doi:10.6041/j.issn.1000-1298.2025.07.052

基于 mBART 的农作物命名实体规范化研究

胡玉雪[1,2]　黄仲强[1,3]　王同官[1,3]　苏东宇[1]　申余丰[4]　沙　瀛[1,5]

(1.华中农业大学信息学院, 武汉 430070; 2.农业农村部智慧养殖技术重点实验室, 武汉 430070;
3.湖北省农业大数据工程技术研究中心, 武汉 430070; 4.中国人民解放军湖北省军区, 武汉 430070;
5.农业智能技术教育部工程研究中心, 武汉 430070)

摘要: 由于地域、文化差异,农业文本中实体名称混乱,使得自动识别和提取信息变得复杂,限制了农业信息化发展。为提高农业信息提取效率,本文提出了基于 mBART 的农业命名实体规范化方法 mJoint。首先,基于农业领域专家的知识经验,构建了一个以农作物为主的农业文本数据集,涵盖豆类、谷物和油料三大农作物,共包含 22 440 条高质量的农业标注数据。其次,农业实体规范化问题涉及农业非规范化实体的检测与识别 2 个问题,本文提出基于 mBART 的统一生成式框架来联合检测、识别出农业非规范实体,直接完成农业命名实体规范化任务。为了提高农业实体规范化效果,在模型中额外引入农业非规范实体检测和农业非规范实体识别 2 个辅助任务。最后,在提出的农作物数据集上进行大量实验,结果表明,本文提出的 mJoint 在农业命名实体规范化任务上的 P、R 与 F1 值都达到 0.99 以上,相较于其他对比方法,各项指标均为最优。与大语言模型相比,本文提出的方法同样具有显著优势。

关键词: 农业文本; 农作物; 命名实体规范化; mBART; 统一生成式框架

中图分类号: TP391.1　　**文献标识码**: A　　**文章编号**: 1000-1298(2025)07-0558-09　　**OSID**:

Crop Named Entity Normalization Based on mBART

HU Yuxue[1,2]　HUANG Zhongqiang[1,3]　WANG Tongguan[1,3]　SU Dongyu[1]　SHEN Yufeng[4]　SHA Ying[1,5]

(1. *College of Informatics*, *Huazhong Agricultural University*, *Wuhan* 430070, *China*
2. *Key Laboratory of Smart Farming for Agricultural Animals*, *Wuhan* 430070, *China*
3. *Hubei Engineering Technology Research Center of Agricultural Big Data*, *Wuhan* 430070, *China*
4. *Hubei Provincial Military Command*, *Wuhan* 430070, *China*
5. *Engineering Research Center of Intelligent Technology for Agriculture*, *Ministry of Education*, *Wuhan* 430070, *China*)

Abstract: Due to geographical or cultural differences, the entity names in agricultural texts are confused, which makes automatic identification and extraction of information complicated and limits the development of agricultural informatization. In view of this, an agricultural entity normalization method based on mBART was proposed. Firstly, based on the knowledge and experience of experts in the agricultural field, a crop-oriented agricultural text dataset was constructed, covering the three major crops of "legumes", "cereals" and "oil crops", with a total of 22 440 pieces of high-quality agricultural labeling data. Secondly, the problem of agricultural entity normalization involved the detection and identification of non-normalized agricultural entities. A unified generative framework was proposed based on mBART to jointly detect and identify agricultural non-normalized entities and directly complete the task of normalizing agricultural named entities. Furthermore, in order to improve the normalization effect of agricultural entities, auxiliary tasks of agricultural non-normalized entity detection and agricultural non-normalized entity recognition were additionally introduced into the model. Finally, extensive experiments were conducted on the proposed crop dataset. The results showed that the proposed method achieved P, R, and F1 above 0.99 in the task of agricultural entity normalization, and all indexes were optimal compared with other methods. Compared with the large language models, the proposed method also had significant advantages.

Key words: agricultural text; crop; named entity normalization; mBART; unified generative framework

收稿日期: 2024-04-14　修回日期: 2024-08-07

基金项目: 国家自然科学基金项目(62272188)、中央高校基本科研业务费专项资金项目(2662021JC008)和内蒙古自治区科技重大专项(2021ZD0046)

作者简介: 胡玉雪(1990—),女,博士生,主要从事农业文本处理和自然语言处理研究,E-mail: hyx@mail.hzau.edu.cn

通信作者: 沙瀛(1973—),男,教授,博士生导师,主要从事农业信息化技术和自然语言处理研究,E-mail: shaying@mail.hzau.edu.cn

0　引言

我国正加快从农业大国向农业强国的转变，农业信息化在其中发挥着重要作用[1]。随着信息化进程推进，农业领域产生了大量涵盖技术、生产和管理的文本数据。如何高效准确地从海量数据中获取所需内容，实现农业信息的有序组织与利用，已成为农业信息化亟待解决的问题之一。然而，农业文本普遍存在非结构化、表达不规范、种类繁多等问题，尤其是农作物别名、病虫害名称等非规范实体广泛存在，表现为缩写、同物异名、异物同名等现象，严重影响了信息的识别与提取效率，制约了农业信息资源的有效利用。因此，开展农业实体规范化研究，对于提升信息组织效率、促进农业知识图谱构建具有重要意义。

农作物命名因地域和时期差异呈现多样性，反映了语言、文化与认知差异[2]。实际交流中，人们常采用简单、直观的方式命名，导致不规范缩写、同物异名、异物同名等非规范实体广泛存在。例如，“棉花黄萎病”可简称为“CW”或“黄萎”，而“黄萎”也可以指“黄萎病”或者“烟草黄萎病”；“玉米”在不同地区被称为“包谷”或“棒子”；“苹果”不仅指水果，也可以指手机。这类现象增加了农业信息处理的复杂性。为提升信息抽取的准确性与系统规范性，亟需建立非规范实体与标准概念之间的映射机制，实现农业实体的规范化。

