

doi:10.6041/j.issn.1000-1298.2026.05.037

基于 DPNA - CASREL 的柑橘病虫害实体关系联合抽取方法

吴叶兰 于宛莹 秦 晴 廉小亲 于重重 吴静珠

(北京工商大学计算机与人工智能学院, 北京 100048)

摘要: 针对柑橘病虫害领域文本数据中存在重叠三元组、嵌套实体和复杂实体抽取困难的问题, 提出一种基于 DPNA - CASREL (Dual-pointer network annotation-cascade binary tagging framework for relational triple extraction) 的柑橘病虫害实体关系联合抽取方法。通过结合预训练模型 RoBERTa - wwm - ext (Robustly optimized BERT pre-training approach with whole word masking and extended training data) 与双向长短期记忆网络 (Bi-directional long short-term memory, BiLSTM) 构建编码器获取文本的多维向量编码, 并根据柑橘病虫害语料特点设计双重指针网络标注的解码网络, 在头实体解码中引入多级指针网络标注方法, 在尾实体解码网络中采用复杂实体标注策略以增强模型对复杂实体的抽取性能, 实现对实体关系三元组的同步抽取, 解决三元组重叠、嵌套实体等问题。在自建柑橘病虫害数据集上的实验结果表明, DPNA - CASREL 模型的精确率、召回率和 F1 值分别为 82.12%、81.97%、82.05%, 优于其他模型, 对嵌套、复杂实体抽取的 F1 值比 CASREL 分别提升 8.16、6.58 个百分点, 有效解决了实体嵌套和实体边界不清晰问题。本文方法可为柑橘病虫害知识图谱构建提供基础。

关键词: 柑橘病虫害; 实体关系联合抽取; 双重指针网络标注; 嵌套实体; 复杂实体

中图分类号: TP391

文献标识码: A

文章编号: 1000-1298(2026)05-0398-09

OSID:



Joint Entity-relation Extraction Method for Citrus Diseases and Pests Based on DPNA - CASREL

WU Yelan YU Wanying QIN Qing LIAN Xiaojin YU Chongchong WU Jingzhu

(School of Computer and Artificial Intelligence, Beijing Technology and Business University, Beijing 100048, China)

Abstract: Aiming at the problems of overlapping triples, the difficulty in extracting nested entities and complex entities in the text data within the field of citrus diseases and pests, a joint extraction method for citrus diseases and pests entity relationships based on dual-pointer network annotation-cascade binary tagging framework for relational triple extraction (DPNA - CASREL) was proposed. By combining the pre-training model robustly optimized BERT pre-training approach with whole word masking and extended training data (RoBERTa - wwm - ext) with the bi-directional long short-term memory (BiLSTM) to construct an encoder, multi-dimensional vector encodings of the text were obtained. According to the semantic characteristics of citrus diseases and pests, a decoding network with dual-pointer network annotation was designed. The multi-level-pointer-network annotation method was introduced in decoding the head entity, and a complex entity labeling strategy was adopted in the decoding network of the tail entity to enhance the model's extraction performance for complex entities. By adopting a complex entity labeling strategy in the tail entity decoding network, the synchronous extraction of entity relationship triples was realized, and the problems of overlapping triples and nested entities were solved. Experimental results on a self-built citrus diseases and pests dataset showed that the precision, recall, and F1-score of the DPNA - CASREL model reached 82.12%, 81.97%, and 82.05%, respectively, which was superior to those of other models. Compared with CASREL, the F1-score of the nested and complex entity extraction were improved by 8.16 percentage points and 6.58 percentage points, respectively. This method can effectively solve the problems of entity nesting and unclear entity boundaries. It can provide a basis for citrus diseases and pests knowledge-graph construction and other downstream tasks.

