

基于曼哈顿注意力机制的水稻病虫害命名实体识别

路 阳¹ 高鹏飞¹ 李昊岩¹ 石业欣¹ 王 鹏^{2,3}

(1. 黑龙江八一农垦大学信息与电气工程学院, 大庆 163319;

2. 东北石油大学黑龙江省网络化与智能控制重点实验室, 大庆 163318;

3. 东北石油大学三亚海洋油气研究院, 三亚 572024)

摘要: 水稻病虫害信息多源自非结构化文本,具有嵌套实体密集、句式冗长、语法结构复杂的特点,使得现有命名实体识别方法难以充分识别相关实体。针对该问题,提出 AgriRoBERTa-BiLSTM-Man-CRF 模型。构建农业领域预训练语料库和水稻病虫害命名实体识别标注数据集,为模型训练提供高质量数据支撑;在农业领域语料上使用全词掩码策略二次预训练 RoBERTa,使其能够专注中文词语整体语义并学习水稻病虫害文本的语言模式,增强模型对中文水稻病虫害专业术语和语义捕捉能力;引入曼哈顿注意力机制,利用 L1 距离捕捉高维空间中稀疏特征并量化特征差异,精准聚焦关键上下文信息,进一步提升实体边界识别精度。试验结果表明,所提算法实体识别 F1 值达 90.69%,精确率为 87.87%,召回率为 93.71%。模型 F1 值较 BiLSTM-CRF、BiLSTM-Attention-CRF、BERT-BiLSTM-CRF 和 IDCNN-CRF 4 个常规模型分别提高 7.8、9.99、1.8、15.9 个百分点,可更有效地识别水稻病虫害文本中各类实体。

关键词: 命名实体识别;水稻病虫害;曼哈顿注意力机制;AgriRoBERTa-BiLSTM-Man-CRF

中图分类号: TP391.1 **文献标识码:** A **文章编号:** 1000-1298(2026)08-0289-10

OSID:



Named Entity Recognition for Rice Pest and Disease Based on Manhattan Attention Mechanism

LU Yang¹ GAO Pengfei¹ LI Haoyan¹ SHI Yexin¹ WANG Peng^{2,3}

(1. College of Information and Electrical Engineering, Heilongjiang Bayi Agricultural University, Daqing 163319, China

2. Heilongjiang Provincial Key Laboratory of Networking and Intelligent Control, Northeast Petroleum University, Daqing 163318, China

3. Sanya Research Institute of Offshore Oil and Gas, Northeast Petroleum University, Sanya 572024, China)

Abstract: Rice pest and disease information mostly originates from unstructured text. These texts contain densely nested entities, lengthy sentences, and complex grammatical structures. Because of this, current named entity recognition (NER) methods struggle to fully identify the relevant entities. To solve this problem, the AgriRoBERTa-BiLSTM-Man-CRF model was proposed. Firstly, a pre-trained corpus in the agricultural domain and a labeled dataset for rice pest and disease named entity recognition were constructed. This provided high-quality data for model training. Secondly, pre-training RoBERTa on agricultural texts was continued by using whole-word masking. This approach enabled the model to focus on the complete meaning of Chinese words and learn the specific language patterns found in texts about rice diseases and pests. Finally, Manhattan attention mechanism was introduced to capture sparse features in high-dimensional space by using L1-distance. This approach quantified feature differences while precisely focusing on critical contextual information, so as to improve the accuracy of entity boundary recognition. Experimental results showed that the proposed algorithm achieved an F1 score of 90.69%, with a precision of 87.87% and a recall of 93.71% for entity recognition. The F1 score was 7.8, 9.99, 1.8, 15.9 percentage points higher than that of four conventional models: BiLSTM-CRF, BiLSTM-Attention-CRF, BERT-BiLSTM-CRF and IDCNN-CRF. This enhanced performance enabled more effective recognition of diverse entities in rice pest and disease texts. This significant

收稿日期: 2025-07-15 修回日期: 2025-10-10

基金项目: 国家自然科学基金项目(U21A2019/62476081)、黑龙江省自然科学基金联合引导项目(LH2024F048)和黑龙江省省属高等学校基本科研业务费科研项目(ZRCQC202403)

作者简介: 路阳(1976—),男,教授,博士生导师,主要从事模式识别与机器学习研究,E-mail: luyanga@sina.com

improvement indicated that the model can recognize various entities in rice pest and disease texts more effectively.

Key words: named entity recognition; rice diseases and pests; Manhattan attention mechanism; AgriRoBERTa – BiLSTM – Man – CRF

0 引言

水稻病虫害相关知识主要存在于田间生产记录、会议报告、专著和学术论文等非结构化文本^[1]中,农业工作者难以快速准确地从中获取信息。因此,如何高效地从海量农业数据中挖掘病虫害防治关键信息,已成为当前研究与实践的重点。

命名实体识别(Named entity recognition,NER)作为自然语言处理的核心技术^[2],可从非结构化文本中提取具有特定意义的实体,在农业、医学及金融等领域^[3]具有广阔应用前景。水稻病虫害研究中,NER能识别病害、虫害和防治措施等实体,并基于提取的实体构建水稻病虫害知识图谱^[4]。通过图结构展示实体及其关系^[5],不仅提升信息检索效率,还为精准农业的病虫害防控决策提供支持。

早期NER主要依赖人工规则。如RAU^[6]结合启发式规则和列表从金融新闻中提取公司名称,精度超95%;方晓珊等^[7]提出基于实例学习的词典-句法规则自动抽取方法,方法虽简单直观,但推广性差,需不断更新规则。21世纪初,机器学习方法兴起。BIKEL等^[8]提出基于隐马尔可夫模型的IdentiFinder系统,在多语言中识别实体效果较好;CURRAN等^[9]用最大熵标记器整合多种信息实现高精度实体识别;LAFFERTY等^[10]提出条件随机场用于序列数据分割和标记。此类方法依赖大量标注数据,且对长距离依赖和复杂语言处理有限。

深度学习的快速发展为NER带来新方向。COLLOBERT等^[11]提出结合多任务学习和半监督学习的卷积神经网络(CNN),用于多种语言处理任务,错误率降低约2%。为解决CNN存在的长距离依赖问题,研究转向循环神经网络(RNN)。HAMMERTON^[12]将长短时记忆网络(LSTM)应用于NER,通过2轮处理提升性能,缓解了传统RNN梯度消失的问题。HUANG等^[13]结合双向LSTM捕捉上下文及CRF优化标签序列,显著提升NER性能。然而,序列模型的线性处理机制在建模全局上下文关联时,难以动态聚焦关键语义单元。VASWANI等^[14]于2017年提出的完全基于自注意力机制的Transformer架构,其多头注意力层可并行捕获不同子空间的语义特征。这种全局感知能力为NER带来革新,尤其改善了嵌套实体识别。