近年来，随着深度学习模型发展和计算性能的提升，基于卷积神经网络（CNN）[3]、长短期记忆网络（LSTM）[4]、注意力机制（Attention）[5-6]等深度学习方法被相继提出，在精准施肥[7]、病虫害防治[8]和农作物监测[9]等农业领域得到了广泛应用。然而，关于农业命名实体相关研究有限。一方面，现有研究更注重模型的性能提升，忽视了其在领域内的适用性；另一方面，农业领域缺乏公开数据集，并且数据收集困难[10]。目前，少量研究涉及农业命名实体识别[11-14]。相比之下，农业命名实体规范化研究尚属空白，仅在生物医学领域有所探索[15-17]。

综上所述，农业命名实体规范化研究主要面临 2 个问题：①缺乏相关数据集。②目前农业领域尚无相关研究，且其他领域的相关研究通常采用先实体检测再识别的管道式架构，容易导致误差传导。为了解决上述问题，本文结合农业专家知识构建一个涵盖豆类、谷物和油料三大类、共 22 440 条标注数据的大规模农作物数据集。基于此，提出一种基于 mBART 的统一生成式模型 mJoint，可同时完成农业非规范实体的检测与识别。为进一步提升模型效果，在农业命名实体规范化主任务基础上，引入农业非规范实体检测和农业非规范实体识别 2 个辅助任务。最后，在提出的农作物数据集上进行大量实验以验证方法性能。

1　数据采集与预处理

1.1　数据采集

农业命名实体规范化研究的首要问题是收集大量相关的农业文本数据。为了获取高质量且多样化的农业命名实体文本，本文从 2 个互联网平台进行数据采集，分别是中国农业信息网（http://www.agri.cn/）和微博“三农”平台。数据采集主要涉及三大农作物：谷物、豆类和油料。根据研究需求，将三大农作物细分下的 10 个作物（详见表 1）对应的规范化实体设定为种子关键词。基于上述种子关键词爬取到的语料，继续搜索语义相近的关键词，进一步扩充语料。最终，共获取了 46 225 条原始数据，其中谷物类 18 275 条、豆类 10 894 条、油料类17 056 条。由于爬取下来的文本数据包含一些噪声、无关信息或者格式不规范的内容，接下来对原始农作物文本数据进行数据初筛、数据清洗和数据标注工作。整个数据获取和处理流程如图 1 所示。

表 1　数据集各类别用语及句子数量

Tab.1　Number of terms and sentences in each category of dataset

类别	规范化实体	非规范化实体	句子总数
谷物	稻谷	稻米、稻子、谷子	7 308
	小麦	麸麦、浮麦、浮小麦、麦子	
	玉米	棒子、包谷、包芦、包米、苞谷、苞芦、苞米、玉茭、玉麦、玉蜀黍、御麦、番麦	
豆类	大豆	黑豆、黄豆、毛豆、秣食豆、泥豆、青豆、青仁黑豆、青仁乌豆、菽	6 081
	红豆	赤豆、红豆米、红小豆、相思豆、小豆	
	绿豆	菉豆、青小豆、植豆	
油料	花生	地豆、番豆、落花生、长生果、泥豆、长寿果	8 467
	向日葵	葵花、太阳花、向阳花、朝阳花	
	油菜籽	菜子、菜籽、油菜子、芸苔子	
	芝麻	胡麻、巨胜、油麻、脂麻	
None			585
共计			22 440

1.2　数据预处理

在本研究中，针对爬取到的原始农业文本数据中存在的冗余和无效信息，采用一系列有效的数据清洗方法进行初筛和数据清洗，以确保后续

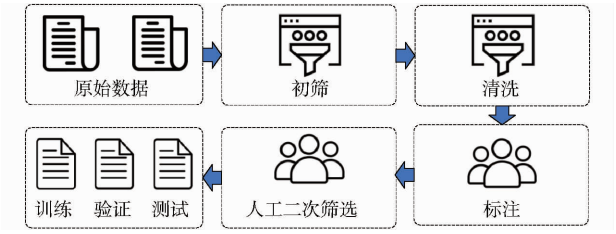

图 1 数据处理流程图

Fig. 1 Flowchart of data construction process

分析的准确性和可靠性。首先,采用基于关键词的筛选方法,通过设定关键词来初步筛选数据,确保选取的内容与研究目的相关。随后,使用 Python 脚本对数据进行二次筛选和更深层次的清洗。具体而言,使用正则表达式的方法,有针对性地去除原始文本数据中的常见标点符号,从而净化文本内容,过滤掉与研究无关的无用信息和重复数据,减少文本中的噪声,使后续分析更加精准和可靠。最后,采用人工审核策略,来保障数据的最终质量,最大程度地排除可能存在的错误,减小实验偏差。

1.3 数据标注

数据标注的目的是找出句子中的非规范化实体并将其标注到对应的规范实体。首先需要判断句中是否包含农业非规范实体,如果包含,将其标注为对应的规范实体,如果不包含,将其标注为“None”。在进行农业实体规范化的数据标注之前,为确保对相同术语和概念采用一致的标注方式,邀请了 1 位农业领域专家指导 5 位植物科学技术学院硕士研究生进行数据标注。这些专业人员在农业领域具有深厚的知识,能够准确理解和标注特定的农业术语。在标注的过程中,采用投票法来解决可能出现的歧义情况,若遇到歧义,通过投票的方式进行决策,选择票数最高的标注结果作为最终标准。这种方法有效避免了个体主观因素对标注结果的影响,确保了标注的一致性和可信度。

