成对抗样本，并将其纳入模型的训练过程中。这样可以使模型在面对具有扰动的输入时更加稳健，并能更好地推广到未处理过的数据上。

对抗训练的一般原理可以用 Min－Max 公式[22]来说明。其中，最小化（Minimize）部分代表了模型的目标，即尽可能减小正常样本的损失；最大化（Maximize）部分代表了对抗样本的损失，即尽可能增大对抗样本的误分类程度。通过不断迭代优化这个公式，模型可以逐步学习到对抗样本的特征，并提高对扰动的抵抗能力。

Min－Max 计算式为

$$\begin{cases}\min\limits_{\theta}\rho(\theta)\\ \rho(\theta)=\mathbb{E}_{(x,y)\sim D}\left[\max\limits_{\delta\in S}L(\theta,x+\delta,y)\right]\end{cases}\tag{1}$$

式中　$L(\theta,x,y)$——损失函数

θ——模型参数集

$\rho(\theta)$——模型风险

$\mathbb{E}_{(x,y)\sim D}$——平均损失

δ——允许扰动

S——允许扰动集

x——输入　　y——标签

计算扰动时，对抗思想为往增大损失的方向增加扰动[23]

$$r_{\mathrm{adv}}=\underset{r,\|r\|\leqslant\varepsilon}{\operatorname{argmin}}\ \mathrm{lb}p(y|x+r;\hat{\theta})\tag{2}$$

式中　r_{adv}——扰动　　r——输入扰动

ε——扰动参数

$\hat{\theta}$——分类器参数常量

实验选取 FGM（Fast gradient method）和 PGD 两种对抗训练方法，FGM 扰动计算式为[23]

$$r_{\mathrm{adv}}=\varepsilon g/\|g\|_2\tag{3}$$

式中　g——梯度

FGM 通过参数 ε 直接计算扰动，存在扰动可能不是最优的可能，相比于 FGM，PGD 做多次迭代，每次迭代都会将扰动映射到规定范围内。PGD 扰动计算式为[22]

$$r_{\mathrm{adv}|t+1}=\alpha g_t/\|g_t\|_2\tag{4}$$

$$\|r\|_2\leqslant\varepsilon\tag{5}$$

式中　α——步长

g_t——迭代 t 步的梯度

FGM 和 PGD 都以相似的方式开始，首先计算输入 x 的前向损失和反向梯度。然后根据嵌入矩阵的梯度计算对抗性扰动 r_{adv}，并将其添加到嵌入向量上。但两者在后续步骤上有所区别。

对于 FGM，它只进行一次计算，生成对抗扰动后将其与嵌入向量相加，得到对抗样本。而对于 PGD，它进行多次迭代的计算过程。每次迭代中，PGD 会根据投影操作，确保生成的对抗样本在一定范围内。也就是说，PGD 会多次计算对抗性扰动并将其与嵌入向量相加，以生成更准确的对抗样本。

两种方法都会计算 $x+r_{\mathrm{adv}}$ 的前向损失并计算其反向对抗梯度。该反向对抗梯度将与初始输入 x 的梯度相加，以获得对抗训练的效果。最后将嵌入向量恢复为初始值，并使用最终得到的梯度来更新模型参数。

1.5.3　语义特征提取层

BiLSTM 由前向 LSTM[24]和后向 LSTM 组成，它是一种用于对时序数据（如文本数据）建模的神经网络模型。LSTM 模型具有 3 个门控，即更新门、遗忘门[25]和输出门，如图 3 所示。LSTM 模型首先利用上一层的输出值计算记忆单元修正值和 3 个门控值，之后通过更新门、记忆单元修正值、遗忘门和上一层的记忆单元值计算新的记忆单元值，最后通过输出门和记忆单元值计算新的输出。公式为

$$\tilde{\boldsymbol{c}}^{(t)}=\tanh(\boldsymbol{W}_c[\boldsymbol{a}^{(t-1)},\boldsymbol{x}^{(t)}]+\boldsymbol{b}_c)\tag{6}$$

