

面向同分布数据的玉米产业链命名实体识别研究

郝燕¹, 杨婉霞¹, 周蓓蓓¹, 芦晶¹, 袁炜炜¹, 汤敏睿^{2,3}

(1. 甘肃农业大学机电工程学院, 甘肃兰州 730070; 2. 新疆大学计算机科学与技术学院; 3. 新疆多语种信息技术重点实验室, 新疆乌鲁木齐 830017)

摘要: 为探究数据分布特征对命名实体识别模型性能的影响, 提出数据同分布特性增强方法, 以提高模型识别性能。首先, 采用K-S检验和KDE分布图对实验数据语料进行同分布验证, 并构建包含2 100份样本的玉米产业链语料库。其次, 设计基于BERT-BiLSTM-Attention-CRF的多层命名实体识别模型, 采用BIO、BMES、BIOES 3种标注策略, 比较验证所提方法与其他方法的识别效果; 同时对玉米产业链实体进行细粒度划分, 验证数据同分布特征对BERT-BiLSTM-Attention-CRF模型识别性能的影响。实验结果表明, 同分布特征最明显的玉米品种的F1分数达到96.61%。为进一步验证所提方法的泛化性, 在6种数据集上分析数据同分布特征的强弱对模型识别效果的影响, 探究pvalue与识别结果的关系。最后, 针对同分布特征不明显的数据(玉米病、虫、草害)进行重采样以强化数据同分布特征, 其F1分数分别提高了5.71%、4.30%、11.78%。实验结果显示, pvalue较大的实体类别对应的识别结果F1分数较高, 反之亦然。通过重采样强化数据同分布特征后, 模型识别效果显著提升, 可见数据的同分布特征直接影响模型的实体识别效果。研究结果为命名实体识别的数据分析提供了思路与方法。

关键词: 玉米产业链; 命名实体识别; 同分布; 重采样

DOI: 10.11907/hjtk.241385

中图分类号: TP391

文献标识码: A

开放科学(资源服务)标识码(OSID):

文章编号: 1672-7800(2025)006-0049-09



Research on Named Entity Recognition of the Corn Industry Chain for Identically Distributed Data

HAO Yan¹, YANG Wanxia¹, ZHOU Beibei¹, LU Jing¹, YUAN Weiwei¹, TANG Minrui^{2,3}

(1. Mechanical and Electrical Engineering College, Gansu Agricultural University, Lanzhou 730070, China; 2. School of Computer Science and Technology, Xinjiang University; 3. Xinjiang Key Laboratory of Multilingual Information Technology, Urumqi 830017, China)

Abstract: To explore the impact of data distribution characteristics on the performance of named entity recognition models, a method is proposed to enhance data homogeneity characteristics to improve model recognition performance. Firstly, K-S test and KDE distribution map were used to verify the homogeneity of the experimental data corpus, and a corn industry chain corpus containing 2100 samples was constructed. Secondly, design a multi-layer named entity recognition model based on BERT-BiLSTM-Attention-CRF, using three annotation strategies: BIO, BMES, and BIOES, to compare and verify the recognition performance of the proposed method with other methods; At the same time, fine-grained segmentation of the corn industry chain entities is carried out to verify the impact of data distribution characteristics on the recognition performance of the BERT-BiLSTM-Attention-CRF model. The experimental results showed that the F1 score of the corn variety with the most obvious distribution characteristics reached 96.61%. To further validate the generalization of the proposed method, the influence of the strength of data homogeneity distribution characteristics on the recognition performance of the model was analyzed on six datasets, and the relationship between p-value and recognition results was explored. Finally, resampling was performed on data with unclear distribution characteristics (such as corn diseases, pests, and grass damage) to enhance the data's distribution characteristics. The F1 scores increased by

收稿日期: 2024-04-18

扫描二维码阅读全文:



基金项目: 科技创新2030-“新一代人工智能”重大项目(2022ZD0115801)

作者简介: 郝燕(1996-), 女, 甘肃农业大学机电工程学院硕士研究生, 研究方向为知识图谱; 杨婉霞(1979-), 女, 博士, 甘肃农业大学机电工程学院教授, 研究方向为信息处理与安全。本文通讯作者: 杨婉霞。

5.71%, 4.30%, and 11.78%, respectively. The experimental results show that entity categories with higher p-values correspond to higher F1 scores in recognition results, and vice versa. After resampling and enhancing the data's identically distributed features, the model's recognition performance is significantly improved, indicating that the identically distributed features of the data directly affect the model's entity recognition performance. The research results provide ideas and methods for data analysis in named entity recognition.

Key Words: corn industry chain; named entity recognition; identical distribution; resampling

0 引言

智慧农业是现代农业和数字农业的发展趋势^[1]。信息技术的进步加快了智慧农业发展的步伐,促进了农业的转型升级。从农业大数据中挖掘信息、提取知识是发展现代农业的关键要素之一。作为主粮作物之一的玉米,其种植、改良以及加工技术越来越受到重视,科学地对其进行种植与管理成为必然。目前,利用机器学习方法从大量且分散的玉米产业链数据中提取信息,将其加工成知识链,进而指导玉米种植者进行规范生产、科学管理是知识服务农业的重要途径。然而,玉米产业链中的实体复杂多样,且缺乏明确的定义和描述,为机器学习正确识别与理解玉米产业链中的实体带来挑战。

命名实体识别(Named Entity Recognition, NER)作为知识抽取、知识图谱、实体链接等研究的核心内容,近年来受到各领域学者的广泛关注,主要分为基于规则的方法、基于统计的机器学习方法和基于深度学习的方法几类^[2-4]。基于规则的方法因需人工制定规则且移植性较差而应用较少。基于统计的机器学习方法一度成为NER的主流,常用模型包括最大熵模型(Maximum Entropy Model, MEM)、隐马尔可夫模型(Hidden Markov Model, HMM)、支持向量机(Support Vector Machines, SVM)和条件随机场(Conditional Random Field, CRF)^[5-8]。例如,钱邵捷^[9]基于CRF对于农业实体进行了识别,并比较了BIO和BMES两种实体标注方式的效果。李想等^[10]根据农作物病虫害及农药命名实体的特点设定自动标注策略,结合CRF对农作物病虫害和农药等实现了实体识别。然而,基于统计的机器学习模型需要人工标注大量语料,耗时耗力。BiLSTM(Bidirectional Long Short-Term Memory)-CRF模型的出现拉开了深度学习技术应用于NER的序幕,研究者开始尝试使用深度神经网络构建模型实现农作物语料的实体识别^[11]。例如,宋林鹏等^[12]将词向量作为BiLSTM-CRF模型的输入,成功识别了农业技术文本中的实体。郭旭超^[13]将Word2vec预训练模型、BiLSTM-Attention与卷积神经网络(Convolutional Neural Networks, CNN)相结合,通过提取上下文和局部特征实现了农业病虫害实体的识别,F1分数达到94.32%。Guo等^[14]采用不同卷积核大小的CNN提取局部上下文特征,利用BiLSTM-CRF模型并结合Self-Attention机制准确识别了农业病虫害领域的实体,在AgCNER数据集上的F1分数达到94.15%。Zhang等^[15]将农业领域字典与相似词汇合并输入到BiLSTM-CRF模型中进

