doi:10.6041/j.issn.1000-1298.2025.11.050

基于 BERT - BiLSTM - CRF 的茶树病虫害命名实体识别方法

李春春[1,2]　丁　鑫[1]　张华扬[1]　王德炅[1]　王子阳[1]　江子曦[1]　雷　雨[1,2]

(1. 安徽大学互联网学院, 合肥 230601;
2. 安徽大学农业生态大数据分析与应用技术国家地方联合工程研究中心, 合肥 230601)

摘要: 在非结构化茶树病虫害文本数据中快速、准确地识别出病害实体,对于茶树病虫害高效管理与防控至关重要。目前茶树病虫害领域缺乏公开高质量数据集,且传统命名实体识别方法存在实体边界模糊、识别准确率低等问题。为此,本文构建了茶树病虫害标注语料库,并提出一种基于 BERT - BiLSTM - CRF 的茶树病虫害命名实体识别方法。利用 BERT 模型获取上下文相关词向量表示,提升茶树病虫害语义理解能力;通过 BiLSTM 模块融合正向与反向信息,捕捉长距离依赖关系,挖掘茶树病虫害的深层语义特征;采用 CRF 层对标签序列进行全局优化,提升茶树病虫害实体边界识别精度。试验结果表明,本文模型在自建数据集上的精确率、召回率和 F1 值分别为 98.42%、98.75% 和 98.58%,达到了 SOTA 性能。在数据集 MSRANER 和 CLUENER2020 上 F1 值分别为 92.4% 和 87.8%,验证了模型泛化能力。研究结果不仅为农业领域知识抽取提供了融合语言学规则与深度学习的范式参考,还为构建茶树病虫害知识图谱与问答系统提供了技术支持,从而有助于茶树种植者和农业技术人员实现高效决策。

关键词: 茶树病虫害; 命名实体识别; BERT; 知识图谱; 问答系统

中图分类号: S435.711; TP391.1　　**文献标识码**: A　　**文章编号**: 1000-1298(2025)11-0517-11　　**OSID**:

Named Entity Recognition Method for Tea Tree Pests and Diseases Based on BERT - BiLSTM - CRF

LI Chunchun[1,2]　DING Xin[1]　ZHANG Huayang[1]　WANG Dejiong[1]　WANG Ziyang[1]　JIANG Zixi[1]　LEI Yu[1,2]

(1. *School of Internet*, *Anhui University*, *Hefei* 230601, *China*
2. *National Engineering Research Center for Agro-Ecological Big Data Analysis and Application*, *Anhui University*, *Hefei* 230601, *China*)

Abstract: Rapidly and accurately extracting disease entities from unstructured tea tree pest and disease text data is crucial for the efficient management and prevention of tea tree pests and diseases. Currently, there is a lack of publicly available, high-quality datasets in the field of tea plant pests and diseases. Moreover, traditional named entity recognition (NER) methods often suffer from issues such as ambiguous entity boundaries and low recognition accuracy. To address these challenges, the annotated corpus specifically for tea plant pests and diseases was constructed and a NER method for tea tree pests and diseases was proposed based on the BERT - BiLSTM - CRF architecture. Firstly, the BERT model was employed to obtain context-related word vector representations to improve the semantic understanding ability of tea tree pests and diseases; the BiLSTM module was used to fuse forward and reverse information to capture long-distance dependencies and deeply explore the deep semantic features of tea tree pests and diseases; the CRF layer was used to globally optimize the label sequence to improve the entity boundary recognition accuracy of tea tree pests and diseases. Experimental results showed that the

收稿日期: 2025-04-10　修回日期: 2025-07-15

基金项目: 安徽省自然科学基金项目(2108085QA35)、安徽省高校科研计划项目(2024AH052212)、国家自然科学基金项目(32301701、71971002)、安徽省高等学校科学研究项目(2022AH050085)、合肥市自然科学基金项目(202309)和河南省重点研发专项(241111110800)

作者简介: 李春春(1990—),女,讲师,博士,主要从事自然语言处理与知识图谱研究,E-mail: licc@ahu.edu.cn

通信作者: 雷雨(1990—),男,讲师,博士,主要从事农情信息获取技术与装备研究,E-mail: leiyu@ahu.edu.cn

proposed model achieved a precision of 98.42%, a recall of 98.75%, and an F1-score of 98.58% on the self-constructed dataset, reaching state-of-the-art performance. Additionally, F1-scores of 92.4% and 87.8% were achieved on the general-domain datasets MSRA NER and CLUENER2020, respectively, demonstrating the model's generalization capability. Research findings not only provided a paradigm reference for knowledge extraction in the agricultural field by integrating linguistic rules and deep learning, but also provided technical support for building a knowledge graph and question-answering system for tea pests and diseases, thereby helping tea growers and agricultural technicians to make efficient decisions.

Key words: tea tree pests and diseases; named entity recognition; BERT; knowledge graph; question-answering system

0 引言

茶叶是我国重要的经济作物[1-2]。然而,茶树病虫害种类繁多,如茶圆赤星病等,严重威胁茶叶产量与品质,并因农药残留问题制约其国际竞争力[3-4]。因此,开展高效的茶树病虫害防控工作,对保障茶叶生产安全、推动茶产业绿色可持续发展至关重要[5-6]。目前,茶树病虫害相关资源分散,整合不足,科技资源与防控需求不匹配,技术服务缺乏统一标准,严重影响病虫害防治措施的科学实施[7]。

随着农业数字化转型的持续推进,农业知识智能化服务在农业生产与管理中的技术支撑作用日益突出[8-9]。然而,在农业信息化过程中,数据资源分散和信息融合困难导致知识利用效率低[10]。近年来,知识图谱技术在农业领域显示出其在语义推理和数据整合方面的优势[11-14]。作为构建知识图谱的关键技术,命名实体识别(Named entity recognition, NER)对农业术语识别和标准化表达至关重要[15-17]。

命名实体识别是指从文本中识别出属于预定义语义类型(如人物、地点等)的实体标识符任务[18]。早期的 NER 系统多采用基于规则的方法,依赖人工构建的领域词典和句法模式识别实体[19-20]。QUIMBAYA 等[21]提出一种结合词典的信息抽取方法,电子健康记录中的命名实体识别,召回率达到 57.3%。CHATTERJEE 等[22]提出的 RENT 方法通过结合正则表达式和自然语言处理技术,并利用领域专家定义的术语模式,在农业文本 F1 值超过 68%,显著提高了术语提取准确性。基于规则的 NER 方法在特定领域表现良好,但因依赖人工规则、成本高、通用性差、召回率低和跨领域适应性差,限制了其在多样化农业文本中的应用。