$$\boldsymbol{\Gamma}_u=\sigma(\boldsymbol{W}_u[\boldsymbol{a}^{(t-1)},\boldsymbol{x}^{(t)}]+\boldsymbol{b}_u)\tag{7}$$

$$\boldsymbol{\Gamma}_f=\sigma(\boldsymbol{W}_f[\boldsymbol{a}^{(t-1)},\boldsymbol{x}^{(t)}]+\boldsymbol{b}_f)\tag{8}$$

$$\boldsymbol{\Gamma}_o=\sigma(\boldsymbol{W}_o[\boldsymbol{a}^{(t-1)},\boldsymbol{x}^{(t)}]+\boldsymbol{b}_o)\tag{9}$$

$$\boldsymbol{c}^{(t)}=\boldsymbol{\Gamma}_u\tilde{\boldsymbol{c}}^{(t)}+\boldsymbol{\Gamma}_f\boldsymbol{c}^{(t-1)}\tag{10}$$

$$\boldsymbol{a}^{(t)}=\boldsymbol{\Gamma}_o\tanh\boldsymbol{c}^{(t)}\tag{11}$$

式中　$\boldsymbol{x}$——输入

$\tilde{\boldsymbol{c}}$——记忆单元修正值

$\boldsymbol{c}$——记忆单元

σ——sigmoid 激活函数

tanh——双曲正切函数

$\boldsymbol{a}$——输出

$\boldsymbol{W}_c$、$\boldsymbol{W}_u$、$\boldsymbol{W}_f$、$\boldsymbol{W}_o$——记忆单元、更新门、遗忘门、输出门的权重矩阵

$\boldsymbol{b}_c$、$\boldsymbol{b}_u$、$\boldsymbol{b}_f$、$\boldsymbol{b}_o$——记忆单元、更新门、遗忘门、输出门的偏置向量

$\boldsymbol{\Gamma}_u$——更新门函数

$\boldsymbol{\Gamma}_f$——遗忘门函数

$\boldsymbol{\Gamma}_o$——输出门函数

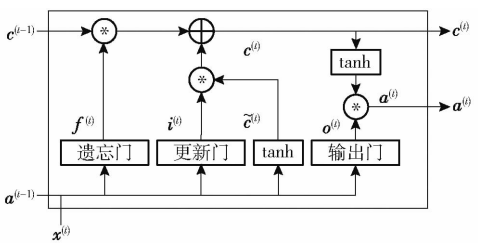

图 3　LSTM 结构示意图

Fig. 3　Schematic of LSTM structure

通过双向 LSTM 的应用,可以更好地获取上下文信息。对于每个句子,分别计算从左至右和从右至左的表示,然后将这 2 个表示拼接起来,得到当前时刻 BiLSTM 的输出。BiLSTM 通过同时考虑前向和后向的信息,提供了更全面的上下文理解能力。例如对于“蛋白质含量 14.43%”和“14.43%的蛋白质含量”等不同的前后叙述形式有更好的处理能力。

1.5.4 条件约束层

CRF(条件随机场)是一种概率模型,用于处理序列标注问题。它建立在无向图上,利用 2 种类型的分数来定义损失函数:发射分数和转移分数。发射分数是 CRF 的输入,通常是来自上一层的输出,它表示状态特征。转移分数是转移特征,用于衡量标签之间的相关性。CRF 模型的优势在于它能够全局地对整个序列进行建模,并且可以通过学习转移概率来调整标签之间的相关性。通过联合优化发射分数和转移分数,CRF 能够在序列标注任务中提供更准确和一致的标签预测结果。在基于 BIO 标注的序列标注任务中,I 标签必须紧跟在 B 标签之后。通过将 CRF 层添加到 BiLSTM 的输出层,模型能够考虑到标签之间的转移关系,从而更好地捕捉句子中相邻词之间的上下文信息。