行字符增强,对苹果病虫害语料库中的21类实体实现了准确识别。任妮等^[16]将BERT(Bidirectional Encoder Representations from Transformers)模型应用于番茄病虫害领域的实体识别中,有效降低了因分词错误带来的影响。李书琴等^[17]提出嵌入词汇信息的BERT-CRF玉米育种实体关系联合抽取方法,F1分数达到93.8%。

目前,深度学习技术在NER领域中的应用正不断完善,然而大多数模型的应用前提为假设训练数据与测试数据同分布。实际上,不同来源的数据分布存在差异。如果测试数据与训练数据的分布不一致,会很大程度上影响模型识别效果。例如,有些模型可能在训练集上表现良好,但在测试集上效果不佳。因此,在模型训练前对训练模型的数据分布进行分析很有必要^[18]。然而,农业领域的NER任务缺乏公开、标准的数据语料库,且农业领域实体的构词复杂多样,而现有NER模型专业性不强,达不到预期效果。

针对农业产业链实体识别难度高、语义复杂,而现有识别方法标注策略单一、数据特征挖掘不够深入、系统知识获取不完整等问题,本研究首先对收集到的玉米产业链语料进行数据同分布特征挖掘,并采用3种标注策略进行标注;其次,使用BERT-BiLSTM-Attention-CRF(本文模型)、BERT-BiLSTM-CRF、Word2vec-BiLSTM-CRF、TENER(Transformer Encoder for NER)和BiLSTM-CRF 5种深度学习模型进行比较实验,分析同分布特征的数据集在不同标注策略下的NER准确性;最后,对同分布特征不明显的数据进行重采样处理,以加强数据同分布特征,提升模型对玉米产业链语料中实体识别的准确率。本研究的主要贡献点为:①构建了玉米产业链专业数据集。由于没有公开的玉米产业链数据语料,本文利用爬虫等技术获取网站及网页的大量非结构化数据,对其进行预处理后通过人工标注构建了包含2 100份样本,共计37万余字的玉米产业链专业语料库,作为本文NER数据集;②研究了数据分布特征对NER模型性能的影响。现有NER研究仅关注模型结构的设计,忽略了对数据分布特征的分析,而数据分布特征对模型识别性能有一定影响。因此,本文在数据预处理阶段采用统计学K-S检验和KDE(Kernel Density Estimation)分布图的方法对数据集进行同分布验证与分析。实验结果表明,数据同分布特征越强,模型在该数据集上的实体识别准确率越高。对玉米病、虫、草害领域数据集进行重采样处理可有效解决数据分布不一致问题,显著提升模型识别效果。

1 NER模型构建

本文构建了BERT-BiLSTM-Attention-CRF深度学习模型用于玉米产业链NER任务,结构如图1所示。该模型主要由BERT层、BiLSTM层、Attention层和CRF层组成。首先,将玉米产业链文本数据输入到BERT中以得到含有丰富语义特征的词向量;其次,词向量经过BiLSTM进行语义编码处理后进一步与Attention层相结合,从而关注重点信息;最后,由CRF计算得到模型预测的标签序列。例如,输入为“宜北京种植。”输出为对应的预测标签。

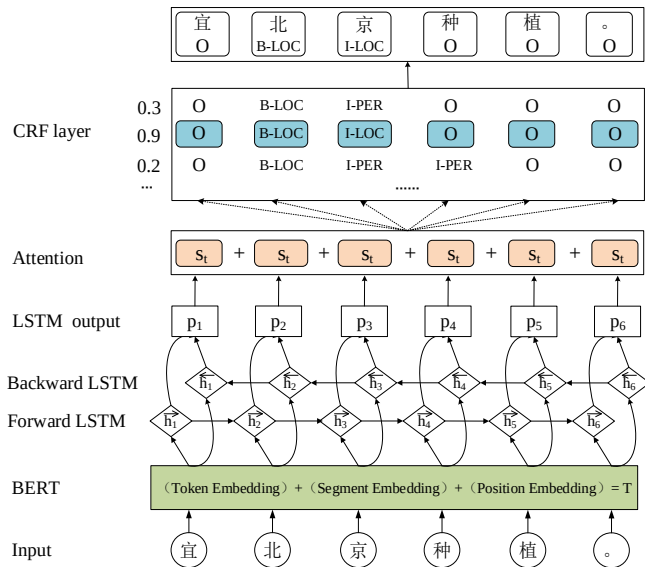


Fig. 1 BERT-BiLSTM-Attention-CRF model structure

图1 BERT-BiLSTM-Attention-CRF模型结构

BERT预训练模型的网络主体为Transformer结构^[19]。该模型对输入语料进行分词处理,在句首添加[CLS],句间使用[SEP]进行分割,得到每个词向量的3个部分,分别为词表征(Token Embedding)、段表征(Segment Embedding)和位置表征(Position Embedding)。将这3部分向量之和输入到Transformer网络中,输出的词向量T作为BiLSTM的输入。BiLSTM的隐藏状态为 $h=(h_1, h_2, \dots, h_n)$,将前、后向状态加权求和后自动提取句子特征,记为矩阵 $p=(p_1, p_2, \dots, p_n)$ 。BiLSTM层通过充分学习词向量中句子序列的上下文信息获取其全局特征,通过门控装置有效缓解了循环神经网络(Recurrent Neural Network, RNN)无法处理长距离依赖的问题,且在一定程度上解决了梯度消失和梯度爆炸问题^[20-22]。