Key words: citrus diseases and pests; joint entity-relation extraction; dual-pointer network annotation; nested entities; complex entities

收稿日期: 2024-10-31 修回日期: 2024-12-20

基金项目: 国家重点研发计划项目(2023YFD2101001)

作者简介: 吴叶兰(1970—), 女, 副教授, 主要从事智能信息处理、高光谱技术和机器人技术研究, E-mail: wuyel@th.btbu.edu.cn

0 引言

我国是柑橘生产、消费大国^[1],柑橘病虫害的防治对柑橘产业至关重要。现有的柑橘病虫害防治知识分散无序^[2],难以得到有效利用^[3],而知识图谱能够实现针对不同类型海量数据高效整合和复杂信息提取^[4]。知识图谱的构建关键在于知识抽取^[5],如何有效抽取实体关系对于柑橘病虫害防治的智能化发展具有重要作用。

基于深度学习的实体关系抽取方法通常分为流水线法(Pipeline)和联合学习方法(Joint learning)两类^[6]。流水线法将实体识别和关系抽取当成2个独立任务,先识别句子中的所有实体,再对实体间存在的关系进行分类。研究者通过参考柑橘病虫害领域农业叙词表,采用开源抽取系统^[7]和双向长短期记忆网络^[8]进行实体抽取和关系抽取。例如,LIU等^[9]通过结合预训练模型、双向长短期记忆网络和条件随机场,从非结构化数据中提取柑橘病虫害实体。但流水线方法容易出现错误传播^[10],实体识别的错误会在后续关系抽取中累积,影响整体知识抽取的准确性;同时,由于2个任务独立进行,难以进行全局优化调整,知识抽取不完善。为了解决这些问题,研究者提出实体关系联合抽取模型,包括基于特征工程的方法^[11-12]和基于神经网络的方法^[13-16],这些方法充分利用2个任务间的内在联系和相互依赖关系,在一个模型或框架中同时进行实体识别和关系抽取,避免了流水线法中错误传播和缺乏全局优化等问题,提高知识抽取的效率和准确性。基于特征工程的联合模型高度依赖人工提取特征规则,成本高且效率低,近年来,基于神经网络的联合模型因可自动学习特征而大幅减少人工参与,应用更为广泛。

ZHENG等^[17]将实体和关系的联合抽取转换为序列标注任务,采用双向长短期记忆网络同时进行实体识别和关系抽取,模型加强了2个子任务的交互性,但无法抽取实体间的多重关系。DAI等^[18]提出了一种基于注意力机制的统一标注策略联合抽取模型,根据关系类型对三元组进行验证,可以解决关系重叠问题。李书琴等^[19]构建了一种基于词嵌入的玉米育种实体关系联合抽取方法,融合字符和词汇特征以增强模型语义能力,设计统一标注方法和实体关系三元组匹配算法,提升了玉米育种知识抽取性能。以上基于序列标注的方法是对文本序列中的每个字符进行标注,不仅需要设计复杂的解码过程,当样本类别分布不均衡时还会影响模型解码效果。

为解决实体重叠问题,WEI等^[20]提出了参数共享的联合实体关系抽取方法,也称层叠指针网络框架,其基本原理是首先在一个句子中识别所有可能的头实体,将关系视为从头实体到尾实体的映射函数,再基于头实体同时识别关系和尾实体,实现三元组的联合抽取。胡滨等^[21]提出了一种基于双重指针网络和条件标准化网络层的联合抽取模型,加强模型对关系重叠问题的处理能力,实现家禽疾病诊疗文本中复杂实体关系三元组的快速抽取。周俊等^[22]根据水稻施肥数据特点提出了改进的层叠指针网络联合抽取模型,提高水稻施肥信息抽取效果。以上方法虽然实现了实体间多重关系的自动抽取,但是文本信息没有充分利用,且未能完全涵盖所有的嵌套实体,易造成实体间的关系缺失。现有柑橘病虫害领域的研究多集中于流水线式抽取研究,而联合抽取研究和相关标注数据集较少,且未对实体关系类型做出统一定义,缺乏对柑橘病虫害知识的完整性描述。同时,柑橘病虫害领域文本包含大量的重叠三元组和嵌套实体,部分实体存在特殊性,边界特征不明显,模型识别此类复杂实体时存在一定的难度。

综上,考虑到柑橘病虫害语料中存在重叠三元组、嵌套实体和复杂实体难以抽取等问题,提出一种基于双重指针网络标注的柑橘病虫害实体关系联合抽取模型(Dual-pointer network annotation-cascade binary tagging framework for relational triple extraction, DPNA-CASREL),以三元组(头实体、关系、尾实体)的形式呈现。将预训练模型(Robustly optimized BERT pre-training approach with whole word masking and extended training data, RoBERTa-wwm-ext)与双向长短期记忆网络(Bi-directional long short-term memory, BiLSTM)相结合,获取柑橘病虫害文本的多维共享编码表示;设计融合复杂实体标注与多级指针网络标注的双重指针网络标注方法,实现实体、实体类型和关系类型的同步抽取,从而获得高质量的柑橘病虫害知识,为构建柑橘病虫害知识图谱,实现柑橘病虫害知识服务系统提供坚实支撑。