1.6 评价方法

实验采用精确率(Precision)、召回率(Recall)、F1 值作为评估指标[26]。

2 实验

实验的运行环境:CPU 为 Intel(R) Xeon(R) Silver 4114 CPU@2.20 GHz,内存 64 GB,GPU 为 NVIDIA Corporation GP104GL(16 GB 显存),Ubuntu 20.04 操作系统,Tensorflow 2.5.0 版本。实验数据集按照比例 6∶2∶2划分为训练集、测试集和验证集,数据集之间无重复,设置批处理大小(batch_size)为 256,学习率 0.001,引入 dropout 机制,值为 0.5,以防止模型的过拟合。

本文的目的是对预训练层、对抗训练层、语义特征提取层、条件约束层 4 层网络的有效性进行验证,并不拘泥于某一特定的模型,因此另外选取 BiGRU、IDCNN 2 个语义特征提取层模型进行实验。BiGRU[27]为双向门控循环单元,是 BiLSTM 的一种简化版本。BiGRU 有 2 个门控,复位门负责记忆单元修正值的计算,更新门负责记忆单元值的计算。BiGRU 首先通过上一层的记忆单元值计算复位门和更新门,之后通过复位门计算记忆单元修正值,最后通过更新门、记忆单元值和记忆单元修正值计算新的记忆单元值即输出值。IDCNN 模型是 CNN 卷积神经网络的改进,相比传统的 CNN,IDCNN 可以捕获更长的上下文信息,并且可以实现并行性[28]。传统 CNN 末层神经元可能只得到了输入数据中一小部分信息,为了得到更多信息就要加入更多的卷积层,导致网络越来越复杂。YU 等[29]在卷积核中增加空洞,扩大模型感受野,提出 Dilated CNN 的模型。将多个 Dilated CNN 模型拼接在一起就是 IDCNN 模型。

2.1 消融实验

BiLSTM + CRF 模型通过双向长短期记忆网络和条件随机场的组合进行信息抽取,词嵌入向量输入 BiLSTM,输出每个单词对应的预测标签的分数,并作为 CRF 层的输入,从而使模型学到标签间的约束信息,可以评判 2 个相邻标签的可能性,例如 B-name、I-name 是对的,而 B-name、I-year 是错误的。BiGRU 与 BiLSTM 类似,IDCNN 则是对输入句子的每个字生成一个分数,并且在加入 CRF 后也取得了更好的效果。

优化算法选择方面,使用 Adam 与 SGD 2 种优化算法进行对比,发现 Adam 优化算法整体上取得了更好的效果。原因可能是 SGD 采用随机梯度下降,收敛速度和学习率相关性较高,在某些情况下可能陷入局部最优,而 Adam 加入了动量梯度下降和自适应学习率,在数据分布稀疏的场景,能更好利用稀疏梯度的信息,比标准的 SGD 算法更有效地收敛。因此后续实验均使用 Adam 优化算法。

以语义特征提取层结合条件约束层为基础,进行消融实验分析,结果如表 2 所示。

由表 2 可以看出,预训练层的加入为模型带来了更好的效果,采用 BERT 模型作为预训练模型时,模型可以学习到字词更完整的语义表示。相比于权重随机初始化的训练,使用 BERT 模型可以更快收敛,并获得一个更好的效果。对模型进行 Fine-tuning 时,由于参数量大幅增加,batch_size 需要调整为 32,学习率调整为 0.000 05。在 BiLSTM-CRF 模型的基础上,BERT 模型的加入使模型精确率提高 2.7 个百分点、召回率提高 2.2 个百分点、F1 值提高 2.4 个百分点;在 BiGRU-CRF 模型的基础上,BERT 模型的加入使模型精确率提高 1.1 个百分点、召回率提高 1 个百分点、F1 值提高 1 个百分点;在 IDCNN-CRF 模型的基础上,BERT 模型的加入使模型精确率提高 3.6 个百分点、召回率提高 8.1 个百分点、F1 值提高 5.9 个百分点。但 Fine-tuning 的使用对模型的效果改进不明显,在 BERT-