基于自注意力机制的突破,预训练语言模型^[15]在NER领域展现出巨大潜力。PETERS等^[16]提出的ELMo通过双向语言模型预训练获取上下文敏感词表示;DEVLIN等^[17]提出的BERT模型采用掩码语言建模和下一句预测任务进行预训练,进一步提升上下文理解能力。融合Transformer与LSTM优势的BERT – BiLSTM – CRF^[18]成为NER主流方法。其核心包括:BERT生成上下文词向量捕捉语义,BiLSTM建模序列时序依赖增强理解,CRF优化标签序列全局一致性。但在水稻病虫害NER任务中,该模型存在不足:①缺乏领域知识深度融合,对专业术语语义理解不足。②相似实体边界容易混淆,难以精准区分。

针对上述不足,提出AgriRoBERTa – BiLSTM – Man – CRF模型,通过领域适配和实体区分提升性能。领域适配方面,以RoBERTa为基础,利用水稻病虫害语料微调,通过掩码语言建模和任务优化增强术语语义表征。实体区分方面,在BiLSTM后引入曼哈顿注意力机制,通过计算令牌间曼哈顿距离突出边界特征,并动态调整注意力权重,以降低相似实体混淆概率。

1 材料与方法

1.1 数据集获取

农业数据,特别是水稻病虫害领域数据,具有高度专业性与特殊性。目前,该领域缺乏大规模、高质量且公开可用的数据集,导致水稻病虫害领域命名实体识别研究及数据支撑不足。因此,构建了2个关键的数据集,一个用于模型增强的预训练语料库,另一个专门用于水稻病虫害NER数据集。

1.1.1 模型预训练语料数据集

为使模型更加理解水稻病虫害领域语义,将语料数据源聚焦于相关专业书籍,广泛搜集相关资料,构建一个内容全面的水稻病虫害领域预训练语料库。

在数据采集过程中,依托图书资源平台,围绕水稻病虫害管理、水稻病虫害防治、水稻病虫害识别等关键词展开搜索筛选。从大量书籍中选取19本相关性强的书籍作为语料主要来源,内容包含水稻病虫害种类、分布、识别和防治等。具体来源如表1所示。

表 1 水稻病虫害语料信息来源

Tab.1 Source information of rice pest and disease corpus

书名	主编	出版社	出版年份
水稻病虫害害彩色图鉴	王玉山	辽宁科学技术出版社	2013
水稻病虫害	工蜂	安徽科学技术出版社	1982
水稻病虫害及防治原色图册	傅强	金盾出版社	2009
水稻病虫害及其防治技术	魏松红	辽宁科学技术出版社	2020
水稻病虫害识别与防治	王琦	宁夏人民出版社	2009
水稻病虫害识别与绿色防控图谱	彭红	河南科学技术出版社	2020
水稻病虫害田间手册-病虫害鉴别与抗性鉴定	陶荣祥	中国农业科学技术出版社	2006
水稻病虫害原色图谱及其诊治技术	陈福如	中国农业科学技术出版社	2012
水稻主要病虫害防控关键技术解析	黄世文	金盾出版社	2010
北方水稻病虫害综合防治	盛广华	中国农业出版社	2008
大田作物病虫害防治图谱-水稻	孙艳梅	吉林出版社	2010
宁夏水稻主要病虫害识别与防治	杨明进	阳光出版社	2018
水稻病虫害鼠害安全用药问答	耿继光	安徽科学技术出版社	2008
水稻病虫害防治图解	董伟	化学工业出版社	2014
水稻病虫害绿色防治	张永礼	吉林人民出版社	2016
水稻病虫害实用防控技术	黄炳超	中国书籍出版社	2011
水稻病虫害诊断与防治原色图谱	傅强	金盾出版社	2005
水稻主要病虫害防控关键技术解析	黄世文	金盾出版社	2010
优质水稻生产与病虫害防治	中央电视台农广天地栏目	上海科学技术文献出版社	2009

获取源书籍数据后,利用 MinerU^[19]工具抽取书籍文本,并统一转换为 txt 格式存储。鉴于部分书籍因年代久远、印刷质量不佳,在 OCR 识别过程中易出现文字错误、乱码等问题,开展了人工核查工作。针对重复、相似的错误予以记录,对其他错误则直接修正。完成上述步骤后,进行文本清洗。通过正则表达式,批量纠正前期记录的重复性错误,并去除因格式转换产生的冗余字符、特殊标记,以及与书籍正

文无关的前言、目录、版权页等内容。同时,借助 jieba 工具对语料进行句子分割。最后,以句子为单位对数据执行去重操作,确保数据的唯一性。

经清洗,构建模型预训练语料数据集。数据集总共约 1.77×10^6 字,包含 47 472 条语料,涵盖水稻病虫害领域专业术语和知识体系。

1.1.2 水稻 NER 标注数据集

针对水稻病虫害命名实体识别任务,主要通过网络爬取获得源数据。利用 Python 的 Scrapy 框架,从百科、农业相关新闻报道及科研论文等资料中,针对性地爬取与水稻病虫害相关的文本信息。数据涵盖水稻病虫害的名称、症状描述、防治方法等内容,具有较高的领域相关性和专业性。对原始文本使用正则表达式进行清洗,去除 HTML 标签、特殊字符处理、断句规范化等,最终形成未标注文本数据包含句子 1 770 条,共 105 295 字。

NER 任务中,需根据预先定义实体进行类别划分。基于水稻病虫害领域知识体系,根据专家指导并结合任务需求,最终定义病害、虫害、寄主等 11 类实体。根据定义的实体类型,使用开源工具 Label-studio 进行手工实体标注。数据集格式遵循 BIOES^[20]规范进行转换,该格式能更精确描述实体边界,尤其适用于中文中空格分隔的连续字符序列,标签定义如表 2 所示。

表 2 BIOES 标签定义
Tab.2 BIOES tag definition

标签	定义
B-XXX	X 实体起始位置 (Begin)
I-XXX	X 实体内部位置 (Inside)
E-XXX	X 实体结束位置 (End)
S-X	单字符实体 X (Single)
O	非实体位置

标注后数据集包含 3 572 条标注文本,总字数 89 456 字,标注实体实例 6 192 个。水稻病虫害实体类别如表 3 所示。

1.1.3 公开数据集

在自然语言处理研究中,跨领域数据集的引入对于评估模型的泛化能力、验证领域适应性以及挖掘知识迁移潜力具有重要意义。如表 4 所示,引入 3 个其他领域公开数据集:Resume 数据集源自简历摘要数据,实体包括 CONT(国籍)、EDU(教育背景)、LOC(地名)、NAME(人名)、ORG(组织)、PRO(专业)、RACE(民族)、TITLE(职称);CCKS2019 数据集源自中文临床文本处理任务,实体包括解剖部位、疾病和诊断、药物、手术、影像检查、实验室检验;Bank 数据集源自银行客户服务评价文本,实体包含