1 语料特征分析

柑橘病虫害语料主要包括病虫害名称、为害症状、发生规律等信息,分析语料表达形式可知,有如下4类特征:

(1)三元组重叠。柑橘病虫害语料中,同一实体常出现在不同关系中。例如,语料“柑橘黄龙病主要为害柑橘叶片,多发生在广东等地区。”按照

柑橘病虫害知识图谱结构特征,“柑橘黄龙病”是重叠实体,存在于(柑橘黄龙病,部位,叶片)和(柑橘黄龙病,地区,广东)2个三元组中。

(2)实体嵌套。实体嵌套是指一个实体中隐藏另一实体。例如,语料“柑橘黄龙病为害叶片的典型症状是初期病树出现叶片黄化的新梢。”同时出现了为害部位实体“叶片”和为害症状实体“叶片黄化”。其中“叶片”为嵌套实体。

(3)实体中英文数字混合,实体复杂度高。柑橘病虫害语料中包括由数字、字母、汉字组成的复杂实体,例如,形态特征实体“成虫体长0.9~1.2 mm”包含了中文、数字、符号和英文字符,边界不清晰,导致实体抽取困难。

(4)实体长度较长,辨析难度大。例如,语料“温度适宜湿度大易发病,且柑橘脂点黄斑病的春梢发病比夏梢严重。”发生规律实体“春梢发病比夏梢严重”有9个中文字符,实体长,难以辨析。

现有研究尚未公开发表柑橘病虫害领域的实体关系抽取数据集,不同资料从结构上重点描述的柑橘病虫害类别不尽相同,内容上的描述各有侧重,文本信息相互区别又相互补充。在实体类型的定义中,从各文献资料的标注规范可知,“虫害”、“病害”、“为害部位”、“为害症状”等病虫害实体类型必不可少;在以防治资料为主的文本中,实体的定义通常会“防治方法”、“防治药剂”等类型再进行细分。

在关系类型的定义中,由于现有语料多聚焦于单一的实体定义,未必都对三元组的头尾实体进行区分,导致关系类型划分不全面,进而弱化了数据原有的语义信息。因此,本文针对柑橘病虫害语料特点,设计13种实体类型,包括柑橘病害、柑橘虫害、柑橘病虫害别称、病害病原、为害地区、为害症状、为害部位、发生规律、柑橘虫害生活习性、柑橘虫害形态特征、生物防治物种、农业防治措施、药剂防治措施。通过对实体间的关系梳理,得到柑橘病虫害实体之间的11种关系类别,如表1所示。

2 实体关系联合抽取模型

CASREL模型采用层次化的二级标注框架,有效地解决了重叠三元组问题。核心思想是先从BERT预训练模型编码的文本向量中检测出头实体,遍历选出的所有头实体,融合头实体特征及共享文本编码,从中解码对应尾实体与关系。但将CASREL模型应用于柑橘病虫害的实体关系抽取时,识别边界模糊的复杂实体效果欠佳,且现有头实体抽取方法无法识别嵌套实体和实体类型等问题。

表1 关系类别及示例

Tab.1 Categories and examples of relation

关系类别	示例
别称	柑橘黄龙病→柑橘黄梢病
病原	柑橘疮痂病→柑橘痂圆孢
地区	玉带凤蝶→广东
症状	柑橘黄龙病→叶片黄化
部位	柑橘溃疡病→果实
规律	柑橘白粉病→高温多雨
生活习性	茶黄蓟马→避光趋湿
形态特征	红蜡蚧→触角淡黄色
生物防治	褐天牛→长尾啮小蜂
农业防治	橘实瘦蚊→果实套袋
药剂防治	褐天牛→乐果乳油

因此,本文提出一种基于双重指针网络标注的DPNA-CASREL联合抽取模型,将RoBERTa-wwm-ext预训练模型引入文本编码,增强对柑橘病虫害领域知识文本的表征能力,得到具有丰富语义的序列向量;采用BiLSTM编码双向处理序列向量,捕获全局上下文语义特征,提升长实体识别性能;设计多级指针网络标注方法,同步解码头实体和实体类型;在尾实体解码网络中采用复杂实体标注策略,通过融合头实体特征、共享文本编码和复杂实体信息,增强模型对中英文、数字和特殊符号组成的模糊边界实体信息的捕捉能力,解码尾实体和关系;最终实现实体关系三元组的同步抽取。DPNA-CASREL模型的整体结构如图1所示。