表 2　消融实验分析

Tab. 2　Analysis of ablation experiments

模型	优化算法	学习率	精确率/%	召回率/%	F1 值/%
BiLSTM－CRF	SGD	0.001	91.7	84.1	87.7
BiLSTM－CRF	Adam	0.001	91.2	93.6	92.4
BERT－BiLSTM－CRF	Adam	0.001	93.9	95.8	94.8
BERT－Fine-tuning－BiLSTM－CRF	Adam	0.000 05	93.0	96.2	94.6
BERT－PGD－BiLSTM－CRF	Adam	0.001	95.4	97.0	96.2
BERT－FGM－BiLSTM－CRF	Adam	0.001	95.1	96.9	96.0
BiGRU－CRF	SGD	0.001	91.2	68.6	78.3
BiGRU－CRF	Adam	0.001	92.6	95.2	93.9
BERT－BiGRU－CRF	Adam	0.001	93.7	96.2	94.9
BERT－Fine－tuning－BiGRU－CRF	Adam	0.000 05	93.9	96.5	95.2
BERT－PGD－BiGRU－CRF	Adam	0.001	94.8	96.9	95.8
BERT－FGM－BiGRU－CRF	Adam	0.001	95.1	96.7	95.9
IDCNN－CRF	SGD	0.001	83.0	73.7	78.1
IDCNN－CRF	Adam	0.001	89.1	87.4	88.2
BERT－IDCNN－CRF	Adam	0.001	92.7	95.5	94.1
BERT－Fine－tuning－IDCNN－CRF	Adam	0.000 05	93.3	96.2	94.7
BERT－PGD－IDCNN－CRF	Adam	0.001	94.1	96.4	95.2
BERT－FGM－IDCNN－CRF	Adam	0.001	94.5	96.6	95.5

BiLSTM－CRF 模型上甚至出现了降低，同时模型训练占用资源增大，训练速度减慢。原因可能在于 BERT 模型的参数是在极其庞大的数据训练下得来的，当对模型进行微调时，如果训练数据无法达到一定数量，可能会对参数造成反效果，因此对抗训练部分不使用 Fine－tuning。

引入对抗训练后，可以在训练过程中向模型中增加一些噪声，从而提高其鲁棒性和泛化性能。实验中使用 PGD 和 FGM 两种对抗训练方法。可以看出 2 种对抗训练方法都对模型效果做出了提升，在 BERT－BiLSTM－CRF 模型中，使用 PGD 方法效果更好，模型精确率提高 1.5 个百分点、召回率提高 1.2 个百分点、F1 值提高 1.4 个百分点。FGM 方法在 BiGRU 模型和 IDCNN 模型中表现更好，在 BERT－BiGRU－CRF 模型中，模型精确率提高 1.1 个百分点、召回率提高 0.7 个百分点、F1 值提高 0.9 个百分点；在 BERT－IDCNN－CRF 模型中，模型精确率提高 1.8 个百分点、召回率提高 1.1 个百分点、F1 值提高 1.4 个百分点。原因可能是数据中难免存在一些噪声和异常值，而对抗训练通过引入扰动，可以让模型学习到这些异常值，从而进行更好地处理。

2.2　不同模型性能对比

以不同的语义特征提取层为基础，选取实验中各模型最好的结果进行对比，结果如表 3 所示，可以看出，BERT－PGD－BiLSTM－CRF 模型取得了最好的效果，但 3 类模型并没有很大的差距，F1 值均在 95% 以上。结果表明，本研究所提方法具有一定的扩展性，通过不同网络层的模型替换，均可以实现较好的效果。