BERT和BiLSTM虽然解决了全局特征信息提取以及长距离依赖问题,但是难以保留局部特征信息,因此本文引入Attention机制,旨在提升局部特征的识别效果,其核心思想为带权求和。首先,计算Query与Key之间的相似度,获得初始权值 e_t ;其次,对 e_t 进行归一化,得到可直接用于后续计算的权重向量 α_t ;最后,利用 α_t 对特征向量集合 h_t 进行加权求和,获得序列中关键信息的表示 s_t ,从而专注

于重要信息。Attention机制的计算公式为:

$$e_t = v_t \tanh(w_t h_t + b_t) \quad (1)$$

$$\alpha_t = \frac{\exp(e_t)}{\sum_{c=1}^n e_c} \quad (2)$$

$$s_t = \sum_{i=1}^k \alpha_i h_i \quad (3)$$

式中: v_t 和 w_t 为权重矩阵; b_t 为偏移量; h_t 为隐层状态向量; e_t 为 t 时刻的能量值。

CRF层用于解决序列标注任务中标签之间的依赖性问题。假设文本序列 $X=(x_1, x_2, \dots, x_n)$ 及其对应标签序列 $y=(y_1, y_2, \dots, y_n)$ 均为随机变量,在给定 X 时, y 的条件概率分布 $P(y|X)$ 构成CRF,满足马尔可夫性,则输入序列的标签转移分数 $\text{Score}(X, y)$ 可表示为式(4)。通过对可能的输出序列 y 的得分进行归一化,得到 y 的概率分布,表示为式(5)。训练过程中采用极大似然估计寻求最大化 $P(y|X)$ 的参数,即 $\log(P(y^*|X))$,计算公式为式(6)。

$$\text{Score}(X, y) = \sum_{i=1}^n (W_{y_{i-1}, y_i} + P_{i, y_i}) \quad (4)$$

$$P(y|X) = \frac{e^{\text{Score}(X, y)}}{\sum_{i=0}^n e^{\text{Score}(X, y)}} \quad (5)$$

$$\log(P(y^*|X)) = \text{Score}(X, y^*) - \log\left(\sum_{i=0}^n e^{\text{Score}(X, y)}\right) \quad (6)$$

式中: P_{i, y_i} 表示第 i 个字被预测为标签 y_i 的概率; W_{y_{i-1}, y_i} 表示标签 y_{i-1} 转移标签 y_i 的分数; y^* 表示极大似然估计。

2 数据集构建

玉米产业链知识库的构建需要大量数据支撑。利用爬虫技术从中国作物种质资源信息网(<https://www.cgris.net/>)的作物病虫害知识模块、种业商务网(<https://www.chinaseed114.com/>)的品种模块等获取玉米产业链数据,构成本文的数据语料。该数据语料具有多样性和复杂性,包括结构化数据、半结构化数据和非结构化数据,本文主要以非结构化数据为研究对象开展实验。非结构化数据需要进一步处理。首先,对爬虫所得数据进行去重、去空格、脱敏等基础预处理;其次,采用人工方式对数据进行标注;最后,将已标注的字符序列转换为模型需要的格式。数据预处理流程如图2所示。

2.1 数据结构分析

为检验数据是否同分布并分析同分布特征强弱对实验结果的影响,本文采用KDE法分析玉米产业链中品种数据语料的同分布情况,因为KDE分布图可以直观展现数据样本的分布特征^[23]。首先,将train_set.txt和test_set.txt转化为稠密的词向量表示;其次,通过核函数计算将转化后的词向量映射到有信息量的空间中;最后,基于以上数据得到KDE分布图。具体示例如图3所示。可以看出,train_set.txt和test_set.txt有较大的重叠区域,峰值位置和数量相似,分布形状也相似,说明二者具有同一分布性。采用K-S检验比较train_set.txt与test_set.txt分布之间的差



Fig. 2 Data preprocessing flow

图2 数据预处理流程

异,评价参数为 statistic 值和 pvalue,其中 statistic 值表示 train_set.txt 与 test_set.txt 之间的最大差异;pvalue 表示在原假设为真的情况下,观察量出现的概率^[24]。假设玉米品种训练集 train_set.txt 与测试集 test_set.txt 的差异不显著,检验结果 pvalue 为 0.220 5(大于显著性水平 0.05),表明在原假设为真的情况下,观察量出现的概率为 0.220 5。因此,在 0.05 的显著性水平下,无法拒绝原假设,说明两组样本分布之间的差异不显著,即 train_set.txt 与 test_set.txt 具有同一分布性。

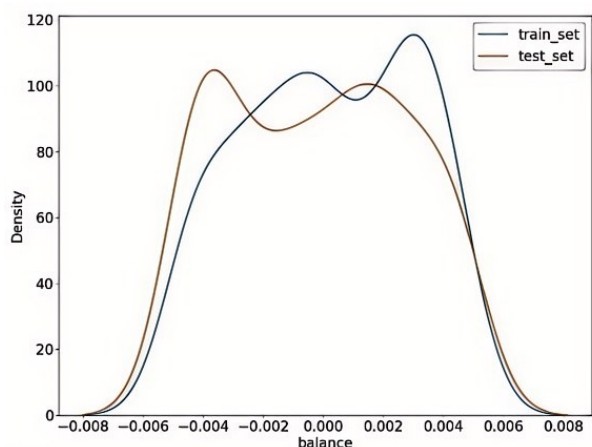


Fig. 3 Example of KDE distribution diagram

图3 KDE 分布图示例

为全面分析数据语料的分布特征,确保后续实验的有效性和严谨性,本文对玉米产业链数据集集中的病害、虫害、草害等数据语料,以及公共数据集 Resume(中国股市上市高管的简历数据集)和 CMeEE(医学实体识别数据集)进行 K-S 检验。结果如表 1 所示。可以看出,6 类数据语料之间差异显著。其中,玉米品种 K-S 检验的 statistic 值为 0.105,表明玉米品种 train_set.txt 与 test_set.txt 之间的差异性较小;K-S 检验的 pvalue 为 0.220 5(>0.05),接受原假设,说明数据差异极不显著,即同分布的概率较大。纵观玉米病害、虫害、草害样本的 statistic 值,均大于玉米品种的 statistic 值,表明玉米病害、虫害、草害数据的 train_set.txt 与 test_set.txt 差异性较小,但略大于玉米品种的差异性。两类公共数据集中 train_set.txt 与 test_set.txt 的差异性大于玉米品种,但小于玉米病害、虫害、草害数据。由此可见,玉米品种数据 train_set.txt 与 test_set.txt 之间的差异性最小,同分布特征最强。