2.1 RoBERTa-wwm-ext的文本编码

RoBERTa-wwm-ext是在BERT的基础上,针对中文语料进行优化得到的预训练语言模型,其特点是采用全词掩码(Whole word masking, WWM)机制,充分考虑中文语料特征,以词为单位进行动态Mask操作,通过对语料潜在结构和规律的学习,有效获取文本语义信息。由于柑橘病虫害文本包含较多专业术语和领域特定表达,需要模型具备较高的语言理解能力,而RoBERTa-wwm-ext在预训练阶段,通过处理大规模的中文文本可以较好地理解柑橘病虫害专业术语的含义。模型结构如图2所示,给定柑橘病虫害文本序列 $x = \{x_1, x_2, \dots, x_n\}$,将经过embedding的词嵌入、句嵌入、位置嵌入叠加后得到的序列编码表示作为模型的输入,经过多个深层双向Transformer编码器,充分挖掘文本中的语义信息,获得文本的多维向量 $x' = (x'_1, x'_2, \dots, x'_n)$ 。

通过分析柑橘病虫害语料特征发现,在长句子中常会出现多类实体,且具有实体长度不一、实体间跨度较大的特点,因此,采用双向长短期记忆网络BiLSTM,通过双向结构同时捕捉文本的前向和后向

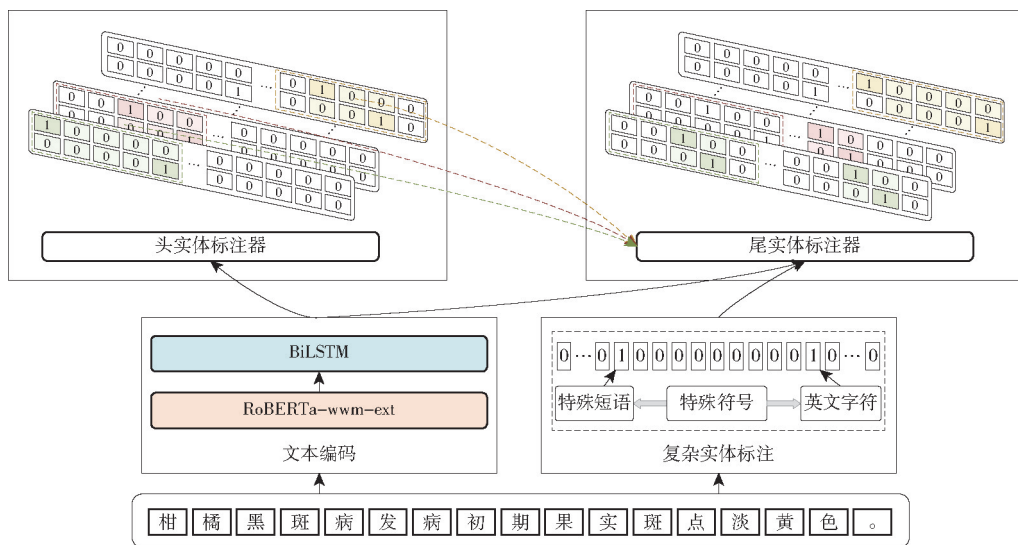


图1 DPNA-CASREL模型整体结构图

Fig.1 DPNA-CASREL model structure

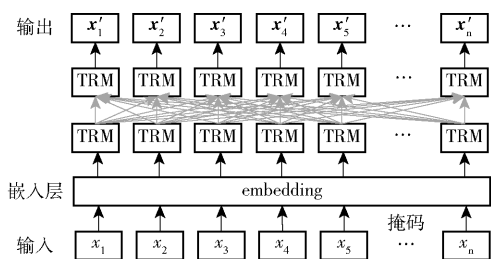


图2 RoBERTa-wwm-ext模型结构图

Fig.2 RoBERTa-wwm-ext model structure

上下文信息,有效地处理文本中的时序依赖和多层次语义,以适应柑橘病虫害不定长序列数据,缓解因长距离依赖带来的信息丢失问题。将 RoBERTa-wwm-ext 输出的编码向量 x' 作为输入, BiLSTM 从前后方向捕获柑橘病虫害文本中实体词的上下文语义特征和长距离依赖关系,得到编码向量 $h = \{h_1, h_2, \dots, h_n\}$, 提升对柑橘病虫害文本中实体词的抽取性能。