表 3　不同模型性能对比

Tab. 3　Comparison of performance between different models

模型	优化算法	学习率	精确率/%	召回率/%	F1 值/%
BERT－PGD－BiLSTM－CRF	Adam	0.001	95.4	97.0	96.2
BERT－FGM－BiGRU－CRF	Adam	0.001	95.1	96.7	95.9
BERT－FGM－IDCNN－CRF	Adam	0.001	94.5	96.6	95.5

随机抽取品种文本信息进行信息抽取分析，使用 BERT－PGD－BiLSTM－CRF 模型识别文本中的各种实体和关系进行信息抽取，模型将文本中包括到的信息进行了正确的抽取，抽取结果如表 4 所示。

3　讨论

利用 4 层结构设计了作物品种信息抽取模型，并利用消融实验证明了每层结构的效果。CRF 层的加入避免了模型学习到错误的相关序列，导致模

表 4 信息抽取结果

Tab. 4 Information extraction results

序号	文本内容	抽取结果
1	春性常规品种,全生育期 165 天。幼苗半直立,株高 86 厘米,长方形穗。	品种类型:春性 生育期:165 天 株高:86 厘米
2	2020—2021 年度参加生产试验,平均亩产 300.4 千克,比对照长 6878 增产 6.3%。	亩产:300.4 千克
3	穗长约 17.2 cm,每穗粒数约 154.4 粒,着粒密度均匀,半散穗,结实率 85.4%;谷粒椭圆粒,颖壳及颖尖黄色,无芒,千粒重 22.8 g。	穗长:17.2 cm 穗粒数:154.4 粒 结实率:85.4% 千粒重:22.8 g

型输出错误的预测结果,对状态特征和转移特征的学习使得模型更加考虑相邻实体的相关性。深度学习离不开大量数据,但农业领域缺乏注释语料库。BERT 预训练层的加入可以很好地解决该问题,BERT 具有庞大的预训练参数,这些训练好的参数可以帮助模型以更高的起点开始学习,朝着精细化的方向改进,从而取得细分领域更好的效果。同时,BERT 模型中 Transformer 和注意力机制的运用可以让模型学到丰富的语义表示并且能够抓住关键信息。对抗训练层的加入可以让模型更加稳健,扰动样本参与到模型的训练过程中避免了模型对训练数据的过度依赖导致的对异常样本的错误处理,同时扰动样本的增加也为模型增多了训练数据量,从而学到更多知识。同时,本研究在所有的消融实验中均加入了 BiGRU 及 IDCNN 模型的对比试验,BiLSTM 和 BiGRU 是传统 RNN 的改进,而 IDCNN 是 CNN 的改进,本研究使用这 3 种模型用不同的方式去处理自然语言文本,验证了 3 种模型在 NLP 领域的效果,也证明了本文所提出的 4 层模型结构的有效性。在本研究中,BiLSTM 略强于 BiGRU 及 IDCNN 模型,但三者差距很小。

最终结果表明,使用 Adam 优化算法的 BERT - PGD - BiLSTM - CRF 模型有着最好的效果,BERT 模型作为预训练模型,提供更丰富的语义表示;PGD 对抗训练方法为训练样本增加扰动,提高模型的鲁棒性和泛化性;BiLSTM 模型考虑长距离语义信息,以应对语言的多种表达方式;CRF 为模型学习约束信息,提高结果的精度。BERT - PGD - BiLSTM - CRF 模型在评估阶段取得了令人满意的结果。具体而言,模型精确率达到 95.4%,即模型正确预测的样本比例很高;召回率达到 97.0%,即模型能够有效地检测出正样本。