1.4 数据分析

经过上述流程的处理,最终得到农业用语数据集的各类别及其数量如表 1 所示。

从表 1 可看出,3 种农作物类别中,谷物共含有 7 308 条句子,豆类有 6 081 条句子,油料有 8 467 条句子,None 类别(表示句中没有农业非规范化实体)有 585 条句子。所有类别共计 22 440 条句子,能够满足实验需要。

为了更直观展示经过以上流程处理后的农作物命名实体数据集,同时介绍本文的主要研究任务,在构建好的数据集中,从 3 种农作物及 None 的类别数据中,挑选出 7 个样例进行展示,如表 2 所示。

表 2 数据集中非规范化用语样本

Tab. 2 Sample of non-standardized terms in dataset

句子样本	非规范化用语	规范用语
但是比起种包谷,葡萄园的收入更多	包谷	玉米
三春不如一秋忙,收完棒子再种粮	棒子	玉米
古人称之为“菽”,五谷之一	菽	大豆
其中上升明显的是红小豆种植面积增加 220 亩	红小豆	红豆
本周来甘肃、新疆等地的葵花籽批量上市了	葵花	向日葵
我们油麻田现在春暖花开,美得很哩	油麻	芝麻
不能东一榔头、西一棒子	棒子	None

在表 2 中,句子“三春不如一秋忙,收完棒子再种粮”中,“棒子”是农业非规范化实体,对应的规范实体应该为“玉米”。本文的任务就是检测出农业文本中的“棒子”并识别出其对应的规范化实体是“玉米”。同样地,句子“我们油麻田现在春暖花开,美得很哩”中的“油麻”也是非规范化实体,其对应的规范化实体应该为“芝麻”。此外,为了提高模型泛化能力,还在数据中添加不含有规范化实体的句子。例如,在“不能东一榔头、西一棒子”这句话中,结合上下文语义,“棒子”一词仅具有字面含义,没有非规范化实体的意思,因此不属于任何一类规范化实体。

2 基于 mBART 的农业用语规范化研究方法

农业实体规范化的问题涉及非规范实体的检测与非规范实体的识别,现有方法使用串联方式容易导致错误传导。因此,本文提出基于 mBART 的统一生成式框架来联合完成农业非规范化实体的检测与识别。

2.1 模型结构

本文提出的模型结构主要由基于 mBART(Multilingual bidirectional and auto-Regressive transformers)[18] 的生成式模型和 2 个辅助任务构成。BART 本身是一种预训练的序列到序列模型,基本结构包括基于 Transformer 的编码器(encoder)和解码器(decoder)[19]。模型首先使用自编码器的方式将输入序列压缩为潜在表示,然后使用自回归方式将这个表示解码为原始序列。mBART 是 BART 模型的多语版本。为了更好地提升农业命名实体规范化效果,本文在实体规范化的主任务上,额外引入 2 个辅助任务,分别是农业非规范化实体检测和农业非规范化实体识别。本文的模型架构如图 2 所示。

2.2 mBART 编码器

mBART 编码器由多层自注意力机制和前馈网络组成,通过残差连接和层归一化技术,有效捕捉输

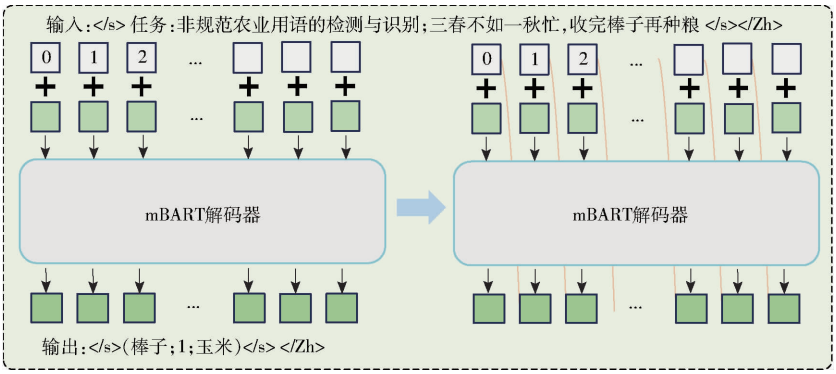

图 2　网络结构图

Fig. 2　Network architecture diagram

入农业文本的内部依赖关系并保持训练过程的稳定性。位置编码的引入使得模型能够理解每个汉字在农业文本中的位置信息。其主要作用是将输入的农业文本映射到一个潜在的语义空间,捕捉输入文本序列的语义信息,并为后续的任务提供一个有意义的表示。在具体实现中,首先将待检测与识别的农业文本输入传递给 mBART 的编码器。输入的农业文本首先经过标记化处理,将其分解为词片段(tokens)。接着将该层每个词片段映射到高维词嵌入空间。然后添加位置编码以保留输入文本中单词的顺序信息。对于输入文本 S,mBART 的编码器可以表示为

$$H_{\text{encoder}} = \text{Encoder}(S) \tag{1}$$

式中　Encoder( )——mBART 编码器

H_{encoder}——文本 S 的语义特征

2.3　mBART 解码器

mBART 的解码器是一个自回归的 Transformer 组件,它通过接收来自编码器的语义表示,并结合自注意力机制和前馈网络,逐步生成目标文本序列。它首先通过编码器-解码器注意力机制接收来自编码器的语义表示,实现源文本到目标文本信息的传递。然后使用自回归方式不断地生成一个一个的 token 来生成目标序列,最终将输入序列的编码表示转换为目标序列。具体地,解码器注意力机制使得模型能够在生成每个词时考虑到已经生成的词语,从而更好地捕捉到词语之间的依赖关系。每个时刻,解码器生成一个词,并使用这个词作为输入继续生成下一个词,这一过程持续到生成结束标志(〈EOS〉)。最终,解码器的输出通过一个线性层和 Softmax 层生成目标农业文本的输出概率分布。mBART 的解码器可以表示为

$$Y_t = \text{Decoder}(H_{\text{encoder}}, Y_{<t}) \tag{2}$$

式中　Decoder( )——mBART 解码器

Y_t——第 t 时刻标记

$Y_{<t}$——t 时刻前标记

2.4　损失函数

为了衡量模型输出概率分布与真实标签分布之间的差异,使模型输出概率分布尽可能接近真实分布,本文使用交叉熵损失函数作为模型的主要损失函数。mBART 接收输入农业文本序列并输出目标农业文本序列。输出序列是词汇表中每个单词的概率分布,并且在生成每个标记时,模型使用 Softmax 函数将输出分数转换为概率分布,以确保生成序列是合法的概率分布。主要损失函数公式为