2.2 基于多级指针网络标注的头实体解码

为解决柑橘病虫害语料中实体嵌套和实体类型缺失等问题,本文设计了一种多级指针网络标注方法来解码头实体及实体类型。将编码得到的文本向量 h 作为头实体标注器的输入,对每种实体类型都分别建立 2 个指针,用以标注头实体在文本序列中的位置,通过 Sigmoid 函数将每个字符向量映射为其作为实体首位置 (a)、尾位置 (z) 的概率,字符在每个实体类型 e 中的概率计算过程为

$$p_i^{a_s} = \sigma(W_a^e h_i + b_a^e) \quad (1)$$

$$p_i^{z_s} = \sigma(W_z^e h_i + b_z^e) \quad (2)$$

式中 W^e, b^e ——头实体训练参数矩阵、偏置

$p_i^{a_s}, p_i^{z_s}$ ——输入序列中第 i 个标识为头实体的开始和结束位置的概率

h_i ——第 i 个字符的编码向量

σ ——Sigmoid 激活函数

分别检查每个字符对应的概率,如果概率值超过阈值,表示该字符为实体边界,标为 1,否则为 0。当识别到同一文本序列中有多个实体边界时,按照就近匹配原则,根据标记为 1 的实体开始位置,查找距离其最近的标记为 1 的实体结束位置,图 3 为标注示意图。

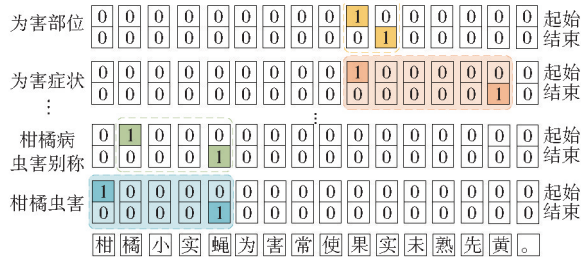


图3 多级指针网络标注

Fig.3 Multilevel pointer network annotation

2.3 融合复杂实体标注的尾实体解码

柑橘病虫害语料中包含大量由数字、英文和特殊字符组成的复杂实体,且该类实体大都出现在尾实体中,本文设计一种复杂实体标注方法,将得到的复杂实体位置信息融入尾实体标注器,提高复杂实体的抽取质量。

首先,利用分词工具将每条语料转换为词元,通过分析柑橘病虫害语料特征,筛选与形态特征、发生规律相关的词元,建立复杂实体语料库。采用正则法对句中出现的数字、英文和特殊字符进行匹配,以匹配到的符号作为中心,扫描复杂实体语料库,将向前扫描得到的词元作为起始标注位,向后扫描得到的英文字符作为结束标注位,将起始和结束标注位字符的位置标记为 1,其余位置标记为 0,实现从文

本序列到包含复杂实体位置信息标注向量的转换,图4给出了标注示例。

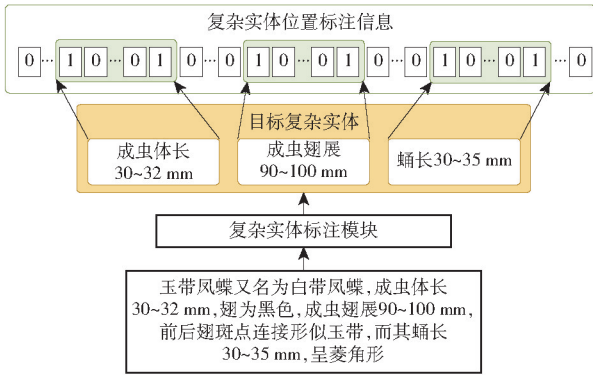


图4 复杂实体标注

Fig.4 Complex entities annotation

具体而言,设 $x = \{x_1, x_2, \dots, x_n\}$ 为文本序列,其中 $x_j (j=1, 2, \dots, n)$ 为文本中的元素。定义映射函数 f , 使得 $f: x \rightarrow C$, 其中 $C = \{C_1, C_2, \dots, C_m\}$ 为词元序列空间,且 $C_i, C_j \in C$, 满足