目前信息抽取领域的研究中并没有关于作物品种信息抽取的研究。CHATTERJEE 等[30]借助 RENT 算法进行了农业领域子域间关系抽取,但工作过程中需要手动识别参与。LI 等[31]将实体和关系联合提取,关系提取任务中取得了优于基线模型的结果,但静态嵌入的 BiLSTM - CRF 模型并不能取得较高的 NER 召回率。HAN 等[32]比较了 CNN、LSTM 和 BERT 3 种深度学习以及 cTAKES 和词袋模型在自动分类 SDOH 类别上的效果,但并未尝试结合使用。蒲攀等[33]优化 Transformer 用于苹果病虫命名实体识别,李书琴等[34]使用 BiGRU - CRF 模型并融合字词语义信息识别猕猴桃种植领域实体,刘新亮等[35]利用 BERT - CRF 模型识别生鲜蛋供应链实体,三者均是农业领域命名实体识别研究,并取得了较好的效果。本研究通过加入对抗训练,提出 BERT - PGD - BiLSTM - CRF 模型,在测试集上取得了良好的表现,并且模型的鲁棒性将更有保障。在下一步工作中,将着力于提高模型的泛化性,以期在多种不同来源的数据上均能表现出良好的信息抽取效果。

4 结论

(1)在 BiLSTM、BiGRU 和 IDCNN 模型的基础上增加预训练层、对抗训练层、条件约束层,提高了模型的语义获取能力、鲁棒性、对错误样本的抗扰动能力以及相邻实体抽取能力。消融实验充分证明了各层的加入均对模型的品种信息抽取能力有一定程度的提高。

(2)使用单独划分的训练集、验证集和测试集对 18 种模型进行了训练和测试,并使用精确率、召回率及 F1 值进行评价,其中 BERT - PGD - BiLSTM - CRF 模型取得了最好的效果,精确率为 95.4%、召回率为 97.0%、F1 值为 96.2%,证明了模型的卓越性能。

(3)提出的 BERT - PGD - BiLSTM - CRF 模型可以较好地处理农作物品种信息抽取任务,模型的抽取结果与人工抽取所得结果基本一致,为利用农业信息提出了一个可行的解决方案。本研究所做工作得到的数据库可以为农业数据的研究提供数据支撑,同时本研究的工作方法也可以为其他缺乏结构化数据源的领域做参考。

参 考 文 献

[1] 张红生, 王州飞. 种子学[M]. 北京:科学出版社, 2021.

[2] 王术坤,韩磊. 中国种业发展形势与国际比较[J]. 农业现代化研究,2022,43(5):814 - 822.