Table 1 K-S test results of six types of data corpus

表 1 6 类数据语料的 K-S 检验结果

数据集	数量	Statistic	pvalue
玉米品种	600	0.105	0.220 5
玉米病害	374	0.189	0.132 1
玉米虫害	213	0.163	0.147 6
玉米草害	104	0.247	0.144 9
Resume	100	0.143	0.188 4
CMeEE	100	0.145	0.189 8

2.2 实体类型定义

农业科学叙词表是具有科学性、规范性和实用性的农业信息检索工具,可为农业文本实体分类提供科学依据^[25]。参考农业科学叙词表,同时结合 AgCNER 本体框架的实体类型定义以及本文数据的特点,将玉米产业链的实体分为 12 类^[26]。具体如表 2 所示。

Table 2 Types of entities in the corn industry chain

表 2 玉米产业链实体类型

实体类型	释义	实例
育种	育种技术	单倍体育种、二环系育种
品种	品种名称	935HA、农华 101、登海 605
整地	种植前对土地的处理	翻地、耙地、耨地
种植	种植地选择、种植方式	宽窄行种植、等行距种植
病害	疾病名称	黑炭病、褐斑病、赤霉病
虫害	虫害名称	红蜘蛛、灰翅夜蛾
草害	草害名称	刺儿菜、白草、马齿苋
农药	杀虫剂、杀菌剂	三唑酮、波尔多液
肥料	有机肥、无机肥	3% 硫酸水溶液
农机	农业生产中所用的机械设备	植保机、拖拉机、收割机
销售	销售方式	批发、零售、电商、物流
加工	加工产品及方式	玉米油、人工及机械加工

2.3 数据标注

数据标注是人工智能算法得以有效运行的关键环节,通常分为人工标注、自动标注、半自动标注、众包标注 4 类。其中,人工标注质量高,适合复杂任务;自动标注处理大数据集速度快且成本低,但准确度一般;半自动标注结合了两者的优点,可提高效率,但需人工验证,成本较高;众包标注有多人参与,能提高标注一致性,但不适合敏感任务。综合考虑,本文采用人工标注方法以确保数据质量,该方法的前提是选择合适的标注工具。本文分析比较了多款文本标注工具,包括 YEDDA、Brat、Doccano、MarkTool、Chinese Annotator 和精灵标注助手。其中,YEDDA 为轻量级的高效开源工具,专注于实体类型标注,支持快捷键标

注,但实体类型定义有限;精灵标注助手功能全面、操作简便、界面友好,但线下读取数据限制多人协作,数据易丢失且路径不可更改;Brat 适用于 Linux 系统,支持多种类型标注,尤其在关系标注方面表现出色;MarkTool 为一款通用标注工具,支持多种标注功能,且持续维护,功能全面,但

需要 docker 镜像部署;Chinese Annotator 为一款智能中文标注工具,标注流程简洁,效率较高,但需要安装 MongoDB;Doccano 功能全面,支持多种标注任务和多人协助标注,但不支持某些事件特定关系标注。基于以上分析,本文选择 Doccano 标注工具,其标注界面如图 4 所示。



Fig. 4 Doccano annotation interface

图4 Doccano标注界面

实验数据标注完成后需要转化为模型所需的格式。针对农业数据结构复杂、专业性强等特点,结合数据量大小以及本文数据集的特征,拟采用 BIO、BMES 和 BIOES 3 种序列标注策略,在 NER 模型上分别测试其性能,进而确定最优标注策略^[27]。3 种标注策略的示例如图 5、图 6 所示。其中,B(Begin)用于标注实体名称的首字,I(Intermediate)用于标注实体名称的中间字;M(Middle)用于标注实体名称的中间字,特别是对于多个字组成的嵌套实体;O(Other)用于标记不属于任何实体的字;E(End)用于标注实体名称的尾字;S(Single)用于标注单个字的实体。

3 实验方法与结果分析

3.1 实验环境与实验超参数配置

具体实验环境设置如表 3 所示。在对 Bert-BiLSTM-Attention-CRF 模型进行训练和调优的过程中设定了主要

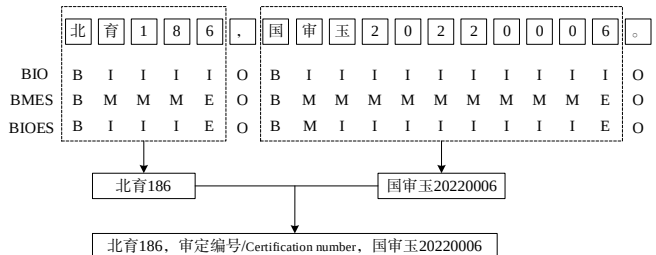


Fig. 5 Example of corpus annotation without single entities

图5 不含单实体语料标注示例

超参数。具体配置详见表 4。

3.2 评价标准

采用精确率(Precision, P)、召回率(Recall, R)和 F1 值评价模型性能。其中,精确率表示标注结果正确的数量占所有结果的比例,反映模型识别的可信度;召回率表示标注结果正确的数量占所有命名实体数量的比例,反映模型识别的完整度;F1 值为以上两者的调和平均值,可综合评

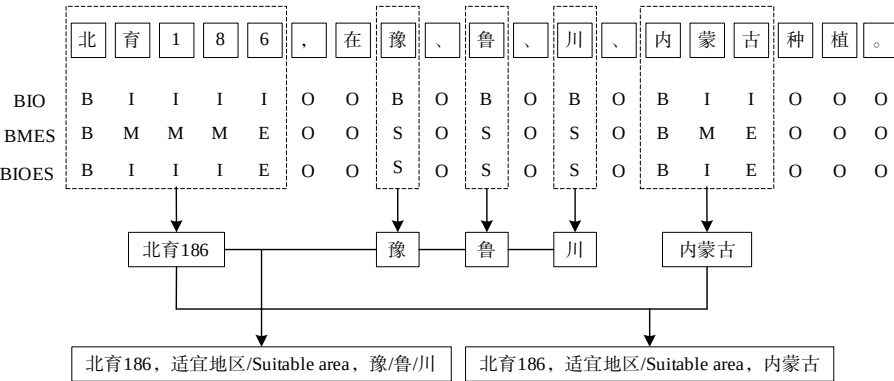


Fig. 6 Example of corpus annotation containing single entities

图6 含单实体语料标注示例

Table 3 Experimental environment settings

表3 实验环境设置

实验环境	配置信息
操作系统	Linux Ubuntu
Python	3.7.0
CUDA	10.1
GPU	NVIDIA GeForce RTX 2080 Ti
TensorFlow	2.2.0
显存	11G