基于机器学习的 NER 方法将实体识别任务转化为分类问题,利用统计学习模型进行建模与求解[23-24]。SZARVAS 等[25]提出一个基于 AdaBoostM1 和 C4.5 决策树的多语言命名实体识别系统,通过大规模特征集和多数投票机制提升识别性能。李想等[26]提出一种基于条件随机场的方法,通过整理和标注农业领域数据集的特征信息,实现农作物和病虫害等实体的自动识别及三元组构建。机器学习在 NER 任务中能够自动提取特征,减少人工依赖;但训练数据依赖性强,泛化能力需提高,尤其在领域迁移、嵌套实体识别和复杂上下文理解方面存在局限。

近年来,深度学习在 NER 任务中因连续向量表示和非线性语义处理而广泛应用,其中 BiLSTM - CRF 架构表现尤为突出[27-30]。YANO[31]提出的基于字符的 BiLSTM - CRF 端到端神经网络模型在提取日语文本中的疾病命名实体及判断确诊状态方面,性能优于传统字符级 CRF 和 SVM 方法。AN 等[32]设计了融合多头自注意力机制的 BiLSTM - CRF 模型,更好地利用全局信息和多级语义特征,有效提升了 NER 性能。蒲攀等[33]引入注意力机制至 BiLSTM - CRF 框架,苹果病虫识别 F1 值提至 92.66%。WANG 等[34]提出的 ASLT 模型通过融合话题特征和注意力机制,增强了农业文本 NER 对复杂构词等问题的鲁棒性。CHEBBI 等[35]研究了大型语言模型在农业等非结构化数据领域命名实体识别中的有效性及其应用。DE 等[36]提出了结合 NER 和关系抽取的方法,利用数据增强和定制模型 AgNER - BERTa 与 AgRE - BERTa 构建农业知识图谱,解决了农业领域文本自动化处理工具不足的问题。综上,BiLSTM - CRF 架构在序列标注任务中表现出色,已广泛应用于多个领域的 NER 任务中。然而,针对茶树病虫害领域的 NER 研究仍较为缺乏。茶树病虫害相关知识通常以碎片化、非结构化的形式分散于各类文献与资料中,且缺乏公开、高质量的标注数据集,导致现有 NER 方法在该领域的直接应用面临诸多挑战,如实体边界模糊、识别准确率低等问题,亟需开展茶树病虫害领域的 NER 方法研究。

为解决上述问题,本文以茶树病虫害为研究对

象，构建专用于该领域的数据集，并提出一种基于 BERT－BiLSTM－CRF 的命名实体识别方法。利用 BERT 模型对输入文本进行语义编码，提取丰富的上下文特征，以增强模型对病虫害相关术语的理解能力；借助 BiLSTM 网络捕捉文本中的长距离依赖关系，进一步提升对语义结构和语义关联的建模效果；采用 CRF 层对实体边界和类型进行精确标注，以实现对茶树病虫害相关命名实体高效识别与分类。

1　茶树病虫害数据集构建

1.1　数据来源

通过多渠道收集茶树病虫害数据，涵盖结构化、半结构化和非结构化数据。其中，结构化数据具有规范化的组织形式，主要来源于关系型数据库。从国家农业科学数据中心（https://www.escience.org.cn）获取结构化数据，该中心提供了常见的 25 种茶树病害和 71 种茶树虫害，以及相关的关键属性信息，部分数据示例如表 1 所示。半结构化数据通常以 HTML 和 XML 等标记语言呈现，具有更高灵活性，主要来源于茶树官方网站和百度百科等平台。为高效获取这些数据，基于 Scrapy 框架编写网页爬虫程序，结合页面结构特点与爬取规则，批量获取相关病虫害页面中的关键信息。

半结构化数据爬取主要针对网页推荐的属性词条进行定向提取，并保留原页面的半结构化表达形式。这种将整段文本作为单一属性值的方式，有助于保持数据完整性和一致性，但也存在信息遗漏与表达模糊问题。例如，在茶圆赤星病的“症状”属性中，往往隐含着尚未被结构化挖掘的丰富信息，如为害部位、别名、分布区域等。这类蕴含实体关系的信息提取，属于非结构化数据的处理范畴。此外，茶树病虫害相关的专业书籍，如《茶树病虫害识别与防治手册》等，也属于典型的非结构化文本数据，亟需通过自然语言处理技术进行进一步的信息抽取与整理。

表 1　茶树病虫害结构化数据部分示例

Tab.1　Partial display of structured data on tea tree pests and diseases

中文病名	中文别名	拉丁学名	为害部位	分布地点
茶轮斑病	茶梢枯死病	*Pestalotiopsis theae* Steyaert	叶片、新梢	长江流域茶区
茶小绿叶蝉	浮尘子、叶跳虫	*Empoasca pirisuga* Matumura	叶片	贵州、安徽、江苏、浙江、湖南
茶尺蠖	拱拱虫、拱背虫、量尺虫、造桥虫	*Ectropis oblique hypulina* Wehrli	叶片	长江流域以南各产茶省

1.2　数据预处理

1.2.1　数据清洗

在茶树病虫害研究中，数据质量对分析结果至关重要。为此，需要对半结构化和非结构化数据进行清洗。预处理包括：去除文本中的无意义字符；定义清洗策略，利用正则表达式处理重复记录和异常值；统一数据格式。针对茶树病虫害专业书籍资源，采用高分辨率扫描仪扫描相关内容，应用光学字符识别技术将扫描文本自动解析转换为可编辑电子文档。结合自然语言处理（Natural language processing，NLP）技术进行自动校对和格式化处理，减少识别错误，为知识提取任务奠定基础。

1.2.2　数据标注

通过 NLP 技术提取茶树病虫害文本中的实体关系，采用 BIO 标注法对实体进行标注，其中，B、I 和 O 分别表示实体开始、实体内部和非实体，如图 1 所示。标注工作由 2 名团队成员独立完成，一致性达到 97.83%，然后由农学专家进行审核，确保数据准确性和专业性。标注好的数据将作为 BERT－BiLSTM－CRF 模型输入，其质量直接影响模型实体抽取效果。

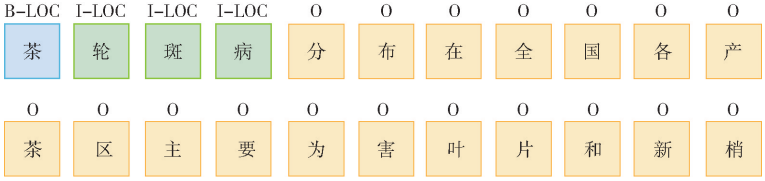


图 1　基于 BIO 方法的实体标注示意图

Fig.1　Schematic of entity annotation based on BIO method

梳理并标注了实体关键属性类别，如表 2 所示。为了处理数据与属性的对应关系，采用独热编码方法，将每个属性类别转换为独特的二进制向量，确保类别数据的独立性和一致性，如表 3 所示。例如，“棉盲蝽”属性向量表示为（0，1，0，0，0）。该向量化方法简化了数据处理流程，解决了分类器处理属性数据时的难题，提高模型准确性。