$$f(x) = C \quad (3)$$

$$q(x, C_i) = m \quad (4)$$

$$q(x, C_j) = n \quad (5)$$

$$u = g(x, a, b) \quad (6)$$

其中 $u = \{u_1, u_2, \dots, u_n\}$
式中 q ——映射函数,得到字符在文本序列中的位置

m ——中文词元位置,即起始标注位

n ——英文词元位置,即结束标注位

g ——映射函数,将文本序列与位置信息映射为包含复杂实体位置的向量,并对齐文本向量维度

u ——文本序列 x 对应的复杂实体标注向量

尾实体标注器的输入由编码器生成的文本编码向量 h 、复杂实体位置标注向量 u 和头实体解码器生成的头实体特征 v 组成,对每一种关系建立 2 个指针,依次预测每个字符作为尾实体首位置 (a)、尾位置 (z) 的概率,若概率超过阈值,则字符被判定为尾实体边界,将字符所在位置标注为 1,否则标注为 0。字符在每个关系类型 r 下的概率计算过程为

$$p_i^{a_o} = \sigma(W_a^r(h_i + u_i + v_s^k) + b_a^r) \quad (7)$$

$$p_i^{z_o} = \sigma(W_z^r(h_i + u_i + v_s^k) + b_z^r) \quad (8)$$

式中 W^r, b^r ——尾实体训练参数矩阵、偏置

u_i ——第 i 个字符的复杂实体位置标注向量

v_s^k ——第 k 个头实体的编码向量

$p_i^{a_o}, p_i^{z_o}$ ——输入序列中的第 i 个标识为尾实体的开始和结束位置的概率

2.4 损失函数

本文模型通过 2 个解码器分别实现头、尾实体

的抽取,对 2 个模块进行联合学习,因此本文模型的整体损失为头实体抽取任务与尾实体任务的损失之和,即

$$L_s = -\frac{1}{N} \sum_{j=1}^N \left[\sum_{s \in T_j} p_{\theta_e}(s|h) + \sum_{r \in R} p_{\phi_r}(o|s, h, u) \right] \quad (9)$$

其中

$$p_{\theta_e}(s|h) = \sum_{t \in (a_s, z_s)} \sum_{i=1}^L [y_i^t \lg p_i^t + (1 - y_i^t) \lg(1 - p_i^t)] \quad (10)$$

$$p_{\phi_r}(o|s, h, u) =$$

$$\sum_{t \in (a_o, z_o)} \sum_{i=1}^L [y_i^t \lg(p_i^t) + (1 - y_i^t) \lg(1 - p_i^t)] \quad (11)$$

$$\theta_e = \{W_a^e, W_z^e, b_a^e, b_z^e\} \quad (12)$$

$$\phi_r = \{W_a^r, W_z^r, b_a^r, b_z^r\} \quad (13)$$

式中 L_s ——整体损失

N ——输入文本的数量

T_j ——句中潜在三元组

$p_{\theta_e}(s|h)$ ——头实体抽取任务的目标函数

$p_{\phi_r}(o|s, h, u)$ ——尾实体抽取任务的目标函数

L ——文本长度

y_i^t ——标签类别,取 0 或 1

p_i^t ——第 i 个字符的实体首尾位置的概率

θ_e, ϕ_r ——参数

3 实验结果与分析

3.1 数据来源与标注

目前,有关柑橘病虫害结构化数据的资源有限,且数据可信度参差不齐。因此,本文有关柑橘病虫害信息的数据集主要从《柑橘病虫害诊断与防治原色图谱》^[23]、《新编柑橘病虫害诊断与防治图鉴》^[24]等书籍上通过摘录获取。从中国惠农网 (<https://www.cnhn.com>)、农业病虫害多媒体数据库 (<http://zhibao.yuanlin.com>) 等农技网站获取知识文本数据作为补充,建立柑橘病虫害知识文本数据集。

对获取到的柑橘病虫害文本数据进行清洗,共整理得到 5 074 条文本语料。根据定义的 13 种实体类型和 11 种关系类型构建数据集,对原始语料以句子为单位进行标注,数据标注格式为: {句子, 关系类型, 头实体类型, 尾实体类型、头实体、尾实体}, 如图 5 所示。所标注的不重复实体总量 6 312 个、关系总量 14 313 个,不同类型的数量分布如图 6 所示。标注完成的数据集按照比例 8:1:1 划分为训练集、验证集和测试集。