Table 4 BERT-BiLSTM-Attention-CRF model hyperparameters

表4 BERT-BiLSTM-Attention-CRF 模型超参数

实验环境	配置信息
embedding_size	128
hidden_size	128
dropout	0.5
epoch	50
batch_size	16
learning-rate	0.001

估模型的整体性能。3个评价指标的计算公式为：

$$P = \frac{TP}{TP + FP} \times 100\% \quad (7)$$

$$R = \frac{TP}{TP + FN} \times 100\% \quad (8)$$

$$F1 = 2 \times \frac{PR}{P + R} \times 100\% \quad (9)$$

式中：TP指模型正确识别并预测为实体的数量；FN指模型未能识别但实际存在的实体数量；FP指模型错误地将非实体预测为实体的数量。

3.3 实验结果与分析

鉴于玉米产业链数据特征的复杂性,本文首先进行不同标注策略对不同模型性能的影响实验,以筛选出识别效果较优的标注策略及适配模型;其次,通过细粒度划分玉米产业链实体进一步验证 BERT-BiLSTM-Attention-CRF 模型对细粒度实体的识别性能;再次,为证明实验数据的同分布特征对NER模型的重要性,针对不同领域的数据集设计了比较实验;最后,对同分布特征不明显的玉米病、虫、草害领域数据进行重采样处理,以强化其同分布特征,并进行了相应实验。本文实验均以随机抽样的方式将数据集以6:2:2的比例划分为训练集(train.txt)、测试集(test.txt)和验证集(dev.txt)。

3.3.1 不同模型在不同标注策略下的性能比较

为获得适合本文数据特征的标注策略,以自建玉米产业链语料库为数据集,分别验证 BIO、BMES、BIOES 3种标注策略在 BERT-BiLSTM-Attention-CRF、BERT-BiLSTM-CRF、TENER、BiLSTM-CRF、Word2vec-BiLSTM-CRF 模型上的识别效果。实验结果如表5所示。TENER是基于Transformer的具有方向与相对位置信息的改进模型,其语义提取特征能力远强于LSTM^[28-29]。Word2vec将输入的文本序列转换为词向量,保留了上下文的语义信息,可帮助模型更好地理解输入序列语义^[30-31]。以上对比模型均为NER的典型模型。为了更好地理解与分析实验结果,对其

进行三维可视化,具体如图7所示。

Table 5 Performance comparison of different models under different annotation strategies

表5 不同模型在不同标注策略下的性能比较 (%)

标注策略	模型	P	R	F1
BIO	TENER	60.25	60.25	60.25
	BiLSTM-CRF	62.94	65.73	64.30
	Word2vec-BiLSTM-CRF	64.25	68.94	66.51
	BERT-BiLSTM-CRF	73.20	65.82	69.31
	BERT-BiLSTM-Attention-CRF	70.54	72.40	71.46
BMES	TENER	58.00	59.24	58.61
	BiLSTM-CRF	67.84	70.73	69.25
	Word2vec-BiLSTM-CRF	70.25	69.78	70.01
	BERT-BiLSTM-CRF	73.19	68.14	70.57
	BERT-BiLSTM-Attention-CRF	69.92	73.25	70.49
BIOES	TENER	57.00	57.00	57.00
	BiLSTM-CRF	60.00	62.14	61.05
	Word2vec-BiLSTM-CRF	65.29	70.68	67.88
	BERT-BiLSTM-CRF	68.17	69.35	68.75
	BERT-BiLSTM-Attention-CRF	58.73	68.45	63.22

由表5和图7可知,本文模型 BERT-BiLSTM-Attention-CRF 在3种标注策略下均有较好表现。采用BIO标注数据时,BERT-BiLSTM-Attention-CRF模型的实体识别率高于TENER、BiLSTM-CRF、Word2vec-BiLSTM-CRF、BERT-BiLSTM-CRF 4个模型,F1分数分别提高了11.21%、7.16%、4.95%和2.15%。虽然BMES和BIOES标注策略能够更加明确实体边界,提供更多信息,但其需要预测的标签数量更多,这在一定程度上增加了模型识别难度,导致识别效果受到影响。因此,在综合考虑标注策略的准确性和模型的复杂性后,本文选择BIO标注策略进行后续实验。

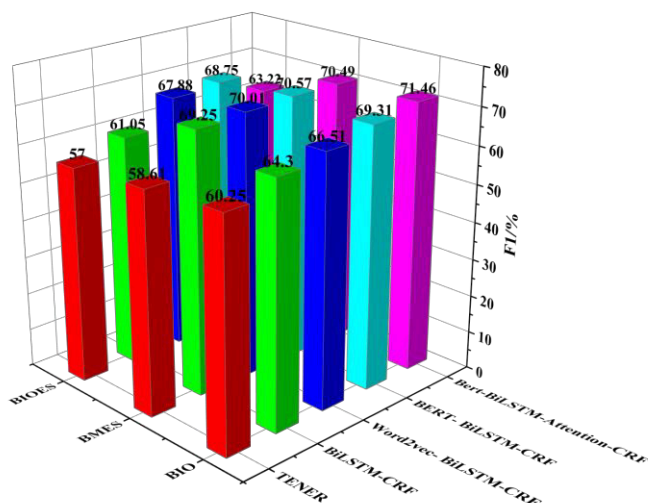


Fig. 7 Visualization of model recognition performance under different annotation strategies