表 3 水稻病虫害实体类别

Tab.3 Rice pests and diseases entity categories				
实体名	实体标签	实体数量	实体描述	实体举例
病害	Disease	480	水病害名称	稻瘟病、纹枯病
虫害	Pest	431	水稻虫害名称	水稻线螨、稻瘰蚊
寄主	Host	757	寄主植物	水稻、小麦、玉
病原	Pathogen	187	致病微生物类别	镰刀菌、稻瘟病菌
学名	Sci_Name	152	病原拉丁学名	<i>Magnaporthe</i>
地区	Geo_Region	833	病害高发区	长江流域、华南地区、黑龙江
症状	Symptom	1598	病虫害致病后的特征	白穗、稍褪绿、失水、枯心
防治药剂	Chemical	680	防治病虫害药剂名称	氯菊酯乳油、敌百虫
防治方法	Agri_Control	794	农业防治措施	灯光诱杀、天敌防治、清除杂草
分类	Taxonomy	156	病原分类学信息	界门纲目科属种
别名	Name	124	病虫害的俗称	红虫、白尖病、铁锈病

表 4 公开数据集
Tab.4 Public datasets

数据集	实体类型	实体数量
Resume	CONT、EDU、LOC、NAME、ORG、	16 461
	PRO、RACE、TITLE	
CCKS2019	解剖部位、疾病和诊断、药物、手术、影像检查、实验室检验	23 655
Bank	BANK、PRODUCT、COMMENTS_N、	16 593
	COMMENTS_ADJ	

BANK（银行机构）、PRODUCT（金融产品）、COMMENTS_N（名词性评价）与 COMMENTS_ADJ（形容词性评价）。通过领域差异揭示模型在知识迁移、实体泛化等方面的优势与局限。

1.2 水稻病虫害命名实体识别模型建立

基于经典 BiLSTM – CRF 架构,引入领域预训练

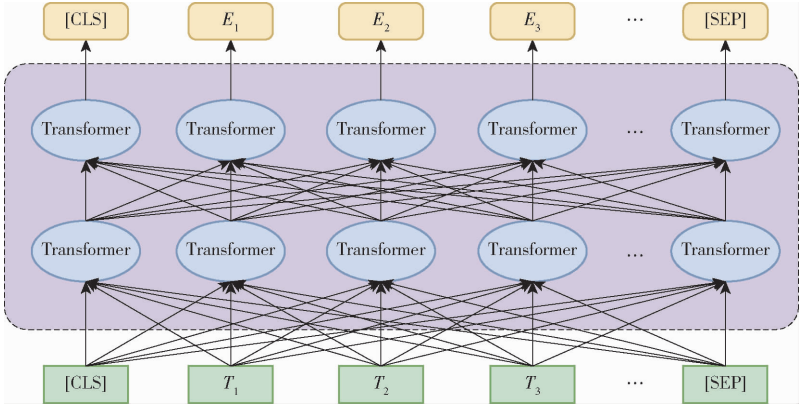


图 2 RoBERTa 结构图

Fig.2 RoBERTa structure

AgriRoBERTa 模块和曼哈顿注意力机制,构建了针对水稻病虫害命名实体识别的增强模型。模型核心模块构成包括:AgriRoBERTa 负责生成词嵌入表示,BiLSTM 用于捕捉长距离序列依赖,曼哈顿注意力加强对关键上下文特征的聚焦,CRF 则用于标签序列的全局优化。整体架构如图 1 所示,其中 h 表示语义特征向量。

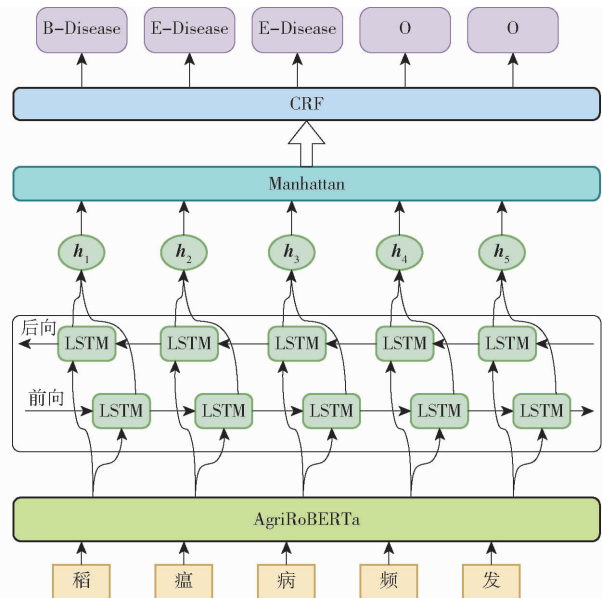


图 1 模型架构图

Fig.1 Model framework

1.2.1 AgriRoBERTa

RoBERTa (Robustly optimized BERT approach)^[21] 是基于 Transformer 架构的预训练语言模型,由 Facebook AI 于 2019 年提出,是 BERT (Bidirectional encoder representations from Transformers) 的深度优化版本。RoBERTa 继承了 BERT 的结构,由多层 Transformer 编码器堆叠组成。其结构如图 2 所示, T 表示模型编码器输入单元, E 表示模型解码器输出单元,每层编码器层包含多头自注意力机制和前馈神经网络。这种结构在模型处理文本时,能更有效地汇聚不同位置单词的信息,从

而更精确地掌握上下文含义。与基础的 BERT 对比, RoBERTa 做了一些关键改良: 每次输入训练数据时, 都重新随机实施掩码步骤, 增加了数据的多样性并提升模型的泛化能力; 其次, 其将 BERT 中“下一句预测”任务移除, 使模型更聚焦核心的掩码语言建模任务; 最后, RoBERTa 扩大了数据集并增加训练轮次。

RoBERTa 的训练数据主要是英文, 因此在掩码任务中采取 WordPiece^[22]方法进行分词。这种分词法基于英语单词边界明确和词形变化规则的特性设计。其使用贪心算法选择高频子词单元, 能很好地解决英语里的词汇变形和未登录词 (Out-of-vocabulary, OOV) 问题。例如, 对于“running”可分解为“run”和“##ning”。然而, 当处理中文文本时就存在着很明显的局限性。中文词汇之间没有明确的分隔, 语义通常依赖于整个词体现, 而不是单个汉字或子词单元。使用 WordPiece 方法, 一个完整的词可能会被分割为多个子词单元。例如, “水稻”可能被分解为“水”和“稻”, 从而造成语义信息的丢失。

为解决这一问题, 全词掩码 (Whole word masking, WWM) 策略被提出。WWM 的核心思想是将整个词汇 (非单个字符或子词单元) 进行 Masking, 从而确保模型在预训练阶段能够学习到词汇级别的语义信息, 实现方式如图 3 所示。这种方法能够有效避免词汇被拆分后导致的语义丢失问题, 保留了词汇完整语义, 还增强了模型对上下文依赖关系的建模能力。



图 3 全词掩码示例
Fig. 3 Example of WWM

为适配水稻病虫害 NER 任务, 在通用 RoBERTa - WWM 基础上构建 AgriRoBERTa。AgriRoBERTa 采用标准的 12 层 Transformer 编码器架构, 每层包含 768 维隐藏状态和 12 个注意力头, 模型使用自建水稻病虫害领域语料库进行预训练。针对领域术语密集、易被拆分导致语义不完整的问题, 采用“领域术语级掩码”策略, 将如“稻瘟病”、“稻飞虱”等完整术语作为掩码单元, 替代 WWM 的通用词汇掩码, 以增强术语语义一致性。