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N} Q(\hat{Y}_i)\ln(P(Y_i;\Theta)) \tag{3}$$

式中　N——样本数　　Θ——模型参数

$Q(\hat{Y}_i)$——第 i 个样本真实标签分布

Y_i——第 i 个样本预测分布

L_{CE}——主要交叉熵损失

为了增强农业非规范实体的检测能力,模型引入了农业非规范化实体检测这一辅助任务。具体地,该任务在编码器的输出端引入了农业非规范化实体的检测损失 L_{ND}。该损失函数用于监督模型检测出农业非规范化实体的能力,具体公式为

$$L_{ND} = -\frac{1}{N}\sum_{i=1}^{N} y_i \ln p_i + (1-y_i)\ln(1-p_i) \tag{4}$$

式中　y_i——真实样本标签

p_i——模型输出概率,取 0 ~ 1

此外,为了进一步增强实现农业规范化实体的识别能力,模型引入了农业非规范化实体识别任务辅助任务,在解码器的输出端,加入农业非规范化实体识别损失 L_{NI}。该损失函数用于评估模型在识别农业非规范化实体的准确性,具体公式为

$$L_{NI} = -\frac{1}{N}\sum_{i}^{N}\sum_{j}^{C} y_{i,j}\ln(p_{i,j}) \tag{5}$$

式中　C——类别数

$y_{i,j}$——样本 i 的真实标签中第 j 个类别

$p_{i,j}$——模型输出的概率分布中第 j 个类别的概率

最后,模型总损失 L_{all} 为

$$L_{all} = \alpha L_{CE} + \beta L_{ND} + \gamma L_{NI} \tag{6}$$

其中 $\alpha + \beta + \gamma = 1$

式中 α、β、γ——3 个损失函数权重

3 实验与结果分析

3.1 数据集

使用第 1 节提出的农作物命名实体数据集来验证所提方法的效果,该数据集共有 22 440 条农业文本数据,将其按照比例 7∶1.5∶1.5 划分为训练集、验证集和测试集,最终得到 15 708 条农业文本数据作为训练集,3 366 条农业文本数据作为验证集,3 366 条农业文本数据作为测试集。

3.2 实验环境

所有实验均在 Ubuntu 18.0.4 LTS 版本的 Linux 服务器上进行,服务器配备 Tesla - V100 32G GPU,CUDA 版本为 11.1。开发语言为 Python 3.8,使用 Pytorch 1.8.1 框架完成模型的构建和训练。其他超参数设置如表 3 所示。

表 3 参数设置

Tab.3 Parameters setting

模型	数值
最大长度(max_len)	128
批量大小(batch size)	32
学习率(learning rate)	0.000 5
优化器(optimizer)	Adam
训练轮数(epoch)	20

3.3 评价指标

农业实体规范化任务涉及到农业非规范化实体检测和识别,采用统一生成式框架直接完成这一任务。基于此,本文设置 2 种任务的评价指标,使用召回率和精确率的调和平均值 F_{1D}、召回率 R_D、精确率 P_D 来评价模型对农业非规范化实体的检测效果,用 F1 值、R 与 P 来评价模型对农业实体规范化效果。F1 值、R 和 P 都在 0 ~ 1 之间,召回率越接近 1 表示模型对正例的查全率越高;精确率越接近 1 表示模型对正例的查准率越高;F1 值是综合考虑了召回率和精确率的指标,是两者的调和平均值。F1 值越大说明模型在平衡了查准率和查全率的情况下表现越好。

3.4 与基线方法对比结果

为了证明本文所提方法相较于其他对比方法具有优越性,将本文提出的生成式方法分别与 2 个判别式方法和 4 个生成式方法 mBART - large - cc25、mBART - large - 50[18]、mT5 - base 和 mT5 - large[20],共计 6 种基线进行比较。为了公平比较,同样将对比模型在本文收集的数据集上进行微调训练。生成式方法端到端进行训练得到统一结果,而判别式方法管道式地先进行检测再进行识别,所有模型超参数设置与表 3 一致,对比结果如表 4 所示。

表中,RoBERTa - wwm - ext 为结合中文 Whole Word Masking 技术以及 RoBERTa 得到的预训练模型,参数量为 1.01×10^8;RoBERTa - wwm - ext - large 为 RoBERTa - wwm - ext 模型的升级版,参数量为 3.26×10^8;mBART - large - cc25 为使用从 CC 语料库提取的 25 种语言进行预训练的模型;mBART - large - 50 在 mBART - large - cc25 模型的基础上,多了 25 种语言标记,在 50 种语言上进行了预训练;mT5 - base 为使用 101 种语言预训练的模型,参数量为 5.8×10^8;mT5 - large 为使用 101 种语言预训练的模型,参数量为 1.2×10^9。