图7 不同标注策略下模型识别效果可视化

3.3.2 BERT-BiLSTM-Attention-CRF 模型对玉米产业链实体的识别性能

为进一步验证BIO标注策略与BERT-BiLSTM-Atten-

tion-CRF 模型的适配度,对玉米产业链实体进行细粒度划分,采用 BERT-BiLSTM-Attention-CRF 模型对 BIO 策略标注后的玉米产业链各实体进行识别。实验结果如表 6 所示。可以看出,在 12 种实体类型中,BERT-BiLSTM-Attention-CRF 模型对玉米品种的识别效果最佳, P 和 $F1$ 分数分别达到 95.93% 和 96.61%,在其他实体上表现一般,主要原因是各实体数据自身分布特征的差异性,这一点可通过表 1 中的实验结果被证明。由于玉米品种的 p value 最大(即同分布特征最强),BERT-BiLSTM-Attention-CRF 模型对其的识别效果最好。其他实体数据来源复杂,识别效果一般。此外,在 BiLSTM-CRF 模型中加入 BERT 后,对数据量的要求随之增加,数据量少的实体不占优势。

Table 6 Recognition effect of BERT-BiLSTM-Attention-CRF model on different entity types

表 6 BERT-BiLSTM-Attention-CRF 模型对不同实体类型的识别效果				
实体类型	数量	$P/\%$	$R/\%$	$F1/\%$
育种	120	54.20	75.53	63.11
品种	600	95.93	97.29	96.61
整地	118	54.84	62.96	58.62
种植	116	56.25	72.97	63.53
病害	374	60.97	76.60	67.90
虫害	213	58.73	73.51	65.29
草害	104	84.34	61.36	71.05
农药	50	37.39	68.25	48.31
肥料	70	70.92	83.28	76.60
农机	100	52.99	68.13	59.62
销售	115	49.75	67.20	57.17
加工	120	56.25	69.23	62.07

3.3.3 不同领域数据集比较

为突出同分布数据的有效性,在同分布特征强弱不同的数据集上进行比较实验。为确保实验的严谨性,选取自建的玉米品种、病害、虫害、草害领域数据集以及公开数据集 Resume、CMeEE。从各领域数据集中随机抽取 100 条样本进行验证,均采用 BIO 标注策略。由于数据量相对较小,选用 Word2vec-BiLSTM-CRF 模型进行实验。实验结果如表 7 所示。各数据集的 K-S 分析与识别结果如图 8 所示。

Table 7 Comparative experiments on datasets from different fields
表 7 不同领域数据集比较实验

数据集	数量	$P/\%$	$R/\%$	$F1/\%$
玉米品种	100	92.68	95.00	93.83
玉米病害	100	53.80	60.90	57.13
玉米虫害	100	61.09	68.58	64.62
玉米草害	100	52.57	61.74	56.79
Resume	100	78.95	88.24	83.33
CMeEE	100	83.33	90.91	86.96

由表 7 和图 8 可知,模型对 6 种不同领域数据集的识别效果差异明显,其中玉米品种的识别效果相较于其他 5 种数据集更为突出。K-S 检验结果中的 p value 大小与模型识别效果有同步变化的趋势,即 p value 越大, $F1$ 分数越高,

表明同分布特征越强的数据集识别效果越好,非同分布数据会影响模型的识别准确率,验证了 NER 过程中挖掘数据特征的重要性。

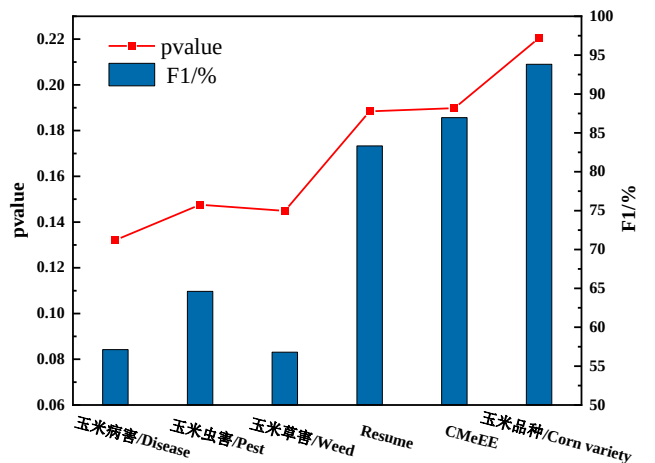


Fig. 8 Comparison of K-S test and model recognition effect
图 8 K-S 检验与模型识别效果比较

3.3.4 经重采样处理的不同领域数据集比较实验

为了实现数据同分布,通常会采用数据预处理、数据增强、重采样、领域适应等方法,其中重采样方法是为了准确获得有代表性的样本^[32]。由于当前大量涌现的高维数据普遍具有内在稀疏性,获得有代表性的样本对于依赖于同分布假设的众多机器学习研究而言具有重要的实际价值。多样化且分布不均的数据特征表现为样本种类多、差异大,因此重采样过程既要维持各类数据原有的分布特性,又要努力平衡各类样本的数量,以确保数据的多样性和均衡性得到妥善处理。本文采用重采样方法处理玉米病、虫害和草害领域数据,对稀有类型与丰富类型数据之间的比例进行微调,以增强数据的同分布特征。重采样的最优比例在很大程度上取决于所使用的数据和模型。本文在实验过程中尝试将稀有类型与丰富类型之间的比例设置为 1:1、1:3 和 2:1,最终使用 1:3 的比例进行微调。不同领域数据集重采样前后实体识别效果如图 9 所示。可以看出,采用重采样方法处理后,3 种数据集的识别效果均有所提升,相较重采样前的 $F1$ 分数分别提高了 5.71%、4.3%、11.78%,其中,玉米草害的 P 、 R 、 $F1$ 分数提升最为显著,表明重采样方法可有效平衡训练集与测试集的数据分布差异,解决了数据分布不一致问题,提高了模型识别性能。

4 结语

本文以玉米产业链数据为研究对象,深入分析了数据分布特征、标注策略对模型识别性能的影响。具体来说,在数据集的预处理阶段采用 K-S 检验和 KDE 分布图的方法对数据进行了同分布验证,构建了玉米产业链的数据语料库;确定了标注策略与模型的最佳匹配模式,即 BIO 标