在参数更新方面采用分层优化策略: 冻结底层编码器 (Encoder 0 ~ 5, 共 6 层), 使其参数在训练过程保持不变, 从而保留模型在大规模通用语料上获得的基础语法和语义特征; 仅微调上层编码器

(Encoder 6 ~ 11, 共 6 层) 和词嵌入层, 允许这些层的参数在训练中更新, 以便更好地适应水稻病虫害领域的语义模式。这样的设计能够有效降低计算冗余, 同时提升模型对领域特征的捕捉能力。

此外, 在词嵌入层引入了水稻病虫害核心术语词典, 为每个术语分配独立可学习的初始向量并通过预训练进行动态优化。这一机制有助于模型精准捕捉“稻瘟病-稻瘟病菌”等领域术语间的语义关联。整个预训练过程以掩码语言建模为核心任务, 使模型在无监督条件下自主学习领域知识, 从而为 NER 任务提供高质量的领域化词嵌入。模型结构如图 4 所示。

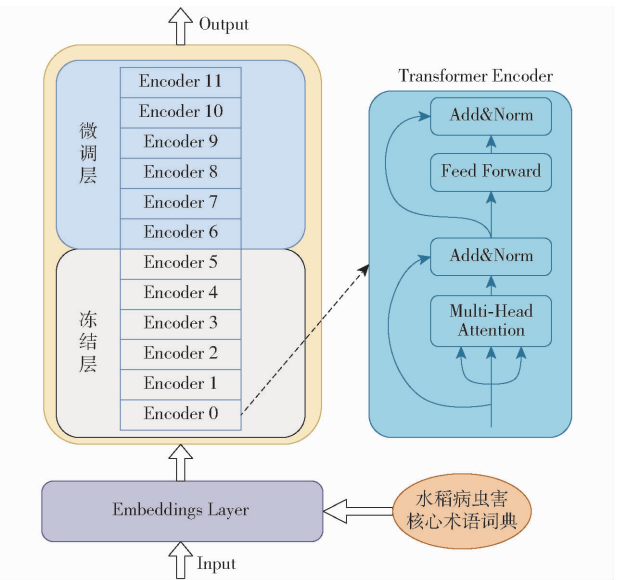


图 4 AgriRoBERTa 结构示意图

Fig. 4 Schematic of AgriRoBERTa architecture

1.2.2 BiLSTM

双向长短期记忆网络 (BiLSTM) 通过结合前向和后向 2 个 LSTM 单元^[23], 分别从序列的开头和结尾处理信息, 从而捕捉序列中的双向上下文依赖。LSTM 主要是通过门控机制解决传统 RNN 梯度消失问题, 核心包括输入门、遗忘门、输出门和记忆单元, 结构如图 5 所示。

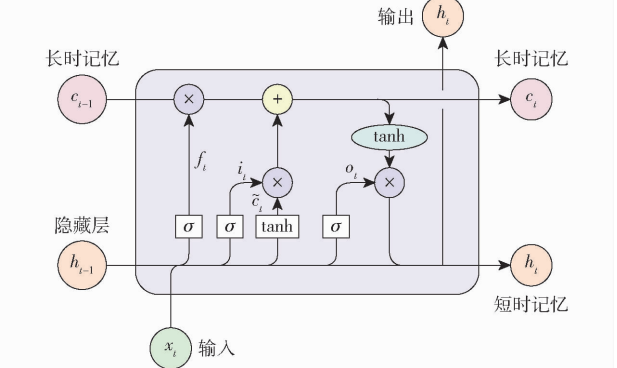


图 5 LSTM 结构图

Fig. 5 LSTM structure

LSTM 单元在时间步 t 详细计算过程为遗忘门

$$f_t = \sigma(\mathbf{W}_f[h_{t-1}, x_t] + b_f) \tag{1}$$

输入门

$$i_t = \sigma(\mathbf{W}_i[h_{t-1}, x_t] + b_i) \tag{2}$$

候选记忆

$$\tilde{c}_t = \tanh(\mathbf{W}_c[h_{t-1}, x_t] + b_c) \tag{3}$$

记忆更新

$$c_t = f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \tag{4}$$

输出门

$$o_t = \sigma(\mathbf{W}_o[h_{t-1}, x_t] + b_o) \tag{5}$$

隐藏状态

$$h_t = o_t \odot \tanh(c_t) \tag{6}$$

式中 $\mathbf{W}_f, \mathbf{W}_i, \mathbf{W}_c, \mathbf{W}_o$ ——权重矩阵

b_f, b_i, b_c, b_o ——偏置

σ ——Sigmoid 激活函数

$\tanh$ ——双曲正切激活函数

x_t ——当前时间步输入

h_{t-1} ——上一时间步隐藏状态

c_t ——当前时间步记忆单元状态

c_{t-1} ——上一时间步记忆单元状态

遗忘门决定上一时刻记忆中哪些信息需要丢弃;输入门和候选记忆共同确定当前输入中哪些信息要加入记忆;记忆更新操作将遗忘门和输入门的结果结合起来更新当前记忆;输出门则控制当前记忆中有多少信息会输出到隐藏状态。这种门控机制使得 LSTM 能够根据输入动态地调整记忆内容,从而有效处理长序列文本。

BiLSTM 模块接收 AgriRoBERTa 输出的嵌入序列 $\mathbf{E} = [e_1 \ e_2 \ \cdots \ e_n]$,并针对每个位置 i ,生成隐藏状态 $\mathbf{h}_i$,其由前向向量 $\vec{h}_i$ 和后向向量 $\overleftarrow{h}_i$ 拼接而成,且 $\mathbf{h}_i \in \mathbf{R}^{2d_h}$,其中 $\vec{h}_i$ 源自前向 LSTM, $\overleftarrow{h}_i$ 来自后向 LSTM。双向结构可强化局部上下文的细粒度建模,弥补 AgriRoBERTa 对局部实体边界捕捉不足的问题,为后续曼哈顿注意力机制提供更精准的序列特征输入。

1.2.3 曼哈顿注意力机制

曼哈顿距离(L1 距离)用于计算输入向量 $\mathbf{x}$ 与参考向量 $\mathbf{u}$ 之间的距离,计算公式为

$$d_{\text{manhattan}} = \sum_{i=1}^D |x_i - u_i| \tag{7}$$

式中 x_i, u_i ——向量第 i 个维度上的值

D ——向量维度

在水稻病虫害 NER 任务中,需要处理高维长序列文本,并应对标注中存在术语边界偏差的情况。

常见的欧氏距离(L2)计算需进行额外平方与开方操作,平方会放大异常值导致注意力偏移;切比雪夫距离仅聚焦单维度最大差异,无法捕捉多维度协同语义;余弦相似度则主要关注向量方向而非特征值大小,难以区分“症状程度”等关键量化差异。相比之下,曼哈顿距离通过坐标差绝对值的累积抑制了异常值的过度影响,提供更稳健的语义差异度量。同时,其跨维度累积差异的特性契合中文实体局部聚集的特点。