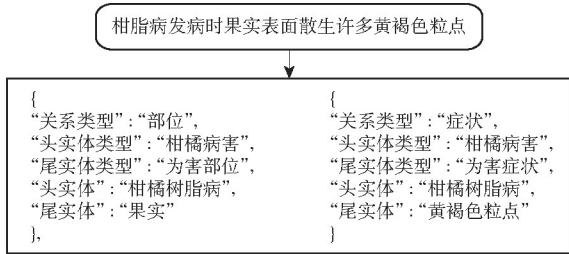


图5 语料标注示例

Fig. 5 Annotated sample

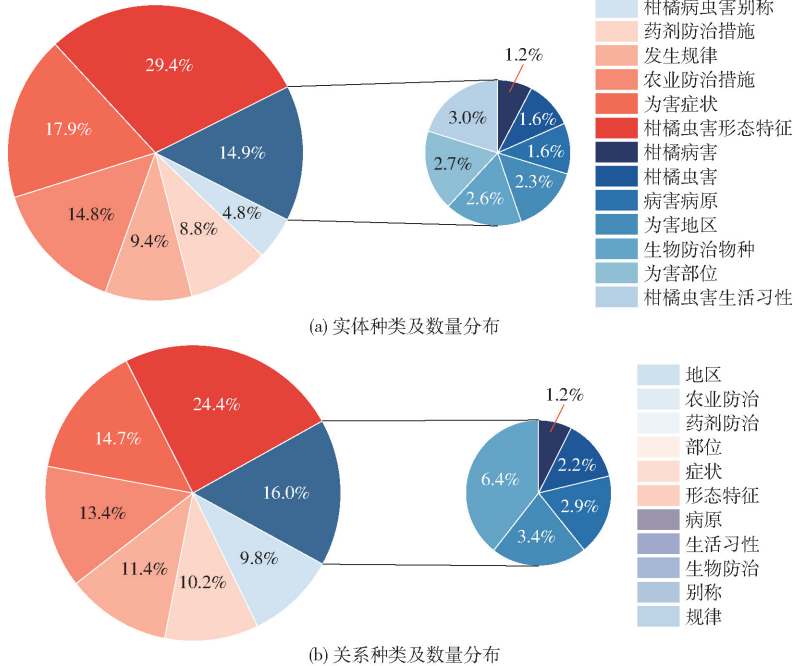


图6 实体和关系种类及数量分布

Fig. 6 Type and quantity distribution of entities and relations

表2 参数设置

Tab.2 Parameters setting

参数	数值
批尺寸	8
最大文本长度	512
迭代次数	150
学习率	0.00001
激活阈值	0.5
随机丢弃率	0.1

能评估。

3.3 模型评估指标

实验采用 Pytorch 深度学习框架,以构建的柑橘病虫害文本数据集为基础,对本文提出的网络模型进行训练及测试,采用精确率(Precision)、召回率(Recall)和 F1 值(F1 score)作为评估指标。精确率用于评估模型预测正样本的能力,召回率可以评估模型识别所有正样本的能力,F1 值是精确率和召回率的加权平均,取值在 0~1 之间,值越高表明模型越稳健。每次实验过程随机重复 5 次,将 5 次实验结果的平均值作为最终结果。

3.2 实验环境与参数设置

本文实验环境为 Intel(R) Core(TM) i7 - 6800K CPU,频率 3.40 GHz,内存为 16.0 GB,GPU 为 NVIDIA GeForce GTX 1080Ti,开发环境为 Python 3.6,CUDA 10.0,Windows 10 64 位操作系统,表 2 给出了实体关系联合抽取模型网络结构参数,并利用验证集对模型超参数进行优化,在具有最佳实验效果的模型参数环境下,基于测试集对模型进行性

3.4 结果分析

3.4.1 不同模型性能分析

为分析模型性能,将本文模型(DPNA - CASREL)和其他基线模型^[25-29]在构建的柑橘病虫害数据集上进行对比,实验结果如表 3 所示。

表3 对比实验结果

Tab.3 Results of comparative experiment %

模型	精确率	召回率	F1 值
DPNA - CASREL	82.12	81.97	82.05
SpERT	77.12	74.57	75.82
TPlinker	76.63	74.26	75.43
OneRel	74.02	64.06	67.09
MultiHead	70.15	64.82	67.38
CasRel	76.67	68.02	72.09
SFRARE	73.27	75.41	74.33