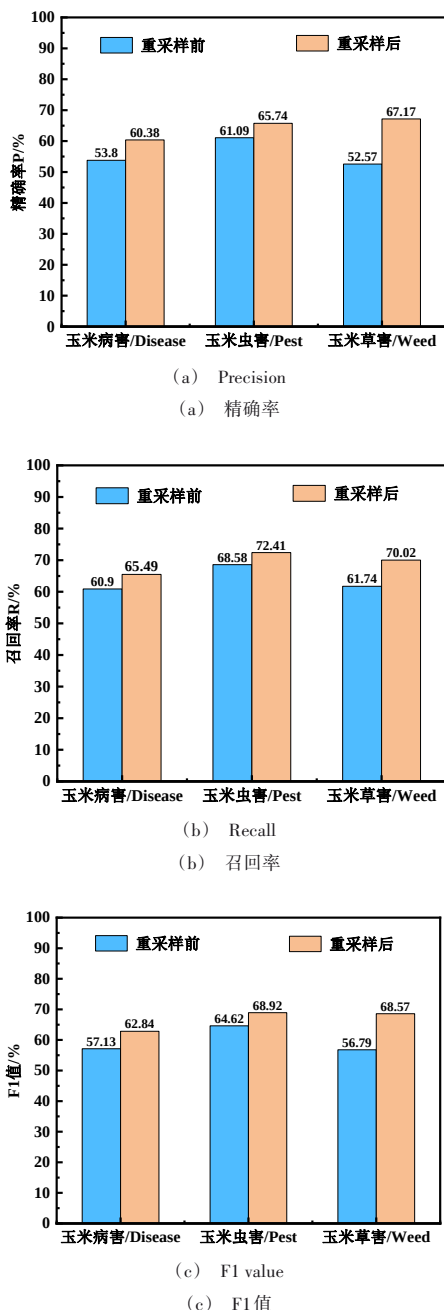


Fig. 9 Comparison of entity recognition performance before and after resampling on datasets from different fields

图9 不同领域数据集重采样前后实体识别效果比较

注策略下的BERT-BiLSTM-Attention-CRF模型;采用最佳标注策略和模型对玉米产业链12类实体进行了识别,同时采用不同种类数据集在同一模型上进行了比较,发现数据同分布特征越强,模型实体识别效果越好;针对同分布特征不明显的玉米病、虫、草害数据集进行重采样处理,有效提高了模型识别性能,证实了数据处理与特征挖掘对于NER任务的重要性。

本文提出的BIO标注策略下的BERT-BiLSTM-Attention-CRF模型可高效地从玉米产业链“耕、种、管、收”大数据中抽取知识,进而构建玉米产业链知识图谱,为玉米的

科学化种植与管理提供专业指导。然而,玉米产业链之间具有高度复杂的相关性,需要更具有针对性的模型和算法。下一步可通过结合迁移学习方法与模型训练技术来构建更准确可靠的玉米产业链NER模型,并将其应用于实际知识图谱构建,助力智慧农业快速发展。

参考文献:

- [1] DEBANGSHI U, SADHUKHAN A, DUTTA D, et al. Application of smart farming technologies in sustainable agriculture development: a comprehensive review on present status and future advancements[J]. International Journal of Environment and Climate Change, 2023, 13(11): 3689-3704.
- [2] LI J, SUN A X, HAN J L, et al. A survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(1): 50-70.
- [3] FRIEDMAN C, ALDERSON P O, AUSTIN J H M, et al. A general natural-language text processor for clinical radiology[J]. Journal of the American Medical Informatics Association, 1994, 1(2): 161-174.
- [4] AJEES A P, K M, IDICULA M S. An improved word representation for deep learning based NER in Indian languages [J]. Information, 2019, 10(6): 186.
- [5] MURATA M, UCHIMOTO K, UTIYAMA M, et al. Using the maximum entropy method for natural language processing: category estimation, feature extraction, and error correction[J]. Cognitive Computation, 2010, 2(4): 272-279.
- [6] GÁMIZ M, NAVAS-CÓMEZ F, RAYA-MIRANDA R, et al. Dynamic reliability and sensitivity analysis based on HMM models with Markovian signal process [J]. Reliability Engineering and System Safety, 2023, 239: 109498.
- [7] ZHU F, LUO K G, CHEN W J, et al. Construction of a water/underwater navigation scene perception model combining SVM and HMM [J]. Journal of Surveying and Mapping, 2023, 52(5): 738-747.
朱锋, 罗科干, 陈惟杰, 等. 联合SVM和HMM的水上/水下导航场景感知模型构建[J]. 测绘学报, 2023, 52(5): 738-747.
- [8] ZHANG L J, CHEN L J, XU S Y, et al. Design of a structured parsing model for corporate bidding documents based on Bi-LSTM and conditional random field (CRF)[J]. International Journal of Information Technologies and Systems Approach (IJITSA), 2023, 16(2): 1-15.
- [9] QIAN S J. Research on data enhancement optimization strategies for named entity recognition[D]. Lanzhou: Lanzhou University, 2022.
钱邵捷. 面向命名实体识别的数据增强优化策略研究[D]. 兰州: 兰州大学, 2022.
- [10] LI X, WEI X H, JIA L, et al. Crop pest and pesticide named entity recognition based on conditional random fields [J]. Journal of Agricultural Machinery, 2017, 48(S1): 178-185.
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(S1): 178-185.
- [11] WEI Z J, SONG L, HU X C, et al. Agricultural named entity recognition based on entity level masking BERT and BiLSTM-CRF [J]. Journal of Agricultural Engineering, 2022, 38(15): 195-203.
韦紫君, 宋玲, 胡小春, 等. 基于实体级遮蔽BERT与BiLSTM-CRF的农业命名实体识别[J]. 农业工程学报, 2022, 38(15): 195-203.
- [12] SONG L P, LIU S H, WANG C. Extracting agricultural technology demand text entities based on word vectors+BiLSTM+CRF [J]. Jiangsu Agricultural Science, 2021, 49(5): 186-193.