基于此,以曼哈顿距离为基础并使用可学习的维度权重参数构建注意力机制。原始曼哈顿距离隐含所有特征维度同等重要的假设,但在农业 NER 任务中,不同维度对任务的重要性不同。引入维度权重 ω ,使模型能够自主学习不同特征维度的重要性分布,即

$$d_w = \sum_{i=1}^D \omega_i |x_i - u_i| \tag{8}$$

其次设计自适应温度调节系统,引入自适应温度参数 τ ,通过一个轻量级神经网络动态控制注意力分布的锐度,即

$$\tau = \tau_{\text{base}} (1 + \alpha f(u_x)) \tag{9}$$

$$d_{\text{scaled}} = \frac{d_w}{\tau} \tag{10}$$

式中 τ_{base} ——基础温度

α ——调节因子

u_x ——输入特征

d_{scaled} ——缩放后的注意力分布

温度 τ 是由基础温度 τ_{base} 和动态调节因子 $\alpha f(u_x)$ 组成,其中 $f(u_x)$ 是一个基于输入特征 u_x 的神经网络,输出范围为 $[0, 1]$ 。

传统注意力机制^[24]原理为输入序列里的每个元素赋予权重,以此动态聚焦对任务最为关键的部分。在水稻病虫害 NER 中存在两大局限:①依赖词向量线性相似性,难以建模“水稻稻瘟病-分生孢子”等非局部语义关联。②计算复杂度 $O(n^2)$,处理长序列(如病虫害症状描述)时易引入噪声,且 Softmax 归一化会导致注意力分散。

针对传统注意力机制的局限,引入曼哈顿注意力机制,其在中文序列标注任务中具有以下优势:①L1 范数对局部差异的敏感性,使得更适配中文实体局部聚集特性。②低复杂度的距离运算降低计算资源消耗。③通过跨维度累积差异,能更好地建模复杂句式语义关联并捕捉非线性关系。

1.2.4 CRF

条件随机场(Conditional random field, CRF)是一种可针对序列标注任务使用的概率图模型,核心

优势在于能对标签序列进行联合建模,保证预测标签的全局合理性与一致性。在水稻病虫害 NER 中,标签间存在依赖关系(例如“B - Pest”后不可接“I - Disease”)。CRF 层通过学习标签间的转移概率,可以有效捕捉此类约束规则,从而降低实体类别误判风险。

具体来讲,对于标签序列 $Y = \{y_1, y_2, \cdots, y_n\}$,其得分定义为

$$s(X, Y) = \sum_{i=1}^n e_{i, y_i} + \sum_{i=1}^{n-1} T_{y_i, y_{i+1}} \tag{11}$$

式中 e_{i, y_i} ——位置 i 发射得分
 $T_{y_i, y_{i+1}}$ —— y_i 到 y_{i+1} 转移得分

标签序列概率为

$$P(Y|X) = \frac{\exp(s(X, Y))}{\sum_{Y'} \exp(s(X, Y'))} \tag{12}$$

2 试验

2.1 试验平台及指标

试验均在 Ubuntu 23.04 系统平台上运行,CPU 为双 Intel(R) Xeon(R) Gold 6326 CPU@2.90 GHz, GPU 是 NVIDIA A100 80 GB,运行内存 64 GB,采用 Python 3.8 编写代码,深度学习框架使用 Pytorch 2.0。涉及的训练参数设置如表 5 所示。

表 5 模型参数设置
Tab.5 Model parameters

参数	数值/形式
初始学习率	3×10^{-5}
优化器	AdamW
批大小	32
训练轮次	100
最大序列长度	256

为全面评估模型性能,采用在 NER 任务中常用的 3 个指标:精确率(Precision)、召回率(Recall)和 F1 值(F1 score)。引入浮点运算量和参数量 2 个指标衡量模型计算复杂度,前者反映模型单次前向传播的运算量,后者统计模型可学习参数总数,二者共同体现模型资源消耗与轻量化水平。

2.2 试验结果分析

为验证所提模型有效性,在自建水稻病虫害 NER 数据集上进行对比试验。选取 NER 领域 4 种具有代表性的模型:BiLSTM - CRF^[25]、BiLSTM - Attention - CRF^[26]、BERT - BiLSTM - CRF、IDCNN - CRF^[27]。训练过程模型 F1 值变化曲线如图 6 所示,所有模型训练结果如表 6 所示。

由图 6 可知,AgriRoBERTa - BiLSTM - Man - CRF 模型收敛速度极快,在第 15 轮次后即趋于稳定

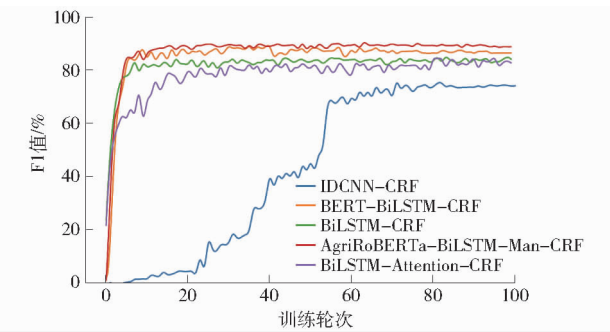


图 6 模型训练 F1 值曲线
Fig.6 Model training F1 curves

表 6 模型训练结果
Tab.6 Model training results

模型	精确率/%	召回率/%	F1 值/%	参数量	浮点运算量
BiLSTM - CRF	83.63	82.17	82.89	4.13×10^6	1.36×10^9
IDCNN - CRF	73.40	76.22	74.79	5.76×10^6	1.47×10^9
BiLSTM - Attention - CRF	80.28	81.12	80.70	4.33×10^6	1.43×10^9
BERT - BiLSTM - CRF	85.71	92.31	88.89	1.12×10^8	2.35×10^{10}
AgriRoBERTa - BiLSTM - Man - CRF (本文)	87.87	93.71	90.69	1.04×10^8	2.36×10^{10}

状态且波动较小,这表明其具备高效的关键特征捕获能力和优异的训练稳定性,有效规避了过拟合风险。BERT - BiLSTM - CRF 模型前期增长较快但波动较大,经过 60 轮次才逐渐稳定,这显示出模型学习能力尚可但对参数调整敏感、稳定性不足。BiLSTM - CRF 和 BiLSTM - Attention - CRF 模型收敛缓慢且训练后期仍有波动,学习效率欠佳,即使后者引入传统注意力机制也未有效改善。表现最弱的是 IDCNN - CRF 模型,其收敛速度最慢且 F1 值最低,因其卷积操作难依赖建模文本序列。

由表 6 可知,AgriRoBERTa - BiLSTM - Man - CRF 模型在各指标均取得最佳表现:精确率为 87.87%,召回率为 93.71%,F1 值为 90.69%。基于预训练语言模型 BERT - BiLSTM - CRF 与本文模型在参数量级上远超其他传统模型。为排除参数量干扰,后续分析将分为 2 个层面:首先与轻量级模型对比,以证明预训练模型相对于传统方法的整体优势;其次与同量级模型进行公平对比,以验证本文模型结构设计的有效性。

(1)与 BiLSTM - CRF、BiLSTM - Attention - CRF 及 IDCNN - CRF 等轻量级模型相比,本文模型在性能上展现出显著优势。这些轻量级模型虽参数规模较小,但受限于结构瓶颈,性能难以进一步提高。IDCNN - CRF 的 F1 值仅为 74.79%,表现最差,其