表 2 茶树病虫害部分属性标签示例

Tab. 2 Example attribute labels for tea plant diseases and pests

属性标签	相关性描述
病原	引起病虫害的病原
别名	病虫害的其他常见名
症状	感染病虫害后的症状
地点	病虫易发生的地区
寄主	害虫的常见寄生对象

表 3 基于 ONE - HOT 方法的属性标签示意

Tab. 3 Attribute label based on ONE - HOT method

数据	寄主	别名	地点	学名	形态
桑树、茶树等	1	0	0	0	0
棉盲蝽	0	1	0	0	0
分布在全国各地	0	0	1	0	0

1.2.3 数据集划分

通过多源异构数据采集、清洗与数据标注后，本文构建了一个覆盖病害、虫害等多维属性的茶树病虫害标注语料库，共包含 11 382 条高质量样本。其中实体病原、别名、症状、地点和寄主数量分别为：1 812、2 346、1 954、2 513、2 757 条。如表 4 所示，数据集按比例 7∶2∶1划分为训练集、测试集和验证集，以支持模型有效训练和评估。

表 4 实验数据集分配信息

Tab. 4 Experimental dataset distribution information

数据种类	训练集	测试集	验证集	数量总计
茶树病害	3 356	959	479	4 794
茶树虫害	4 611	1 318	659	6 588
总计	7 967	2 277	1 138	11 382

2 面向茶树病虫害的命名实体识别模型

通过对茶树病虫害语料分析，发现数据特征主要围绕实体和属性展开，且结构较为清晰。因此，将实体识别与关系抽取统一为命名实体识别任务，简化了知识抽取过程并提升了效率。

针对茶树病虫害语料中实体边界模糊、识别准确率低等问题，构建一个融合 BERT 模型与 BiLSTM - CRF 框架的 BERT - BiLSTM - CRF 多层网络模型，用于茶树病虫害命名实体识别，以提高识别准确性和效率。BERT - BiLSTM - CRF 模型结构如图 2 所示，该模型包括 BERT 层、BiLSTM 层和 CRF 层。由 BERT 层对输入文本进行处理，将其转换为富含上下文信息的向量表示；BiLSTM 层在此基础上精准提取与命名实体识别任务紧密相关的特征信息；CRF 层通过学习标签之间的依赖关系，生成最终的预测标注序列，确保实体识别结果准确性和一致性。

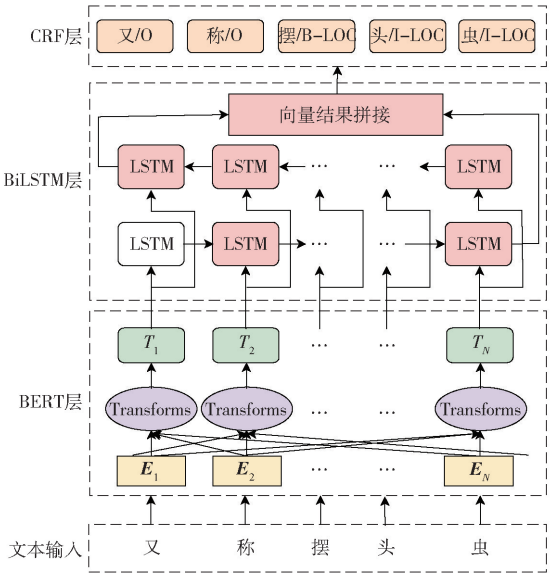


图 2 BERT - BiLSTM - CRF 模型结构图

Fig. 2 BERT - BiLSTM - CRF model structure diagram

2.1 词向量嵌入层

BERT(Bidirectional encoder representations from transformers)模型是一种基于 Transformer 架构的预训练语言模型，在 NLP 领域取得了显著成就。利用 BERT 模型对文本数据进行嵌入处理，以捕获与上下文相关的词向量表示。当文本输入 BERT 模型后，在 Embedding 层进行编码，转换成模型可识别的词单元。这些词的索引被映射到一个固定维度的向量空间，形成词向量的初始部分。利用多头注意力机制评估序列中每个 Token 与其他 Token 的相关度。在此过程中，各个注意力头分别关注于序列的不同部分，捕捉不同序列位置的上下文特征，并进行整合。将这些具有依赖关系的特征输入到前向传播层，引入非线性因素的同时也进一步增强了特征的融合。其中，运用残差连接将输入与输出相加，确保深层网络的有效学习。同时，也采用层归一化操作保证训练的稳定，即

$$E_{attn} = \mathrm{LayerNorm}(E_{i-1} + \mathrm{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})) \tag{1}$$

$$E_i = \mathrm{LayerNorm}(E_{attn} + FFN(E_{attn})) \tag{2}$$

式中 E_{i-1}——上一层输出，作为第 i 层输入

$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$——对每个输入向量进行变换得到的查询矩阵、索引矩阵和值矩阵

$\mathrm{MultiHead}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})$——利用 Concat 操作对多个注意力头结果进行拼接

图 2 中 BERT 层展示了对文本“又称摆头虫”进行嵌入表示过程。每个词被映射为固定维度的向量，以表示其语义信息。通过句向量区分不同句子结构，表示为 $E_1 \sim E_N$。此外，还为每个词添加位置

信息，使得 BERT 即便不使用循环结构也能感知词序。每个词表示 3 部分向量总和，形成完整的输入序列。这种融合多维度信息的方法使 BERT 能够精确理解语言语义、句子结构和词序。

2.2　特征提取层

双向长短时记忆网络（Bidirectional long short term memory，BiLSTM）是一种先进的循环神经网络，与传统单向网络不同，能够从序列 2 个方向捕捉信息，从而更全面地建模序列数据。这种双向机制使 BiLSTM 能够更好地捕捉长距离依赖和复杂时序特征。BiLSTM 利用 LSTM 的门控机制有效防止梯度消失，确保网络稳定学习深层特征。尽管 BERT 在自然语言处理中表现出色，但在时序建模方面仍有局限。BiLSTM 的加入不仅补充了 BERT 在时序建模上的不足，还为后续的 CRF 层提供了准确的发射分数，增强了模型性能。

图 2 中的 BiLSTM 层可以看出，它由 2 个并行的 LSTM 层组成，一个负责正向处理序列，另一个处理反向序列。在每个时间步中，正向 LSTM 生成一组隐藏向量，而反向 LSTM 生成另一组。这 2 组隐藏向量随后被拼接，形成每个时间步的完整上下文信息。这种设计结构使 BiLSTM 能够同时考虑序列中每个元素前后文信息，从而更有效地捕捉长距离依赖关系。

2.3　序列预测层

条件随机场（Conditional random field，CRF）是一种用于序列标注的统计模型，能够通过考虑标签间的依赖关系增强一致性和准确性。本文模型利用 CRF 层进行全局路径优化。