- 宋林鹏,刘世洪,王翠. 基于词向量+BiLSTM+CRF的农业技术需求文本实体提取[J]. 江苏农业科学, 2021, 49(5): 186-193.
- [13] GUO X C, TANG Z, DIAO L, et al. Named entity recognition of pests and diseases based on radical embedding and attention mechanism [J]. Journal of Agricultural Machinery, 2020, 51(S2): 335-343.
- 郭旭超,唐詹,刁磊,等. 基于部首嵌入和注意力机制的病虫草害命名实体识别[J]. 农业机械学报, 2020, 51(S2): 335-343.
- [14] GUO X, ZHOU H, SU J, et al. Chinese agricultural diseases and pests named entity recognition with multi-scale local context features and self-attention mechanism [J]. Computers and Electronics in Agriculture, 2020, 179(5): 105830.
- [15] ZHANG J, GUO M, GENG Y, et al. Chinese named entity recognition for apple diseases and pests based on character augmentation [J]. Computers and Electronics in Agriculture, 2021, 190: 106464.
- [16] REN N, BAO T, SHEN G Y, et al. Research on fine-grained named entity recognition based on deep learning - taking tomato diseases and pests as an example [J]. Intelligence Science, 2021, 39(11): 96-102.
- 任妮,鲍彤,沈耕宇,等. 基于深度学习的细粒度命名实体识别研究——以番茄病虫害为例[J]. 情报科学, 2021, 39(11): 96-102.
- [17] LI S Q, PANG W T. BERT-CRF maize breeding entity relationship joint extraction method with embedded vocabulary information [J/OL]. Journal of Agricultural Machinery, 2024-04-25. <http://kns.cnki.net/kcms/detail/11.1964.S.20230919.1113.006.html>.
- 李书琴,庞文婷. 嵌入词汇信息的BERT-CRF玉米育种实体关系联合抽取方法[J/OL]. 农业机械学报, 2024-04-25. <http://kns.cnki.net/kcms/detail/11.1964.S.20230919.1113.006.html>.
- [18] WANG Y Y, LIU F G, PENG L X, et al. Fusion of AP clustering algorithm and width learning system for out of distribution hard disk fault prediction [J/OL]. Computer Science, 2024-04-25. <http://kns.cnki.net/kcms/detail/50.1075.TP.20231201.1407.018.html>.
- 王屹阳,刘发贵,彭玲霞,等. 融合AP聚类算法和宽度学习系统的分布外硬盘故障预测[J/OL]. 计算机科学, 2024-04-25. <http://kns.cnki.net/kcms/detail/50.1075.TP.20231201.1407.018.html>.
- [19] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems, 2017: 6000-6010.
- [20] ZHANG R T, MA X L, DING W P, et al. MAP-FCRNN: multi-step ahead prediction model using forecasting correction and RNN model with memory functions [J]. Information Sciences, 2023, 646: 119382.
- [21] YIN J, LIU B, SUN G B, et al. Prediction of remaining life of lithium ion batteries based on transfer learning and DAE-LSTM [J/OL]. Journal of Electrical Engineering Technology, 2024-04-25. <https://doi.org/10.19595/j.cnki.1000-6753.tces.221890>.
- 尹杰,刘博,孙国兵,等. 基于迁移学习和DAE-LSTM的锂离子电池剩余寿命预测[J/OL]. 电工技术学报, 2024-04-25. <https://doi.org/10.19595/j.cnki.1000-6753.tces.221890>.
- [22] TANG L, LI Y, TANG Y F, et al. Application of LSTM model based on deep learning in X-ray fluorescence spectroscopy [J]. Nuclear Technology, 2023, 46(7): 80-86.
- 唐琳,李勇,唐羽锋,等. 基于深度学习的LSTM模型在X荧光光谱中的应用[J]. 核技术, 2023, 46(7): 80-86.
- [23] ZHOU Y, XIA H, LIU H Y, et al. DPC clustering algorithm based on K-nearest neighbors and kernel density estimation [J/OL]. Journal of Beihang University, 2024-04-25. <https://doi.org/10.13700/j.bh.1001-5965.2023.0342>.
- 周玉,夏浩,刘虹瑜,等. 基于K互近邻与核密度估计的DPC聚类算法[J/OL]. 北京航空航天大学学报, 2024-04-25. <https://doi.org/10.13700/j.bh.1001-5965.2023.0342>.
- [24] ZHENG Z, YAN Y J, CAI J S, et al. Non specific TVLA energy leakage assessment method based on dual sample KS test [J/OL]. Journal of Communications, 2024-04-25. <http://kns.cnki.net/kcms/detail/11.2102.TN.20230523.1603.012.html>.
- 郑震,严迎建,蔡爵嵩,等. 基于双样本KS检验的非特定TVLA能量泄漏评估方法[J/OL]. 通信学报, 2024-04-25. <http://kns.cnki.net/kcms/detail/11.2102.TN.20230523.1603.012.html>.
- [25] QIAO B. Research on knowledge graph construction technology based on agricultural thesaurus [D]. Changsha: Hunan Agricultural University, 2019.
- 乔波. 基于农业叙词表的知识图谱构建技术研究[D]. 长沙: 湖南农业大学, 2019.
- [26] GUO X, ZHOU H, SU J, et al. Chinese agricultural diseases and pests named entity recognition with multi-scale local context features and self-attention mechanism [J]. Computers and Electronics in Agriculture, 2020, 179: 105830.
- [27] MAI C C, LIU J, HUANG Y H. A Chinese named entity extraction method based on multi annotation framework and fusion features [P]. China, CN202110511025. 8. 2021-07-30.
- 麦丞程,刘健,黄宜华. 一种基于多标注框架与融合特征的中文命名实体抽取方法[P]. 中国, CN202110511025. 8, 2021-07-30.
- [28] DENG Y Y, WU C X, WEI Y F, et al. A review of named entity recognition based on deep learning [J]. Chinese Journal of Information Science, 2021, 35(9): 30-45.
- 邓依依,郭昌兴,魏永丰,等. 基于深度学习的命名实体识别综述[J]. 中文信息学报, 2021, 35(9): 30-45.
- [29] LI L, XI X F, SHENG S L, et al. Research progress on deep learning Chinese named entity recognition [J]. Computer Engineering and Applications, 2023, 59(24): 46-69.
- 李莉,奚雪峰,盛胜利,等. 深度学习中文命名实体识别研究进展[J]. 计算机工程与应用, 2023, 59(24): 46-69.
- [30] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [DB/OL]. <https://arxiv.org/abs/1301.3781>.
- [31] HUANG Z H, XU W, YU K. Bidirectional LSTM-CRF models for sequence tagging [DB/OL]. <https://arxiv.org/abs/1508.01991>.
- [32] SHI M H, WU G C. Unbalanced data classification based on cluster mixed sampling [J]. Computer and Modernization, 2020(5): 34-38.
- 史明华,吴广潮. 基于聚类混合采样的不平衡数据分类[J]. 计算机与现代化, 2020(5): 34-38.

(责任编辑:尹晨茹)