表 4 与基线方法指标对比结果

Tab.4 Comparison of results with baselines

模型	F_{1D}	R_D	P_D	F1 值	R	P
RoBERTa - wwm - ext	0.939 1	0.939 5	0.938 6	0.915 9	0.914 8	0.917 1
RoBERTa - wwm - ext - large	0.942 1	0.935 9	0.948 5	0.924 9	0.929 9	0.927 2
mBART - large - cc25	0.987 8	0.986 5	0.989 0	0.986 8	0.985 6	0.988 1
mBART - large - 50	0.982 7	0.989 0	0.986 5	0.978 5	0.984 8	0.982 3
mT5 - base	0.982 9	0.987 1	0.988 7	0.971 9	0.976 2	0.977 7
mT5 - large	0.984 3	0.979 8	0.969 1	0.982 5	0.988 2	0.967 3
mJoint	0.998 6	0.998 7	0.998 4	0.997 5	0.995 9	0.999 1

由表 4 可以看出,相较于其他 6 种对比方法,本文提出的方法在农业非规范化实体检测任务和农业命名实体规范化的任务上,P、R 和 F1 值 3 个指标表现都最优。这说明本文通过使用基于 mBART 的方法,同时使用 3 种损失函数对模型进行约束,可以在农业非规范化实体的检测与识别任务上取得最优的效果。同时从表 4 中也可以看出:①判别式方法先检测后识别的方法效果不如生成式方法,可能一方面先检测后识别管道式方法容易导致误差传导,另一方面同时统一完成检测和识别可以相互补充信

息。②生成式方法中 2 个 mT5 模型的表现不如 mBART 模型，这是因为 2 个模型的预训练方法不同，如果输入是“A[mask]B[mask]E”，mBART 模型输出标记为“ABCDE”，而 mT5 模型输出标记为“CD”。mBART 模型执行的任务难度更大，在农业非规范化实体检测和识别任务中更有效。

3.5　与大语言模型对比结果

由于大语言模型(LLMs)的繁荣发展，目前与大模型的实验对比已经成为主流[21-23]，可以作为一种参考，有助于推动领域内的研究发展。将本文提出的方法与 mPLUG-Owl[21]、LLaMA2[22]、Falcon[23]、StableLM[24]、GPT-3.5[25]等大语言模型进行比较。由于计算资源限制，没有对 LLMs 进行微调，直接使用上述 LLMs 进行测试。

mPLUG-Owl 为阿里达摩院发布的多模态大语言模型，拥有 7.0×10^9 个参数；LLaMA2 为 MetaAI 正式发布最新一代的开源大模型，拥有 7.0×10^{10} 个参数；Falcon 由阿联酋的技术创新研究所发布，是中东首个世界顶级的大模型产品，其中包含 4.0×10^{10} 个参数；StableLM 由 Stable Diffusion 的初创公司 Stability AI 发布并开源，该团队训练的大语言模型，拥有 7.0×10^9 个参数；GPT-3.5 为 OpenAI 公司设计的自然语言处理模型系列中的第 4 个模型，其参数量尚未正式公布，最高可达 1.750×10^{11}。

由表 5 可以看出，相较于 5 种大语言模型，本文提出的方法依旧能够取得优秀的效果。在所有大语言模型的结果中，GPT-3.5 模型结果最好，它拥有 1.750×10^{11} 个参数，是参数量最大的模型。同时可以注意到，mPLUG-Owl、Falcon 与 StableLM 这 3 种大语言模型，在非规范农业实体检测任务上的 P_D 为 1，说明模型能够准确地识别出部分正例，但错过了大量真正的正例。LLaMA2 大语言模型在统一农业非规范化用语检测与识别任务中为 0，其他除 GPT-3.5 以外的模型在该任务中准确率也较低。这说明对于大语言模型来说，农业非规范化用语的检测和识别任务仍然具有挑战性。尽管如此，本文提出的方法在农业非规范化用语检测与识别任务上的性能仍然能够超越大语言模型。

表 5　与大模型方法指标对比结果

Tab. 5　Comparison of results with LLMs

模型	F_{1D}	R_D	P_D	F1 值	R	P
mPLUG-Owl	0.238 1	0.136 5	1.000 0	0.109 1	0.063 1	0.418 5
LLaMA2	0.176 5	0.098 4	0.857 1	0.000 0	0.000 0	0.000 0
Falcon	0.373 3	0.229 5	1.000 0	0.160 1	0.098 4	0.428 6
StableLM	0.205 9	0.114 8	1.000 0	0.058 8	0.032 8	0.285 7
GPT-3.5	0.811 3	0.704 9	0.955 6	0.641 5	0.557 4	0.755 6
mJoint	0.998 6	0.998 7	0.998 4	0.997 5	0.995 9	0.999 1

3.6　消融实验

为了验证 2 个辅助任务对农业命名实体识别任务的影响，分别对 2 个辅助任务进行消融研究，将农业非规范化实体检测任务去掉记为 w/o LND，将农业非规范化实体识别任务去掉记为 w/o LDI，将 2 个任务都去掉记为 w/o LND & LDI。最终得到结果如表 6 所示。

表 6　2 个辅助损失函数对结果的影响

Tab. 6　Effect of two auxiliary loss functions on results

模型	F_{1D}	R_D	P_D	F1 值	R	P
w/o LND	0.977 3	0.975 7	0.978 7	0.976 7	0.975 3	0.978 1
w/o LDI	0.988 9	0.988 4	0.989 4	0.988 6	0.988 1	0.989 1
w/o LND & LDI	0.967 6	0.967 7	0.967 4	0.967 3	0.967 4	0.967 1
mJoint	0.998 6	0.998 7	0.998 4	0.997 5	0.995 9	0.999 1

由表 6 可以看出，当 2 个辅助任务分别去掉之后，本文提出模型的效果有所下降，将 2 个辅助任务都去掉后，结果下降明显，说明 2 个辅助任务在农业命名实体规范化任务中都起到有效作用。