在 CRF 层中，利用 BiLSTM 层提供的发射分数（即每个 Token 对应每个标签的得分）和字符标注之间的概率转移矩阵，计算全局得分，计算式为

$$S(y)=\sum_{t=1}^{L}V_{t,y_t}+\sum_{t=1}^{L+1}E_{y_{t-1},y_t} \tag{3}$$

式中　y——标签序列

$S(y)$——给定标签序列 y 的全局得分

V_{t,y_t}——在时间步 t 处，Token t 被标注为 y_t 时的发射分数

E_{y_{t-1},y_t}——从标签 y_{t-1} 转移到标签 y_t 的转移分数

L——序列长度

通过动态规划计算所有可能路径总分，取其对数，再减去真实路径得分，得到输出损失值，计算式为

$$L=\lg\left(\sum_{y'}e^{S(y')}\right)-S(y_{\text{true}}) \tag{4}$$

式中　L——模型损失值

$\sum_{y'}e^{S(y')}$——所有可能标签序列 y' 的全局得分 $S(y')$ 进行指数运算后求和

$S(y_{true})$——真实标签序列 y_{true} 的全局得分，模型对真实标签序列评分

模型损失值 L 是所有可能标签序列得分的对数和与真实标签序列得分之差，该差值越小，表示模型对真实标签序列的预测越准确。

CRF 层在序列标注中采用维特比算法选取得分最高的标签序列，实现标注优化。该方法不仅评估单个标签的得分，还综合考虑标签间转移概率，从而提升整体标注性能。

2.4　知识融合

数据源包括结构化、半结构化和非结构化数据，该数据中存在模糊和重复信息。知识融合旨在整合这些数据，提高知识图谱质量。例如，网络爬取得到的“茶轮斑病”与关系数据库中的“茶梢枯死病”指代同一实体，但名称不同，且各自拥有不同关系和属性集合。结构化数据源中关于茶饼病属性包括“地理分布”和“为害症状”，而半结构化数据源则涵盖“流行规律”和“分级标准”属性。因此，需要对提取的实体进行评估，确认是否指向同一对象，并处理实体间的相似性和差异性，以实现有效合并。

在提取的三元组中，不同名称的实体可能属于同一类型，这种模糊性会影响数据质量，需要实体对齐。为解决这一问题，使用余弦距离计算实体相似性，以提高测量准确性，计算式为

$$Align_{cosine}(e_1,e_2)=\frac{|A(e_1)\cap A(e_2)|}{\sqrt{|A(e_1)||A(e_2)|}} \tag{5}$$

式中　$A(e_1)$、$A(e_2)$——实体 e_1 和 e_2 属性集

通过设定阈值判断实体是否匹配，实现知识融合，提升数据一致性。

2.5　试验

2.5.1　试验环境

采用 Python 语言和 PyTorch 框架进行试验，相关参数见表 5。为了评估模型性能，选用精确率、召回率和 F1 值作为评价指标。

表 5　试验运行环境参数

Tab.5　Experimental operating environment parameters

运行硬件/软件参数	数值/型号
CPU	Intel(R) Core(TM) i7－10700 CPU@2.90 GHz
GPU	GeForce RTX 3090
内存/GB	32
PyTorch	2.1.2
Pycharm	2024.1.1
Python	3.11

2.5.2 BERT－BiLSTM－CRF 模型参数设置

BERT－BASE－Chinese 模型包含 12 层结构和 12 个注意力头，参数量为 1.1×10^9。BERT 在中文 NLP 任务中应用广泛，且在诸多研究中表现出色，提供了坚实的预训练模型基础。

在模型训练阶段，采用 PyTorch 框架中的 Adam 优化器（版本 2.5.1），这是一种基于随机梯度下降的优化方法，通过自适应调整学习率加速模型的收敛并提高训练效率。实施线性学习率衰减策略，随着训练轮数增加逐渐降低学习率，直至为 0，有助于初期快速收敛并在后期避免过拟合。此外，为了进一步提升模型性能，对超参数进行了细致调整，本文模型以及对比模型最优参数配置如表 6 所示。

表 6 不同模型最佳超参数设置

Tab.6 Optimal hyperparameter settings for different models

模型	批次尺寸	最大序列长度	丢弃率	学习率	LSTM 单元数
TextCNN－CRF	32	128	0.3	3×10^{-3}	256
BiLSTM－CRF	32	128	0.3	3×10^{-3}	256
IDCNN－CRF	512	128	0.3	1×10^{-2}	256
BERT－Softmax	512	128	0.3	1×10^{-2}	128
ALBERT－BiLSTM－CRF	128	128	0.3	3×10^{-3}	512
BERT－BiLSTM－CRF	32	128	0.3	3×10^{-3}	256

3 结果与分析

通过对比试验，从超参数分析、收敛测试、实体识别效果以及模型泛化能力等 4 个维度，将所提出的联合模型与其他 5 种模型进行性能对比分析。

3.1 超参数分析

超参数调整对提升模型泛化能力和预测精度至关重要，是实现高效学习的关键环节。因此，本文探讨 5 个超参数：批次尺寸、最大序列长度、丢弃率、学习率和 LSTM 单元数，以优化模型性能，试验结果如图 3 所示。由图 3a 可知，当批次尺寸为 32 时，模型达到最佳性能。若批次尺寸过大，可能会由于过度限制模型的泛化能力而导致性能下降。图 3b 指出，当最大序列长度设为 128 时，模型表现最优。若序列过长，可能会因为难以捕捉关键信息和计算资源的不必要消耗而影响模型性能。图 3c 显示，当丢弃率为 0.3 时，模型性能最优；如果丢弃率过高，模型性能会有所下降。图 3d 表明，当学习率定为 3×10^{-3}时，模型性能最佳。学习率过高可能导致模型在训练过程中震荡难以收敛，而过低则可能导致收敛速度过慢。图 3e 显示，当 LSTM 单元数为 256 时，模型精确率达到最高。然而，当 LSTM 单元数量增加至 1 024 时，精确率有所下降，这可能是由于模型出现了过拟合现象。