WANG Shukun, HAN Lei. Development situation and international comparison of China's seed industry[J]. Research of Agricultural Modernization, 2022, 43(5): 814 - 822. (in Chinese)
[3] 白岩,高婷婷,卢实,等.近四十年来我国玉米大品种的历史沿革与发展趋势[J].作物学报,2023,49(8):2064 - 2076.
BAI Yan, GAO Tingting, LU Shi, et al. A retrospective analysis of the historical evolution and developing trend of maize mega varieties in China from 1982 to 2020[J]. Acta Agronomica Sinica, 2023, 49(8): 2064 - 2076. (in Chinese)
[4] 张天真.作物育种总论[M]. 北京:中国农业出版社, 2003.
[5] 王儒敬. 我国农业信息化发展的瓶颈与应对策略思考[J]. 中国科学院院刊, 2013,28(3):337 - 343.
WANG Rujing. Bottleneck of agricultural informatization development in China and the thinking of coping strategies[J]. Bulletin of Chinese Academy of Sciences,2013,28(3):337 - 343. (in Chinese)
[6] YANG Y, WU Z, YANG Y, et al. A survey of information extraction based on deep learning[J]. Applied Sciences, 2022, 12(19): 9691.
[7] YAMASHITA R, BIRD K, CHEUNG P Y C, et al. Automated identification and measurement extraction of pancreatic cystic lesions from free-text radiology reports using natural language processing[J]. Radiology: Artificial Intelligence, 2021, 4(2): e210092.
[8] HAN X, WANG J. Earthquake information extraction and comparison from different sources based on web text[J]. ISPRS International Journal of Geo-Information, 2019, 8(6): 252.
[9] YUAN W, YANG L, YANG Q, et al. Extracting spatio-temporal information from Chinese archaeological site text[J]. ISPRS International Journal of Geo-Information, 2022, 11(3): 175.
[10] XIE K, GALLAGHER R S, CONRAD E C, et al. Extracting seizure frequency from epilepsy clinic notes: a machine reading approach to natural language processing[J]. Journal of the American Medical Informatics Association, 2022, 29(5): 873 - 881.
[11] TIFTIKCI M, ÖZGÜR A, HE Y, et al. Machine learning-based identification and rule-based normalization of adverse drug reactions in drug labels[J]. BMC Bioinformatics, 2019, 20(21): 1 - 9.
[12] PENG H, LI Q. Research on the automatic extraction method of web data objects based on deep learning[J]. Intelligent Automation & Soft Computing, 2020, 26(3): 609 - 616.
[13] MIWA M, BANSAL M. End-to-end relation extraction using LSTMS on sequences and tree structures[J]. arXiv Preprint, arXiv:1601.00770, 2016.
[14] ZHENG C, WU Z, WANG T, et al. Object-aware multimodal named entity recognition in social media posts with adversarial learning[J]. IEEE Transactions on Multimedia, 2020, 23: 2520 - 2532.
[15] WANG J, XU W, FU X, et al. ASTRAL: adversarial trained LSTM - CNN for named entity recognition[J]. Knowledge-Based Systems, 2020, 197: 105842.
[16] LAI Q, ZHOU Z, LIU S. Joint entity-relation extraction via improved graph attention networks[J]. Symmetry, 2020, 12(10): 1746.
[17] 朱红,牛浩然,朱彤.基于字词融合与对抗训练的行业人物实体识别[J].计算机工程,2023,49(5): 56 - 62.
ZHU Hong, NIU Haoran, ZHU Tong. Entity recognition of industry figures based on character and word fusion and adversarial training[J]. Computer Engineering, 2023, 49(5): 56 - 62. (in Chinese)
[18] HUANG Z, XU W, YU K. Bidirectional LSTM - CRF models for sequence tagging[J]. arXiv Preprint, arXiv:1508.01991, 2015.
[19] LI W, DU Y, LI X, et al. UD_BBC: named entity recognition in social network combined BERT - BiLSTM - CRF with active learning[J]. Engineering Applications of Artificial Intelligence, 2022, 116: 105460.
[20] JIANG M, ZHANG X, CHEN C, et al. Leveraging part-of-speech tagging features and a novel regularization strategy for Chinese medical named entity recognition[J]. Mathematics, 2022, 10(9): 1386.
[21] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv Preprint, arXiv: 1412.6572, 2014.
[22] MADRY A, MAKELOV A, SCHMIDT L, et al. Towards deep learning models resistant to adversarial attacks[J]. arXiv Preprint, arXiv:1706.06083, 2017.
[23] MIYATO T, DAI A M, GOODFELLOW I. Adversarial training methods for semi-supervised text classification[J]. arXiv Preprint, arXiv:1605.07725, 2016.
[24] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation, 1997, 9(8): 1735 - 1780.
[25] GERS F A, SCHMIDHUBER J, CUMMINS F. Learning to forget: continual prediction with LSTM[J]. Neural Computation, 2000, 12(10): 2451 - 2471.
[26] NISMI MOL E A, SANTOSH KUMAR M B. Review on knowledge extraction from text and scope in agriculture domain[J]. Artificial Intelligence Review, 2023, 56(5): 4403 - 4445.
[27] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation[J]. arXiv Preprint, arXiv:1406.1078, 2014.
[28] STRUBELL E, VERGA P, BELANGER D, et al. Fast and accurate entity recognition with iterated dilated convolutions[J]. arXiv Preprint, arXiv:1702.02098,2017.

(下转第 337 页)