为了验证 2 个辅助任务对农业命名实体规范化结果的影响程度，对 2 个辅助任务的损失系数进行超参数分析，即保证 3 个权重总和不变，分别对 α、β 和 γ 取不同的值，结果如图 3 所示。

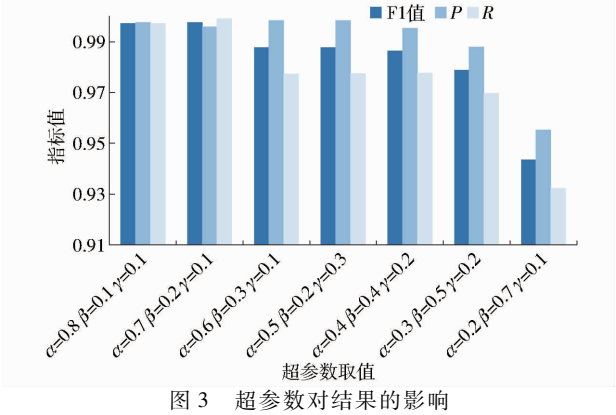

图 3　超参数对结果的影响

Fig. 3　Effect of hyperparameters on results

由图 3 可以看出，对不同损失函数提供不同的权重时，模型取得不同结果，当 $\alpha=0.7$，$\beta=0.2$，$\gamma=0.1$ 时，模型整体效果最佳。这说明本文模型的主要作用来自主任务命名实体规范化的交叉熵损失，其次是农业非规范化实体检测损失，最后是农业非规范化实体识别损失。并且从结果可以发现，当 β 与 γ 和大于 0.5 时，模型效果反而有所下降。这表明农业非规范化实体检测损失与农业非规范化实体识别损失的作用是有限的，如果这 2 种损失占比太高，反而起到抑制作用。因此，选择合适的损失权重是确保模型发挥出最优效果的关键。

3.7 数据集设置

为了进一步验证所提出模型的鲁棒性,分别对数据集规模和数据集组成进行实验分析。

(1)数据集规模影响。对模型的数据效率进行研究,随机抽取原始数据集的 30%、50%、80% 进行对比实验,结果如表 7 所示。总体来看,随着数据量减少,模型的性能呈下降趋势。然而,从 100% 数据规模下降到 30% 的数据规模,模型 F1 值仅降低约 0.02。由此可见,在有限的数据规模下,mJoint 模型仍然可以通过学习语言中的模式和关系来获得更好的泛化性能。

表 7 不同数据规模对比结果

Tab.7 Results of different data scales

数据规模	F_{1D}	R_D	P_D	F1 值	R	P
30%	0.979 9	0.973 4	0.986 5	0.978 3	0.973 3	0.983 4
50%	0.982 4	0.968 3	0.996 8	0.979 2	0.968 1	0.990 5
80%	0.984 9	0.976 6	0.993 4	0.983 4	0.976 6	0.990 3
100%	0.998 6	0.998 7	0.998 4	0.997 5	0.995 9	0.999 1

(2)数据集组成影响。研究不同作物类别下模型的性能,将原始数据集中的豆类、谷物和油料 3 个类别数据分别剥离开,得到 3 个子数据集,并将这些子数据集两两组合以形成混合数据集。模型在这些数据集上的实验结果如表 8 所示。结果表明,模型对数据集组成和数据集大小相对不敏感,结果相差在 0.005 以内,证明了模型的鲁棒性。

表 8 不同数据组成对比结果

Tab.8 Results of different data compositions

数据组成	F_{1D}	R_D	P_D	F1 值	R	P
谷物	0.995 9	0.998 2	0.993 7	0.994 8	0.997 4	0.992 3
豆类	0.997 3	0.998 2	0.996 4	0.996 8	0.998 2	0.995 5
油料	0.994 5	0.989 8	0.999 2	0.993 2	0.988 3	0.998 2
谷物 + 豆类	0.998 2	0.998 2	0.998 2	0.997 3	0.998 4	0.996 1
谷物 + 油料	0.997 6	0.998 4	0.996 9	0.997 0	0.997 3	0.996 7
豆类 + 油料	0.998 0	0.998 2	0.997 9	0.997 1	0.997 7	0.996 5
谷物 + 豆类 + 油料	0.998 6	0.998 7	0.998 4	0.997 5	0.995 9	0.999 1

3.8 案例分析

3.8.1 正确案例

为了展示农业非规范化实体规范化效果,从成功的农业非规范化实体检测和识别案例中随机挑选出 3 个案例进行展示。

从表 9 可以看出,在案例 1 中,“稻米”在句子中属于农业非规范化实体,其对应的农业规范化实体为“稻谷”,模型检测成功,输出结果为(稻米;1;稻谷);在案例 2 中,“黑豆”在句中属于农业非规范化实体,其对应的农业规范化实体为“大豆”,因此模型的输出结果为(黑豆;1;大豆);在案例 3 中,“葵花”在句中属于农业非规范化实体,其对应的农业规范化实体为“向日葵”,因此模型输出结果为(葵花;1;向日葵)。

表 9 正确案例展示

Tab.9 Demonstration of correct cases

案例 1	输入:米的订单价格为每公斤 8 ~ 16 元,比普通稻米的价格提高了近三倍 输出:(稻米;1;稻谷)
案例 2	输入:容易生长,各地区在这个时间段都是葡萄发生黑豆病的高峰期 输出:(黑豆;1;大豆)
案例 3	输入:葵花籽进口量预计为 120 万吨,高于上年的 115 万吨 输出:(葵花;1;向日葵)

3.8.2 错误案例

虽然本文提出的方法取得了优越的效果,评价指标超过了 6 种对比方法与 5 个大语言模型。但是在少数农业文本中仍然存在无法正确检测或识别出农业非规范化实体的情况。将对这些检测和识别失败的案例进行分析,主要分为 2 类,如表 10 所示。