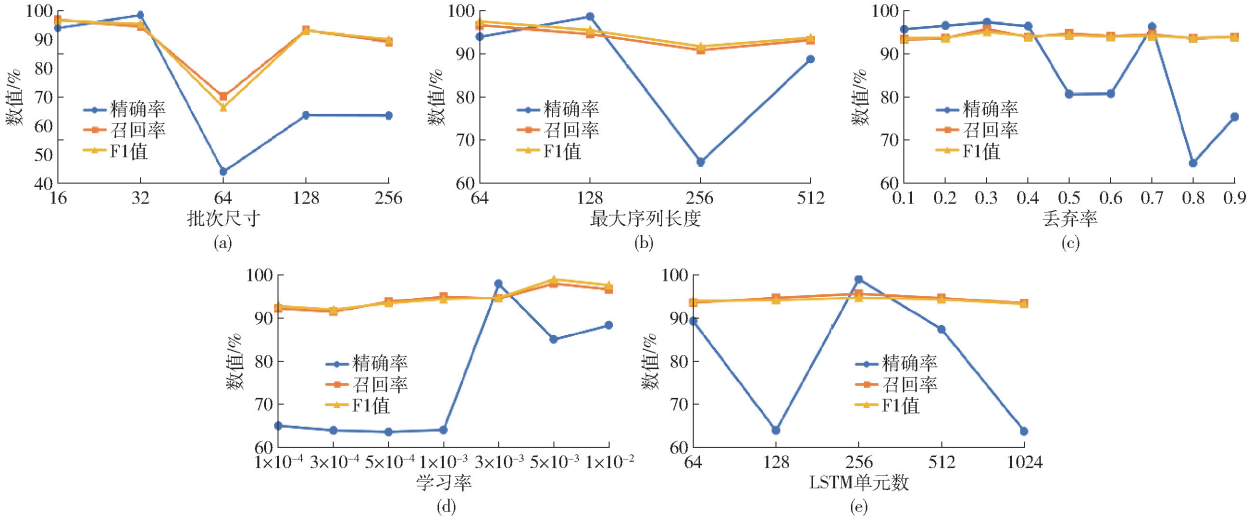


图 3 不同超参数对 BERT－BiLSTM－CRF 模型性能的影响

Fig.3 Impact of different hyperparameters on performance of BERT－BiLSTM－CRF model

3.2 收敛测试

在深度学习领域，损失函数通过评估模型预测输出与实际标签之间的差异，为模型训练过程提供指导。通过监测不同学习率下模型在训练集和验证集上的损失值，分析其过拟合情况。

如图 4 所示，当学习率为 3×10^{-3}时，训练集损失值迅速降低并趋于稳定，验证集损失也显示出类似的下降趋势，表明在此学习率下，模型不仅优化效率高，而且具有良好的泛化能力。相较之下，当学习率为 3×10^{-4}时，训练集和验证集损失下降速度变慢，收敛值较高，说明泛化能力不如 3×10^{-3}时表现。当学习率为 3×10^{-5}时，训练集损失下降缓且

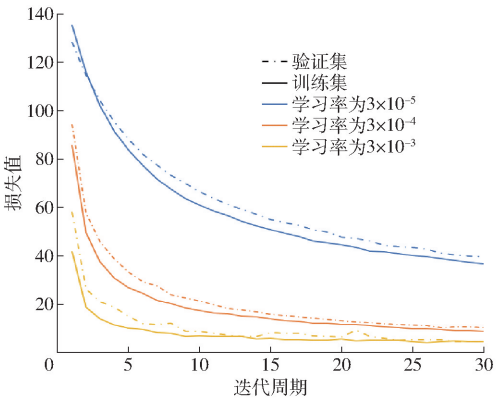


图 4　不同学习率下的损失值变化曲线

Fig. 4　Loss change curves under different learning rates

未达到较优状态，但验证集损失未出现过拟合现象。综合来看，当学习率为 3×10^{-3} 时，模型在训练集和验证集上的损失值下降趋势一致，且未出现明显的过拟合现象，显示出最佳泛化性能。

3.3　预测结果

3.3.1　性能对比试验

基于自构建的茶树病虫害数据集，对命名实体识别领域的几种经典模型进行了对比试验。不同模型性能对比结果如表 7 所示。

表 7　不同命名实体识别模型试验结果性能对比

Tab. 7　Performance comparison of experimental results of different named entity recognition models

%

模型	精确率	召回率	F1 值
TextCNN－CRF	82. 22	82. 22	82. 22
BiLSTM－CRF	88. 72	84. 36	86. 41
IDCNN－CRF	93. 62	95. 15	94. 35
BERT－Softmax	95. 02	93. 64	94. 39
ALBERT－BiLSTM－CRF	95. 16	94. 58	95. 22
BERT－BiLSTM－CRF	98. 42	98. 75	98. 58

由表 7 可知，TextCNN－CRF 和 BiLSTM－CRF 均依赖 Word2Vec 进行文本向量化，前者因缺乏深度序列建模能力，F1 值较低(82. 22%)；后者凭借双向序列建模能力 F1 值提升至 86. 41%，但仍然受限于词向量表示能力。IDCNN－CRF 通过膨胀卷积处理字符级输入，避免分词误差并扩展感受野，F1 值达到 94. 35%，然而，由于缺乏长距离依赖建模，虽然召回率较高，但精确率受到一定限制。

BERT－Softmax 凭借 BERT 强大的上下文表示能力，能够自动学习文本语义信息，整体表现优于传统模型，F1 值达到 94. 39%。然而，Softmax 层无法建模标签之间的依赖关系，限制了其识别性能。ALBERT－BiLSTM－CRF 通过参数共享和因子分解优化了 BERT 计算效率，结合 BiLSTM 和 CRF 后，在命名实体识别任务中优于 BERT－Softmax，F1 值达到 95. 22%。本文提出的 BERT－BiLSTM－CRF 融合了 BERT 语言理解能力、BiLSTM 长序列依赖建模能力以及 CRF 的全局优化能力，在试验中取得了最佳性能，精确率为 98. 42%，召回率为 98. 75%，F1 值达 98. 58%，F1 值相比 ALBERT－BiLSTM－CRF 提升 3. 36 个百分点，较 BERT－Softmax 提升 4. 19 个百分点，验证了该方法的有效性和优越性。

3.3.2　消融试验

通过依次移除关键模块 BERT 和 BiLSTM，探讨各个模块对模型整体性能的贡献，结果如表 8 所示。由表 8 可知，BiLSTM－CRF 表现最弱，凸显了 BERT 预训练语言模型在上下文表示能力方面的优势，提升了 BERT－BiLSTM－CRF 在自然语言处理任务中的性能。BERT－CRF 结果表明 BiLSTM 模块在文本特征提取中的积极作用，该模块对 BERT－BiLSTM－CRF 整体性能提升也做出了显著贡献。

表 8　模型消融试验结果

Tab. 8　Results of model ablation experiments　%

模型	精确率	召回率	F1 值
BiLSTM－CRF	88. 72	84. 36	86. 41
BERT－CRF	95. 77	96. 12	97. 21
BERT－BiLSTM－CRF	98. 42	98. 75	98. 58