卷积结构不擅长处理具有强序列依赖性的文本,无法准确识别复杂的实体边界。BiLSTM – CRF 的 F1 值为 82. 89%,虽能捕捉序列信息,但由于缺乏领域先验知识,对专业术语和复杂语义结构的识别能力较弱,召回率仅为 82. 17%。而在 BiLSTM – CRF 中引入传统注意力机制后,模型性能反而下降,F1 值仅为 80. 70%。这表明传统注意力机制在农业文本中不仅未能强化对关键实体的关注,还引入了噪声干扰。综上,仅依靠传统序列模型难以胜任专业领域命名实体识别任务,而引入领域预训练模型是提升性能的关键。

(2) 与参数量和浮点运算量相近的 BERT – BiLSTM – CRF 模型对比,进一步凸显本文模型的结构优势。结果显示,在参数量减少约 7% 且计算量相当的情况下,本文模型 F1 值仍达到 90. 69%,较对比模型(88. 89%)高 1. 8 个百分点。这证明性能提升并非依赖参数堆叠,而是源于更优秀的架构设计。具体而言,BERT 通过海量通用语料学习

了强大的语义表征,但因缺乏领域知识导致专业术语误判,AgriRoBERTa 通过领域预训练弥补了这一短板。同时,曼哈顿注意力机制的引入相比于 BiLSTM 与 CRF 的简单组合,能更有效地融合特征、抑制噪声。

通过对动态训练过程和静态训练结果分析,验证了所提模型在水稻病虫害 NER 任务中具有更优的性能。这主要得益于多模块协同:领域预训练模型 AgriRoBERTa 提供精准语义表征,解决通用模型领域知识缺失的问题;BiLSTM 细化局部关联以明确实体边界;曼哈顿注意力抑制噪声以提升准确率;CRF 避免标签逻辑矛盾以保障全局一致性。

2.3 具体实体识别性能分析

为进一步验证所提模型的实体级识别优势,表 7 为本文模型与 4 个对比模型在 11 类实体上的精确率、召回率及 F1 值。整体来看,本文模型在绝大多数实体类型上性能均优于其他对比模型,且不同实体类别间的识别稳定性更强。

表 7 各实体类别训练结果
Tab. 7 Training results of each entity category %

实体类型	AgriRoBERTa – BiLSTM – Man – CRF			BiLSTM – Attention – CRF			BiLSTM – CRF			IDCNN – CRF			BERT – BiLSTM – CRF		
	精确率	召回率	F1 值	精确率	召回率	F1 值	精确率	召回率	F1 值	精确率	召回率	F1 值	精确率	召回率	F1 值
Disease	91. 7	91. 0	91. 4	72. 7	78. 7	75. 6	84. 9	84. 2	84. 5	81. 4	68. 0	74. 1	88. 4	87. 7	88. 1
Pest	90. 4	95. 0	92. 6	74. 5	73. 7	74. 1	84. 5	87. 9	86. 1	84. 4	76. 8	80. 4	87. 6	92. 9	90. 2
Host	89. 0	95. 2	92. 0	76. 2	84. 9	80. 3	90. 9	87. 4	89. 1	86. 6	83. 7	85. 1	92. 1	91. 3	91. 7
Pathogen	98. 5	100	99. 3	95. 1	98. 7	96. 9	98. 5	99. 5	99. 0	98. 5	97. 0	97. 7	99. 0	99. 5	99. 3
Sci_Name	96. 5	99. 1	97. 8	85. 8	93. 9	89. 7	94. 4	97. 3	95. 8	95. 9	94. 2	95. 1	96. 1	98. 7	97. 4
Geo_Region	96. 1	97. 9	97. 0	85. 4	89. 2	87. 3	88. 4	92. 4	90. 4	92. 6	90. 6	91. 6	93. 5	95. 8	94. 6
Symptom	66. 2	65. 4	65. 8	40. 1	47. 2	43. 4	46. 0	52. 8	49. 2	56. 8	28. 0	37. 5	52. 4	61. 0	56. 4
Chemical	96. 8	100	98. 4	93. 5	96. 7	95. 1	98. 4	100	99. 2	100	93. 3	96. 6	98. 3	96. 7	97. 5
Agri_Control	85. 2	95. 8	90. 2	81. 8	81. 8	81. 8	92. 0	95. 8	93. 9	77. 8	87. 5	82. 4	95. 8	95. 8	95. 8
Taxonomy	97. 6	99. 5	98. 6	95. 7	98. 1	96. 9	97. 6	99. 0	98. 3	99. 0	93. 2	96. 0	98. 1	99. 0	98. 6
Name	93. 8	100	96. 8	87. 5	93. 3	90. 3	93. 8	100	96. 8	100	73. 3	84. 6	93. 8	100	96. 8

本文模型在 Chemical、Pathogen、Sci_Name 和 Taxonomy 4 类核心专业实体类型上表现尤为突出,F1 值分别达到 98. 4%、99. 3%、97. 8% 和 98. 6%,精确率和召回率均保持在 96% 以上。这一优异性能主要源于 2 方面:AgriRoBERTa 在农业语料上的预训练赋予了模型强大的专业术语和分类层级信息提取能力;曼哈顿注意力机制有效聚焦关键上下文信息。两者共同确保了模型对科学命名规则和领域特定术语的高精度识别。

除核心实体外,模型在其他实体类型上的性能表现同样稳健。Disease、Geo_Region、Host、Name、Pest 和 Agri_Control F1 值分别达到 91. 4%、97. 0%、92. 0%、96. 8%、92. 6% 和 90. 2%,反映了模型对常

规实体边界的准确划分能力。与常规实体的表现不同,所有模型在 Symptom 实体上表现相对较低,原因在于症状描述的语义复杂性、特征模糊性以及训练数据中样本分布不均。即便如此,所提模型较对比模型 F1 值提升 9. 4 ~ 28. 3 个百分点,进一步印证其整体性能的优越性。

为全面评估所提模型,对测试集中错误样本进行深入分析,尤其聚焦于识别效果相对较弱的实体类型。分析发现,Symptom 类实体的识别是目前模型主要短板,其精确率与召回率分别为 66. 2% 和 65. 4%,显著低于其他类别。此类错误中,误报现象主要源于症状表述与常规农艺描述间的语义重叠,例如“叶片营养性黄化”在非病害语境下易被误判

为症状实体。漏报现象则常见于复杂或复合型症状表述,如“叶缘出现水浸状褪绿环带”这类长实体,模型因难以处理其边界和长距离依赖关系而无法完整捕获。该问题与训练数据中 Symptom 样本分布不均及语义多样性不足有直接关联。相比之下,Agri_Control 与 Host 等实体识别效果较为稳定,F1 值分别达到 90.2% 和 92.0%,但仍存在边界划分偏差。例如,Agri_Control 类实体“喷施叶面肥”可能被错误切分为“喷施”和“叶面肥”;Host 类实体“感病水稻品种植株”在复杂修饰结构下也易因语义干扰出现边界识别不准。此外,在 Disease 与 Pest 类实体中,尽管领域预训练模型 AgriRoBERTa 已有效缓解了大部分术语歧义,但在特定上下文中,如“纹枯病菌”仍可能被误判为疾病本身而非病原体。这反映出模型对深层上下文的理解能力仍有待加强。