表 10 失败案例展示

Tab.10 Demonstration of failure cases

案例 4	输入:饮等渠道品尝到贵州的“一杯茶、一条鱼、一包米”,叫响贵州鱼米茶品牌 输出:(包米;0;小麦)
案例 5	输入:请人工,省时省力,趁着晴好天气晒‘长生果’咯 输出:(长生果;0;花生)

失败的案例主要分为 2 种情况:农业非规范化实体检测正确但识别错误,以及农业非规范化实体检测错误但识别正确。从表 10 可以看出,在案例 4 中,文本中的“包米”应当与其前边的“一”字组成量词短语“一包米”,故而此处“包米”应被视为规范化实体,虽然模型可以正确将其检测为规范实体,但是将其错误对应到了“小麦”类别(案例 4 正确的输出应为:(包米;0;None))。在案例 5 中,模型并未检测出农业非规范化实体“长生果”。尽管如此,模型还是正确地识别了其对应的规范化实体“花生”(案例 5 正确的输出应为:(长生果;1;花生))。尽管本文提出的方法取得了良好的识别效果,但是农业文本本身存在的非结构化问题、同物异名、异名同物,以及中文分词困难等,仍在农业非规范化用语的检测与识别任务中带来了挑战。

4 结论

(1)面向农业规范化用语领域,本文构建了一个

应用于农业命名实体研究的数据集。该数据集涵盖农作物豆类、谷物和油料 3 个大类，共 10 个小类，包含 22 440 条高质量带标注的农业命名实体文本，有效缓解了农业实体规范化研究语料库不足的问题。

(2)针对农业命名实体规范化研究，提出了一种基于 mBART 的统一生成式模型，来联合检测与识别农业命名实体。通过在本文构建的数据集上进行大量的对比实验与消融实验，验证了该模型的有效性。在农业非规范化实体检测任务和农业命名实体规范化任务上 P、R 和 F1 值都能达到 0.99 以上。即使与大语言模型相比，本文提出的方法同样具有明显的优势。

参 考 文 献

[1] 薛洲，高强. 从农业大国迈向农业强国：挑战、动力与策略[J]. 南京农业大学学报(社会科学版)，2023，23(1)：1－15.
XUE Zhou，GAO Qiang. Moving from a large agricultural country to an agricultural power house challenges，drivers and strategies[J]. Journal of Nanjing Agricultural University (Social Sciences Edition)，2023，23(1)：1－15. (in Chinese)

[2] 卢方舒. 河南商丘农村方言传统农业词汇变化研究[D]. 汕头：汕头大学，2022.
LU Fangshu. Study on the changes of traditional agricultural vocabulary in Shangqiu rural dialect[D]. Shantou：Shantou University，2022. (in Chinese)

[3] LECUN Y，BOTTOU L，BENGIO Y，et al. Gradient-based learning applied to document recognition[J]. Proceedings of the IEEE，1998，86(11)：2278－2324.

[4] HOCHREITER S，SCHMIDHUBER J. Long short-term memory[J]. Neural Computation，1997，9(8)：1735－1780.

[5] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is all you need[J]. Advances in Neural Information Processing Systems，2017，30：5998－6008.

[6] DEVLIN J，CHANG M W，LEE K，et al. Bert：pre-training of deep bidirectional transformers for language understanding[J]. arXiv Preprint，arXiv：1810.04805，2018.

[7] 朱凤磊，张立新，胡雪，等. 基于蝙蝠优化 BP－PID 算法的精准施肥控制系统研究[J]. 农业机械学报，2023，54(增刊 1)：135－143，171.
ZHU Fenglei，ZHANG Lixin，HU Xue，et al. Precision fertilizer application control system based on BA Optimization BP－PID Algorithm[J]. Transactions of the Chinese Society for Agricultural Machinery，2023，54(Supp. 1)：135－143，171. (in Chinese)

[8] 李紫洲. 基于机器学习的玉米病虫害智能问答系统设计与实现[D]. 武汉：华中师范大学，2022.
LI Zizhou. Design and implementation of the intelligent question answering system for maize diseases and pests based on machine learning[D]. Wuhan：Central China Normal University，2022. (in Chinese)

[9] 吴慧. 基于卷积神经网络的农作物生长识别监测系统[D]. 南京：南京信息工程大学，2020.
WU Hui. Crop growth recognition monitoring system based on convolutional neural network[D]. Nanjing：Nanjing University of Information Science and Technology，2020. (in Chinese)

[10] 徐润方. 面向农业领域的命名实体识别研究[D]. 武汉：华中农业大学，2023.
XU Runfang. Research on named entity recognition for agriculture[D]. Wuhan：Huazhong Agricultural University，2023. (in Chinese)

[11] 蒲攀，张越，刘勇，等. 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，et al. 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)

[12] GUO X，LU S，TANG Z，et al. CG－ANER：Enhanced contextual embeddings and glyph features-based agricultural named entity recognition[J]. Computers and Electronics in Agriculture，2022，194：106776.

[13] 胡乔，赵春江，吴华瑞，等. 结合对抗训练和注意力机制的蔬菜种植领域命名实体识别[J]. 计算机工程与应用，2025，61(9)：343－352.
HU Qiao，ZHAO Chunjiang，WU Huarui，et al. Named entity recognition in vegetable cultivation combining adversarial training and attention mechanism[J]. Computer Engineering and Applications，2025，61(9)：343－352. (in Chinese)

[14] 余克健，张程，乐毅，等. 基于 GPT 修正农业病虫害命名实体识别方法[J]. 内蒙古农业大学学报(自然科学版)，2023，44(5)：34－43.
YU Kejian，ZHANG Cheng，LE Yi，et al. Recognition method of named entities of agricultural pests and diseases based on GPT corrections[J]. Journal of Inner Mongolia Agricultural University (Natural Science Edition)，2023，44(5)：34－43. (in Chinese)

[15] 娄银霞. 面向生物医学文本的实体识别和规范化研究[D]. 武汉：武汉大学，2020.
LOU Yinxia. Research on named entity recognition and normalization from biomedical text[D]. Wuhan：Wuhan University，2020. (in Chinese)

[16] JI Z，WEI Q，XU H. Bert-based ranking for biomedical entity normalization[J]. AMIA Summits on Translational Science Proceedings，2020，2020：269.