3.3.3　不同实体类别评估

为了更细粒度地评估模型性能，对茶树病虫害数据中的 5 个实体类别召回率进行了分析，结果如图 5 所示。由图 5 可知，BERT－BiLSTM－CRF 模型在“寄主”、“地点”和“别名”3 个实体类别上的表现尤为突出，召回率分别达到 98. 08%、96. 77% 和 92. 50%，显示出模型对这些类别的正样本具有很高的识别能力。而各模型在“病原”类别上的召回率都相对较低，可能是由于该类别的数据样本较少，导致模型难以充分捕捉特征。此外，该类别与其他类别间可能存在一定程度的特征相似性，增加了模型识别过程中的混淆，进一步影响了召回率。

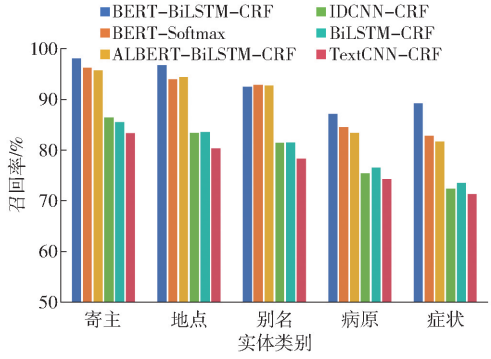


图 5　不同实体类别识别结果

Fig. 5　Recognition results for different entity categories

3.3.4 模型时间效率

对比 6 种模型训练时间和推理时间。训练时间指模型在训练集上运行 30 个周期的时间,推理时间为处理单个验证集样本所需时间,结果如表 9 所示。基于 BERT 框架的模型在训练时间上表现优异,是因为 BERT 架构的高度并行性和预训练模型的微调策略,这使得训练过程高效且避免了从零开始训练的耗时。相比之下,非 BERT 模型训练时间较长。在推理时间方面,BERT 系列模型同样表现出色,本文模型推理时间仅为 1.72 ms,适合快速处理大量数据的场景。尽管本文模型在训练集和验证集上的时间略有延长,但其在精确率、召回率和 F1 值上显著优于其他模型。这些性能提升突显了模型稳健性,使其在需要高精度和深度语义理解的应用中更具优势。

表 9 不同模型训练时间和推理时间

Tab.9 Training and inference times for different models

模型	训练时间/s	推理时间/ms
TextCNN - CRF	371.01	7.57
BiLSTM - CRF	395.29	8.33
IDCNN - CRF	433.47	6.71
BERT - Softmax	217.78	1.46
ALBERT - BiLSTM - CRF	227.29	1.55
BERT - BiLSTM - CRF	231.78	1.72

3.3.5 模型统计显著性检验

通过统计显著性试验进一步验证不同模型的稳定性和性能差异。试验中,每个模型在训练集上进行 3 次独立训练,每次 30 个周期,在验证集上计算精确率、召回率和 F1 值,最终统计每个模型平均值和标准差,显著性试验结果如表 10 所示。

表 10 不同模型统计显著性试验结果

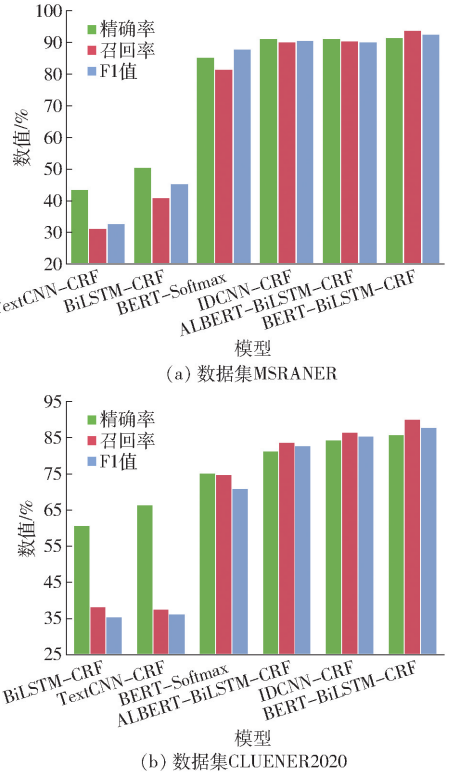
Tab.10 Statistical significance experiments for different models %

模型	精确率	召回率	F1 值
TextCNN - CRF	81.10 ± 2.62	81.50 ± 3.47	83.11 ± 6.04
BiLSTM - CRF	87.23 ± 1.50	84.53 ± 0.97	85.54 ± 1.04
IDCNN - CRF	92.04 ± 3.43	94.99 ± 3.23	93.84 ± 4.17
BERT - Softmax	94.62 ± 0.82	93.92 ± 0.94	95.48 ± 0.40
ALBERT - BiLSTM - CRF	95.27 ± 0.59	94.82 ± 0.76	94.27 ± 0.50
BERT - BiLSTM - CRF	98.45 ± 0.36	98.53 ± 0.49	98.62 ± 0.28

从表 10 可看出,BERT 系列模型在所有性能指标上表现优异,主要得益于其在模型复杂度、训练数据多样性及优化算法上的优势,使得训练过程更加稳定,减少性能波动。BERT - BiLSTM - CRF 模型在平均值和标准差上表现尤为突出,标准差分别为 0.36%、0.49% 和 0.28%,表明其在多次训练中性能变化极小,具有极高稳定性。相比之下,非 BERT 模型如 TextCNN - CRF 因结构较为简单,难以捕捉深层语义信息,且对数据变化敏感,导致性能波动较大。

3.4 模型泛化能力验证

将 BERT - BiLSTM - CRF 模型与对比模型在 MSRANER[37] 和 CLUENER2020[38] 2 种公开数据集上进行对比试验。如图 6 所示,本文模型在 2 个数据集上 F1 值达到 92.4% 和 87.8%,与其他模型相比均最优,显示其较佳的跨数据集稳定性。数据集 MSRANER 由于其标准化的标注和较为清晰的实体边界,使得模型获得较高的 F1 值。而数据集 CLUENER2020 中实体类别间的语义差异较小且边界模糊,增加了类别区分难度。此外,该数据集中标签分布不均衡,高频类别样本数量远超低频类别,导致模型训练时可能偏向高频类别,影响低频类别的召回率。尽管如此,本文模型在复杂数据环境下仍展现出良好的鲁棒性,能够更准确地识别实体边界,并在类别区分上取得更优结果。



(a) 数据集MSRANER

(b) 数据集CLUENER2020

图 6 不同模型在数据集 MSRANER 和 CLUENER2020 上性能对比

Fig.6 Performance comparison of different models on MSRANER and CLUENER2020 dataset

3.5 基于茶树病虫害的知识图谱可视化

采用 Neo4j 图数据库存储和管理茶树病虫害实体及关系数据,实现高效的知识管理和灵活的数据操作,Neo4j 主要由节点和边构成[39-40]。节点作为数据的核心,以圆形表示,用于存储实体信

息，如茶树病虫害名称。边则表示实体间关系，通过带箭头的线展示关联方向和类型，例如“地点”和“症状”。这些相互连接的实体和关系构成复杂的数据网络，支持对实体及其关联关系进行增删改查操作。