2.4 消融试验

为系统探究所提领域预训练 AgriRoBERTa 模型和曼哈顿注意力机制对于水稻病虫害命名实体识别任务性能的影响,以 BiLSTM - CRF 作为基准模型设计消融试验,结果如表 8 所示。

表 8 消融试验结果

Tab.8 Results of ablation experiments			%
模型	精确率	召回率	F1 值
BiLSTM - CRF	83.63	82.17	82.89
BiLSTM - Man - CRF	83.85	85.31	84.58
AgriRoBERTa - BiLSTM - CRF	85.29	91.26	88.18
AgriRoBERTa - BiLSTM - Man - CRF	87.87	93.71	90.69

2.4.1 领域预训练模型对性能的影响

将 BiLSTM - CRF 作为基准,对比引入 AgriRoBERTa 后的 AgriRoBERTa - BiLSTM - CRF 模型。基准模型 BiLSTM - CRF 的 F1 值为 82.89%,而 AgriRoBERTa - BiLSTM - CRF 模型的召回率从 82.17% 大幅提升至 91.26%,F1 值提高 5.29 个百分点。这一提升主要因为 AgriRoBERTa 通过领域预训练学习到水稻病虫害实体的语义关联,增强了其解决上下文歧义和识别领域特定实体的能力。

2.4.2 曼哈顿注意力机制对模型性能的影响

为探究曼哈顿注意力机制对实体边界识别的影响,将其嵌入 BiLSTM - CRF 层间,构建了 BiLSTM - Man - CRF 模型。由表 8 可知,BiLSTM - Man - CRF 模型相较 BiLSTM - CRF 基线模型,F1 值提升 1.69 个百分点,召回率从 82.17% 增至 85.31%。这些性能提升表明,曼哈顿注意力机制通过 L1 范数约束特征交互范围,在长序列症状描述中精准聚焦实体边界,降低了噪声对“实体-上下文”关联判定的干扰。

2.4.3 模块组合对性能的协同影响

为深入评估 AgriRoBERTa 与曼哈顿注意力机制协同优化的效果,构建了完整的 AgriRoBERTa - BiLSTM - Man - CRF 模型,并将其与之前的消融模型进行对比。由表 8 可知,AgriRoBERTa - BiLSTM - Man - CRF 模型所有指标均表现出色,和未集成注意力机制的 AgriRoBERTa - BiLSTM - CRF 模型相比,精确率提升至 87.87%,F1 值提高 2.51 个百分点。这一性能提升证实了曼哈顿注意力机制与 AgriRoBERTa 的有效融合能进一步优化模型,通过抑制长程噪声干扰和增强局部特征提取,显著提升了实体识别精度。

2.4.4 消融试验总结

上述 3 组消融试验分别验证了 AgriRoBERTa 和曼哈顿注意力机制在水稻病虫害 NER 任务中的重要作用。AgriRoBERTa 解决领域语义理解问题,曼哈顿注意力解决实体边界捕捉问题,两者的结合实现了优势互补,共同推动了模型精确率和召回率的提升。最终,AgriRoBERTa - BiLSTM - Man - CRF 模型 F1 值为 90.69%,充分验证了各模块的必要性及协同有效性。

2.5 其他数据集结果分析

在 3 个公开数据集上进行测试,以评估模型的泛化能力,结果如表 9 所示。模型在 Resume 数据集上表现最佳,F1 值达 94.80%。该数据集包含人名、组织、职称等常见且语义边界清晰的实体类型。模型在此类结构化数据上能准确识别特征模式,实现高精确率与召回率,这反映了模型在处理规范文本时能力突出。相比之下,Bank 数据集则包含大量主观性较强且不规则的评价性实体,如名词性评价和形容词性评价。文本环境更具变化性,上下文依赖性更强。面对这种挑战,模型 F1 值仍达到 84.47%,表明其具备较好的泛化能力。在 CCKS2019 数据集,文本涉及大量专业术语和复杂的上下文关系。模型未经过该领域预训练,其 F1 值达到 80.70%。进一步证明了模型优异的性能和泛化能力。

表 9 公开数据集训练结果

Tab.9 Training results of public dataset			%
数据集	精确率	召回率	F1 值
Resume	94.32	95.28	94.80
CCKS2019	80.28	81.12	80.70
Bank	82.80	86.20	84.47

3 结束语

聚焦于水稻病虫害领域的命名实体识别任务,

针对该领域高质量数据稀缺的问题,构建了包含 11 类实体的 NER 数据集及领域语料库,为后续相关研究提供了可靠的数据支撑;同时,为提高水稻病虫害实体识别准确率,提出 AgriRoBERTa - BiLSTM - Man - CRF 模型,通过领域适配二次预训练

RoBERTa 强化术语语义理解,并借助曼哈顿注意力机制优化实体边界区分。试验结果显示该模型在自建数据集上的精确率、召回率和 F1 值分别达到 87.87%、93.71% 和 90.69%,性能优于现有主流模型。

参 考 文 献

[1] 邵舜东. 基于水稻病虫害知识图谱的问答系统设计与实现[D]. 合肥:安徽农业大学,2024.
SHAO Shundong. Design and implementation of question answering system based on knowledge graph of rice pests and diseases [D]. Hefei: Anhui Agricultural University, 2024. (in Chinese)

[2] 丁建平,李卫军,刘雪洋,等. 命名实体识别研究综述[J]. 计算机工程与科学,2024,46(7):1296 - 1310.
DING Jianping, LI Weijun, LIU Xueyang, et al. A review of named entity recognition research[J]. Computer Engineering and Science, 2024, 46(7): 1296 - 1310. (in Chinese)

[3] 吴磊,汪杭军. 基于预训练模型的基层治理敏感实体识别方法[J]. 电子技术应用,2023,46(9):109 - 114.
WU Lei, WANG Hangjun. Identification method of sensitive entities in grassroots governance based on pretraining models[J]. Application of Electronic Technique, 2023, 49(9): 109 - 114. (in Chinese)

[4] 聂啸林,张礼麟,牛当当,等. 面向葡萄知识图谱构建的多特征融合命名实体识别[J]. 农业工程学报,2024,40(3):201 - 210.
NIE Xiaolin, ZHANG Lilin, NIU Dangdang, et al. Multi-feature fusion named entity recognition method for grape knowledge graph construction[J]. Transactions of the CSAE, 2024, 40(3): 201 - 210. (in Chinese)

[5] 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)

[6] RAU L F. Extracting company names from text[C] // Proceedings the Seventh IEEE Conference on Artificial Intelligence Application. IEEE Computer Society, 1991: 29 - 32.

[7] 方晓珊,盛焕烨. 中国人名识别中规则抽取的一种基于实例的方法(英文)[C] // 中国中文信息学会. 第一届学生计算语言学研讨会论文集, 2002: 114 - 121.