[17] 冯凤翔,任慧玲,李晓瑛,等.融合相似度算法与预训练模型的中文电子病历实体映射方法研究[J].医学信息学杂志,2023,44(5):45-50.
FENG Fengxiang, REN Huiling, LI Xiaoying, et al. Study on Chinese electronic medical record entity mapping method by fusing similarity algorithms and pre-trained[J]. Journal of Medical Informatics, 2023,44(5):45-50. (in Chinese)
[18] LIU Y, GU J, GOYAL N, et al. multilingual denoising pre-training for neural machine translation[J]. Transactions of the Association for Computational Linguistics, 2020, 8: 726-742.
[19] LEWIS M, LIU Y, GOYAL N, et al. Bart: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[J]. arXiv Preprint, arXiv:1910.13461, 2019.
[20] XUE L, CONSTANT N, ROBERTS A, et al. mT5: a massively multilingual pre-trained text-to-text transformer[J]. arXiv Preprint, arXiv:2010.11934, 2020.
[21] YE Q, XU H, XU G, et al. mplug-owl: modularization empowers large language models with multimodality[J]. arXiv Preprint, arXiv:2304.14178, 2023.
[22] TOUVRON H, MARTIN L, STONE K, et al. Llama 2: open foundation and fine-tuned chat models[J]. arXiv Preprint, arXiv:2307.09288, 2023.
[23] PENEDO G, MALARTIC Q, HESSLOW D, et al. The refined web dataset for Falcon LLM: outperforming curated corpora with web data, and web data only[J]. arXiv Preprint, arXiv:2306.01116, 2023.
[24] BELLAGENTE M, TOW J, MAHAN D, et al. Stable LM 2 1.6 B technical report[J]. arXiv Preprint, arXiv:2402.17834, 2024.
[25] KOUBAA A. GPT-4 vs. GPT-3.5: a concise showdown[J]. Preprints. org, 2023, 2023030422.

(上接第 501 页)

[27] NTURAMBIRWE J F I, PEROLD W J, OPARA U L. Classification learning of latent bruise damage to apples using shortwave infrared hyperspectral imaging[J]. Sensors (Basel), 2021, 21(15):4990.
[28] KIM G, KIM G H, PARK J M, et al. Application of infrared lock-in thermography for the quantitative evaluation of bruises on pears[J]. Infrared Physics and Technology, 2014, 63: 133-139.
[29] VARITH J, HYDE G M, BARITELLE A L, et al. Non-contact bruise detection in apples by thermal imaging[J]. Innovative Food Science and Emerging Technologies, 2003, 4(2): 211-218.
[30] DOOSTI-IRANI O, GOLZARIAN M R, AGHKHANI M H, et al. Development of multiple regression model to estimate the apple's bruise depth using thermal maps[J]. Postharvest Biology and Technology, 2016, 116: 75-79.
[31] BARANOWSKI P, MAZUREK W, WITKOWSKA-WALCZAK B, et al. Detection of early apple bruises using pulsed-phase thermography[J]. Postharvest Biology and Technology, 2009, 53(3): 91-100.
[32] MERZLYAK M N, SOLOVCHENKO A E, GITELSON A A. Reflectance spectral features and non-destructive estimation of chlorophyll, carotenoid and anthocyanin content in apple fruit[J]. Postharvest Biology and Technology, 2003, 27(2): 197-211.
[33] ABBASPOUR-GILANDEH Y, SABZI S, HERNáNDEZ-HERNáNDEZ M, et al. Nondestructive estimation of the chlorophyll b of apple fruit by color and spectral features using different methods of hybrid artificial neural network[J]. Agronomy, 2019, 9(11): 735.
[34] CENTENO C A R, ALBERTO M C R, WASSMANN R, et al. Assessing diel variation of CH_4 flux from rice paddies through temperature patterns[J]. Atmospheric Environment, 2017, 167: 23-39.
[35] HAYS G C, CHIVERS W J, LALOË J O, et al. Impact of marine heatwaves for sea turtle nest temperatures[J]. Biology Letters, 2021, 17(5): 20210038.
[36] ZHANG Zhewen, WU Lifeng. Graph neural network-based bearing fault diagnosis using granger causality test[J]. Expert Systems with Applications, 2024, 242: 122827.
[37] ANANDAN S, RUDOLPH A, SPECK T, et al. Comparative morphological and anatomical study of self-repair in succulent cylindrical plant organs[J]. Flora, 2018, 241: 1-7.
[38] SPECK O, SPECK T. An overview of bioinspired and biomimetic self-repairing materials[J]. Biomimetics, 2019, 4(1): 26.
[39] SPECK O, LANGER M, MYLO M D. Plant-inspired damage control-an inspiration for sustainable solutions in the anthropocene[J]. The Anthropocene Review, 2021, 9(2): 220-236.
[40] GONZALEZ M E, BARRETT D M. Thermal, high pressure, and electric field processing effects on plant cell membrane integrity and relevance to fruit and vegetable quality[J]. Journal of Food Science, 2010, 75(7): 121-130.
[41] BAUM C F, HURN S, OTERO J. Testing for time-varying granger causality[J]. The Stata Journal, 2022, 22(2): 355-378.