使用“LOAD CSV”命令将结构化数据导入图数据库，数据以“实体-属性-属性值”的三元组形式存储，例如“(茶白星病，地点，云南)”。图 7 为“茶白星病”相关信息，其中橙色圆圈表示病害实体，线表示实体与属性值之间关系，红色圆圈表示具体属性值。这种知识图谱直观地呈现了病害特征和相关信息，为研究人员提供了便捷的信息获取方式，并为构建茶树病虫害问答系统打下了基础。

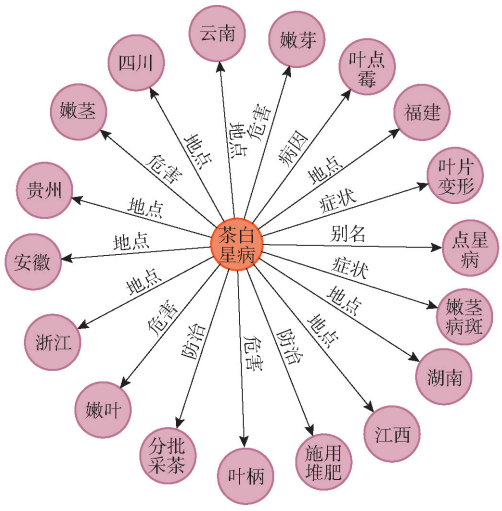


图 7　知识图谱的部分实例

Fig. 7　Examples of knowledge graph instances

3.6　基于茶树病虫害知识图谱的智能问答系统

设计并开发了一个基于茶树病虫害知识图谱的智能问答系统，旨在提高农户和农业专家获取专业知识的效率。

茶树病虫害智能问答系统设计步骤为：用户提交问题后，系统使用 BERT－BiLSTM－CRF 模型识别问题中与茶树病虫害相关的实体。如果问题中未明确属性信息，系统将利用 BertForSequenceClassification 模型进一步分析，以识别属性内容。识别出属性后，系统在 Neo4j 数据库中检索相关信息并适当润色，以自然语言形式提供答案。该属性分类任务精确率、召回率和 F1 值达 82.32%、83.86% 和 81.53%。图 8 为一个问答示例，其中用户询问“茶轮病的症状是什么”，系统生成并返回了相应的自然语言回答。

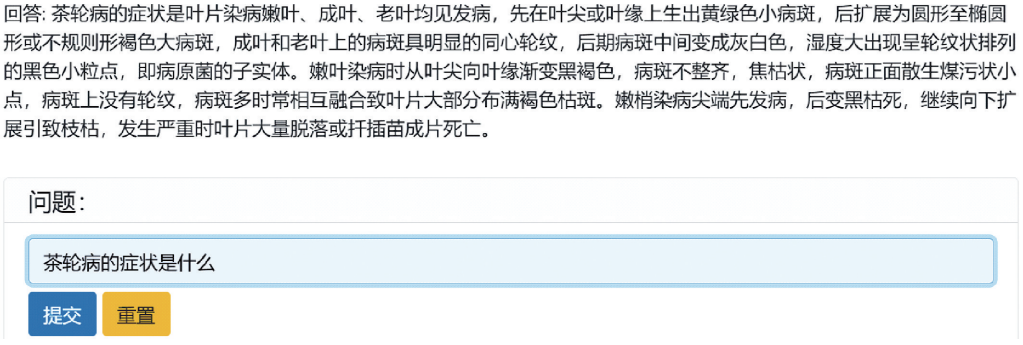


图 8　示例问答结果展示

Fig. 8　Example question and answer results display

4　结论

(1)提出了一种基于 BERT－BiLSTM－CRF 的茶树病虫害命名实体识别方法，利用 BERT 模型获取上下文相关的词向量表示，提升了茶树病虫害语义理解能力，通过 BiLSTM 模块融合正向与反向信息，深入挖掘茶树病虫害的深层语义特征，采用 CRF 层对标签序列进行全局优化，提升了茶树病虫害实体边界识别精度。同时，构建了一个覆盖病害、虫害、防治知识等多维属性的茶树病虫害标注语料库。所建模型对数据集命名实体识别精确率、召回率、F1 值达 98.42%、98.75% 和 98.58%，均达到 SOTA 水平。

(2)数据集 MSRANER 和 CLUENER2020 泛化性试验结果表明，BERT－BiLSTM－CRF 相较于其他对比模型，精确率、召回率及 F1 值均到达最优，F1 值分别为 92.4% 和 87.8%，充分验证了其对不同数据集的适应能力和泛化性能。

(3)构建了茶树病虫害知识图谱，并设计开发了智能问答系统。该成果不仅为垂直领域知识图谱构建与问答系统开发提供了有益的参考与实践案例，也为茶树种植者和农业技术人员提供了有力的决策支持，助力实现茶树病虫害精准防控与科学管理。

参 考 文 献

[1] PAN S, NIE Q, TAI H, et al. Tea and tea drinking: China's outstanding contributions to the mankind[J]. Chinese Medicine, 2022, 17(1): 27.

[2] XUE Z, XU R, BAI D, et al. YOLO-tea: a tea disease detection model improved by YOLOv5[J]. Forests, 2023, 14(2): 415.

[3] 蔡晓明, 边磊, 罗宗秀, 等. 2023 年茶树病虫害防控研究进展[J]. 中国茶叶, 2024, 46(10): 1 - 7.
CAI Xiaoming, BIAN Lei, LUO Zongxiu, et al. Research progress of tea pest control in 2023[J]. Chinese Tea, 2024, 46(10): 1 - 7. (in Chinese)

[4] 陈军, 李佳琪, 李静娴, 等. 茶树病虫害的发生与防治[J]. 河南林业科技, 2023, 43(4): 51 - 53.

[5] 曾德新, 朱林, 王抄抄, 等. 茶产业发展现状及常见茶叶病虫害防控研究进展[J]. 安徽农业科学, 2024, 52(15): 5 - 12.
ZENG Dexin, ZHU Lin, WANG Chaochao, et al. The current development status of the tea industry and research progress in the prevention and control of common tea diseases and pests[J]. Anhui Agricultural Sciences, 2024, 52(15): 5 - 12. (in Chinese)

[6] 余嘉伟. 茶树主要病虫害和综合防控技术的深入研究[J]. 种子世界, 2024(10): 228 - 230.

[7] 旭日. 茶叶种植中病虫害预警的数据挖掘技术分析[J]. 福建茶叶, 2016, 38(3): 28 - 29.

[8] FOUNTAS S, ESPEJO-GARCÍA B, KASIMATI A, et al. The future of digital agriculture: technologies and opportunities[J]. IT Professional, 2020, 22(1): 24 - 28.