[8] BIKEL D M, SCHWARTZ R, WEISCHEDEL R M. An algorithm that learns what's in a name[J]. Machine Learning, 1999, 34(1): 211 - 231.

[9] CURRAN J R, CLARK S. Language independent NER using a maximum entropy tagger[C] // Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, 2003: 164 - 167.

[10] LAFFERTY J, MCCALLUM A, PEREIRA F C N. Conditional random fields: probabilistic models for segmenting and labeling sequence data[C] // Proceedings of the 18th International Conference on Machine Learning, 2001: 282 - 289.

[11] COLLOBERT R, WESTON J. A unified architecture for natural language processing: deep neural networks with multitask learning[C] // Proceedings of the 25th International Conference on Machine Learning, 2008: 160 - 167.

[12] HAMMERTON J. Named entity recognition with long short-term memory[C] // Proceedings of the Seventh Conference on Natural Language Learning at HLT - NAACL 2003, 2003: 172 - 175.

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

[14] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C] // Proceedings of the 31st Annual Conference on Neural Information Processing Systems, 2017: 5999 - 6009.

[15] RADFORD A, NARASIMHAN K. Improving language understanding by generative pre-training[R]. 2018.

[16] PETERS M E, NEUMANN M, IYYER M. Deep contextualized word representations [J]. arXiv Preprint, arXiv:1802.05365, 2018.

[17] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[C] // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4171 - 4186.

[18] DAI Z, WANG X, NI P, et al. Named entity recognition using BERT - BiLSTM - CRF for Chinese electronic health records [C] // 2019 12th International Congress on Image and Signal Processing, Biomedical Engineering and Informatics (CISP-BMEI). IEEE, 2019: 1 - 5.

[19] WANG B, XU C, ZHAO X, et al. Mineru: an open-source solution for precise document content extraction[J]. arXiv Preprint, arXiv:2409.18839, 2024.

[20] MENGLIEV D, BARAKHNIN V, ABDURAKHMONOVA N, et al. Developing named entity recognition algorithms for Uzbek: dataset insights and implementation[J]. Data in Brief, 2024, 54: 110413.

waveguide[J]. Spectroscopy and Spectral Analysis, 2015, 35(3): 841 – 845. (in Chinese)

[18] GATTINGER P, ZORIN I, RANKL C, et al. Spectral-coding-based compressive single-pixel NIR spectroscopy in the sub-millisecond regime[J]. Sensors, 2021, 21(16): 5563.

[19] JEON J, PARK J W, KIM G B, et al. Visible to near-infrared single pixel microspectrometer using electrothermal MEMS grating[J]. Optics Express, 2023, 31(9): 14583 – 14592.

[20] 高浩,王潇,尚林伟,等. 基于二向色及透射准直的小型近红外拉曼光谱仪[J]. 光谱学与光谱分析, 2018, 38(6): 1933 – 1937.

GAO Hao, WANG Xiao, SHANG Linwei, et al. Design and application of small NIR – Raman spectrometer based on dichroic and transmission collimating[J]. Spectroscopy and Spectral Analysis, 2018, 38(6): 1933 – 1937. (in Chinese)

[21] 陈潭轩,杨怀栋,陈科新,等. 宽光谱 Czerny – Turner 光谱仪中的彗差与分辨率[J]. 光谱学与光谱分析, 2010, 30(6): 1692 – 1696.

CHEN Tanxuan, YANG Huaidong, CHEN Kexin, et al. Coma and resolution in wide spectral region Czerny – Turner spectrometer[J]. Spectroscopy and Spectral Analysis, 2010, 30(6): 1692 – 1696. (in Chinese)

[22] ZHANG Y, HUANG Y, LI W, et al. Application of artificial neural network and DS algorithm to calibration transfer of rice protein powder[C]//2016 Sixth International Conference on Instrumentation & Measurement, Computer, Communication and Control (IMCCC), 2016: 822 – 825.

[23] WANG M, ZHENG K, YANG G, et al. A robust near-infrared calibration model for the determination of chlorophyll concentration in tree leaves with a calibration transfer method[J]. Analytical Letters, 2015, 48(11): 1707 – 1719.

[24] GARCÍA N P J, GARCÍA-GONZALO E, SÁNCHEZ L F, et al. Cyanotoxin level prediction in a reservoir using gradient boosted regression trees: a case study[J]. Environmental Science and Pollution Research, 2018, 25(23): 22658 – 22671.

[25] 樊梦丽,赵越,刘言,等. 近红外光谱水光谱组学[J]. 化学进展, 2015, 27(2/3): 242 – 250.

FAN Mengli, ZHAO Yue, LIU Yan, et al. Aquaphotomics of near infrared spectroscopy[J]. Progress in Chemistry, 2015, 27(2/3): 242 – 250. (in Chinese)

[26] HUANG Z, SANAEIFAR A, TIAN Y, et al. Improved generalization of spectral models associated with Vis – NIR spectroscopy for determining the moisture content of different tea leaves[J]. Food Engineering, 2021, 293: 110374.

[27] LI C, ZHANG S, SUN H, et al. Study on a two-dimensional correlation visible-near infrared spectroscopy kinetic model for the moisture content of fresh walnuts stored at room temperature[J]. Food Process Engineering, 2020, 43:e13551.

(上接第 298 页)

[21] LIU Y, OTT M, GOYAL N, et al. Roberta: a robustly optimized bert pretraining approach[J]. arXiv Preprint,arXiv:1907.11692,2019.

[22] DE MELLO H R, DE PAIVA A C, SILVA A C, et al. Application of natural language modeling techniques in natural gas segmentation in seismic reflection images[J]. Neural Computing and Applications, 2025, 37(4): 2383 – 2409.

[23] 杨璐,张恬,郑丽敏,等. 兽药致病命名实体 Att – Aux – BERT – BiLSTM – CRF 识别[J]. 农业机械学报,2022,53(3): 294 – 300.

YANG Lu, ZHANG Tian, ZHENG Limin, et al. Recognition of animal drug pathogenicity named entity based on Att – Aux – BERT – BiLSTM – CRF[J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(3): 294 – 300. (in Chinese)

[24] 孙艳歌,吴飞,姚建峰,等. 多尺度自注意力特征融合的茶叶病害检测方法[J]. 农业机械学报,2023,54(12):308 – 315.

SUN Yange, WU Fei, YAO Jianfeng, et al. Tea disease detection method with multi-scale self-attention feature fusion[J]. Transactions of the Chinese Society for Agricultural Machinery,2023,54(12):308 – 315. (in Chinese)

[25] 廖天正,林晓兰,陈永辉,等. 基于 BiLSTM – CRF 的命名实体识别在临床电子病历的研究与应用[J]. 电脑知识与技术, 2024,20(28):17 – 19.

[26] 程名,于红,冯艳红,等. 融合注意力机制和 BiLSTM + CRF 的渔业标准命名实体识别[J]. 大连海洋大学学报,2020, 35(2):296 – 301.

[27] YU B, WEI J. IDCNN – CRF-based domain named entity recognition method[C]//2020 IEEE 2nd International Conference on Civil Aviation Safety and Information Technology (ICCASIT). IEEE, 2020:542 – 546.