[9] ADLI H, REMLI M, WONG K, et al. Recent advancements and challenges of AIoT application in smart agriculture: a review[J]. Sensors, 2023, 23: 3752.

[10] 张梓强. 智慧农业领域人工智能大模型的应用探索[J]. 中国农机装备, 2025(2): 94 - 96.
ZHANG Ziqiang. Application exploration of artificial intelligent large models in the field of smart agriculture[J]. China Agricultural Machinery and Equipment, 2025(2): 94 - 96. (in Chinese)

[11] 吴华瑞, 郭威, 邓颖, 等. 农业文本语义理解技术综述[J]. 农业机械学报, 2022, 53(5): 1 - 16.
WU Huarui, GUO Wei, DENG Ying, et al. Review of semantic analysis techniques of agricultural texts[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(5): 1 - 16. (in Chinese)

[12] 侯琛, 牛培宇. 农业知识图谱技术研究现状与展望[J]. 农业机械学报, 2024, 55(6): 1 - 17.
HOU Chen, NIU Peiyu. Review of research status and prospects of agricultural knowledge graphs[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(6): 1 - 17. (in Chinese)

[13] 张吉祥, 张祥森, 武长旭, 等. 知识图谱构建技术综述[J]. 计算机工程, 2022, 48(3): 23 - 37.
ZHANG Jixiang, ZHANG Xiangsen, WU Changxu, et al. Survey of knowledge graph construction techniques[J]. Computer Engineering, 2022, 48(3): 23 - 37. (in Chinese)

[14] HUANG Q, TAO Y, DING S, et al. An intelligent Q&A module for tea diseases and pests based on automatic knowledge graph construction[C]// IEEE International Workshop on Metrology for Agriculture and Forestry. IEEE, 2023: 66 - 70.

[15] ABU-SALIH B. Domain-specific knowledge graphs: a survey[J]. Journal of Network and Computer Applications, 2021, 185: 103076.

[16] PENG C, XIA F, NASERIPARSA M, et al. Knowledge graphs: opportunities and challenges[J]. Artificial Intelligence Review, 2023,56(11): 13071 - 13102.

[17] JI S, PAN S, CAMBRIA E, et al. A survey on knowledge graphs: representation, acquisition, and applications[J]. IEEE Transactions on Neural Networks and Learning Systems, 2021, 33(2): 494 - 514.

[18] KERAGHEL I, MORBIEU S, NADIF M. A survey on recent advances in named entity recognition[J]. arXiv preprint arXiv: 2401. 10825, 2024.

[19] MANSOURI A, AFFENDEY L S, MAMAT A. Named entity recognition approaches[J]. International Journal of Computer Science and Network Security, 2008, 8(2): 339 - 344.

[20] SEKINE S, NOBATA C. Definition, dictionaries and tagger for extended named entity hierarchy[C]// Proceedings of the Fourth International Conference on Language Resources and Evaluation, 2004: 1977 - 1980.

[21] 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.

[22] CHATTERJEE N, KAUSHIK 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: ICDECT, 2016: 511 - 522.

[23] BENDER O, OCH F J, NEY H. Maximum entropy models for named entity recognition[C]// Proceedings of the Seventh Conference on Natural Language Learning at HLT - NAACL, 2003: 148 - 151.

[24] WU Y C, FAN T K, LEE Y S, et al. Extracting named entities using support vector machines[C]// International Workshop on Knowledge Discovery in Life Science Literature, 2006: 91 - 103.

[25] SZARVAS G, FARKAS R, KOCSOR A. A multilingual named entity recognition system using boosting and c4.5 decision tree learning algorithms[C]// Discovery Science:9th International Conference, 2006: 267 - 278.

[26] 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]. 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]. Transactions of the Chinese Society for Agricultural Machinery，2017，48(Supp.)：178－185.(in Chinese)

[27] 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.

[28] 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，2022，34(1)：50－70.

[29] HUANG Z，XU W，YU K. Bidirectional LSTM－CRF models for sequence tagging[J]. arXiv preprint arXiv：1508. 01991，2015.

[30] PETERS M E，AMMAR W，BHAGAVATULA C，et al. Semi-supervised sequence tagging with bidirectional language models[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics，2017：1756－1765.

[31] YANO K. Neural disease named entity extraction with character-based BiLSTM + CRF in Japanese medical text[J]. arXiv preprint arXiv：1806. 03648，2018.

[32] AN Y，XIA X，CHEN X，et al. Chinese clinical named entity recognition via multi-head self-attention based BiLSTM－CRF[J]. Artificial Intelligence in Medicine，2022，127：102282.

[33] 蒲攀，张越，刘勇. Transformer 优化及其在苹果病虫命名实体识别中的应用[J]. 农业机械学报，2023，54(6)：264－271.
PU Pan，ZHANG Yue，LIU Yong. Transformer optimization and application in named entity recognition of apple diseases and pests[J]. Transactions of the Chinese Society for Agricultural Machinery，2023，54(6)：264－271.(in Chinese)

[34] WANG C，GAO J，RAO H，et al. Named entity recognition (NER) for Chinese agricultural diseases and pests based on discourse topic and attention mechanism[J]. Evolutionary Intelligence，2024，17(1)：457－466.

[35] CHEBBI A，KNIESEL G，ABDENNADHER N，et al. Enhancing named entity recognition for agricultural commodity monitoring with large language models[C]//Proceedings of the 4th Workshop on Machine Learning and Systems，2024：208－213.

[36] DE S，SANYAL D K，MUKHERJEE I. Fine-tuned encoder models with data augmentation beat ChatGPT in agricultural named entity recognition and relation extraction[J]. Expert Systems with Applications，2025，277：127126.

[37] LEVOW G A. The third international Chinese language processing bakeoff：word segmentation and named entity recognition[C]//Proceedings of the Fifth SIGHAN Workshop on Chinese Language Processing，2006：108－117.

[38] XU L，DONG Q，LIAO Y，et al. CLUENER2020：fine-grained named entity recognition dataset and benchmark for Chinese[J]. arXiv preprint arXiv：2001. 04351，2020.

[39] 袁立存，周俊，戈为溪，等. 基于多模态知识图谱的水稻施肥期判别方法[J]. 农业机械学报，2024，55(9)：163－173.
YUAN Licun，ZHOU Jun，GE Weixi，et al. Rice fertilization period discrimination method based on multi-modal knowledge graph[J]. Transactions of the Chinese Society for Agricultural Machinery，2024，55(9)：163－173.(in Chinese)

[40] 陈明，朱珏樟，席晓桃. 基于知识图谱的花卉病虫害知识管理方法[J]. 农业机械学报，2023，54(3)：291－300.
CHEN Ming，ZHU Juezhang，XI Xiaotao. Knowledge management method of flower diseases and pests based on knowledge graph[J]. Transactions of the Chinese Society for Agricultural Machinery，2023，54(3)：291－300.(in Chinese)