分析结果可知,DPNA - CASREL 的精确率、召回率和 F1 值均优于其他模型,这是因为本文模型充分利用了双向文本特征,同时针对复杂实体设计了标注策略,采用双重指针网络标注方法抽取实体,对比采用单步抽取策略的 CASREL 模型,精确率、召回

率和 F1 值分别提升了 5.45、13.95、9.96 个百分点。SpERT 对每一种实体对都进行关系抽取,导致大量的信息冗余,引入不必要的噪声,精确率、召回率和 F1 值比本文模型分别低了 5.00、7.40、6.23 个百分点。TPlinker 构建的关系矩阵过于稀疏,模型学习效果较差。OneRel 本质上同样忽略实体识别与关系抽取任务间的联系,存在负样本过多的问题。MultiHead 模型为先识别实体,再抽取实体间关系的两阶段过程,导致误差累积,比本文模型精确率、召回率和 F1 值分别低 11.97、17.15、14.67 个百分点。SFRARE 模型没有充分提取嵌套头实体,进而造成了尾实体和对应关系的缺失。综上,本文构建的 DPNA - CASREL 模型在柑橘病虫害实体关系抽取任务中达到了最优效果。

3.4.2 消融实验

为了验证本文模型(DPNA - CASREL)每部分对整体模型的影响,对头实体多级指针网络标注模块 multilevel、复杂实体标注模块 complex 和 BiLSTM 模块进行消融实验,实验结果如表 4 所示。

表 4 消融实验结果

Tab.4 Ablation experiment result %

模型	精确率	召回率	F1 值
DPNA - CASREL	82.12	81.97	82.05
-multilevel	77.47	70.62	73.89
-complex	79.31	71.98	75.47
-BiLSTM	75.81	77.83	76.81

注:“-”表示移除本文模型中的某个组件。

柑橘病虫害数据集中存在大量嵌套实体,一个字符可能会对应多个标签,多级指针网络标注能够针对每个字符按照不同的实体类型分别预测实体边界位置,避免嵌套实体的边界出现交叉或混淆。从表 4 可以看出,移除头实体多级指针网络标注 multilevel 模块,模型的精确率、召回率、F1 值分别降低 4.65、11.35、8.16 个百分点,验证了本文方法在引入多级指针网络标注模块后,通过结合实体类型能够有效提高嵌套实体识别性能,增强后续尾实体抽取任务处理能力。

表 4 还给出了复杂实体标注模块 complex 对模型的影响,相比移除 complex 模块的结果,DPNA - CASREL 模型的精确率提升 2.31 个百分点,召回率和 F1 值分别提升 9.99、6.58 个百分点,多预测了 51 个三元组,预测正确的三元组增加了 29 个,例如,输入文本序列“玉带凤蝶又名白蝶,成虫体长 30 ~ 32 mm,翅为黑色,成虫翅展 90 ~ 100 mm,前后翅斑点连接形似玉带,而其蛹长 30 ~ 35 mm,呈菱形。”双重指针网络标注方法可正确识别“成虫体长

30 ~ 32 mm”、“成虫翅展 90 ~ 100 mm”、“蛹长 30 ~ 35 mm”3 个复杂实体,表明将复杂实体位置向量融入文本特征可以帮助模型更加充分地学习和理解复杂实体数据的内在结构和规律,增强复杂实体在文本向量中的表达能力,提高模型抽取复杂实体的性能。

与移除 BiLSTM 模块的结果相比,DPNA - CASREL 模型的精确率、召回率和 F1 值分别提升 6.31、4.14、5.24 个百分点,这是因为柑橘病虫害文本中头实体和尾实体可能间隔较远,存在长距离依赖问题,而 BiLSTM 通过双向信息处理,能够缓解长距离依赖带来的信息丢失问题,可以有效解析和利用来自编码模块中的词语表征。因此,证明了模型添加头实体多级指针网络标注、复杂实体标注策略和 BiLSTM 模块有利于模型性能的提升。

3.4.3 不同预训练模型对比

预训练模型能够提取语料中的文本特征,将 BERT - base - chinese、BERT - wwm、BERT - wwm - ext、RoBERTa - wwm - ext 模型分别作为柑橘病虫害文本序列的编码模型进行对比。BERT - base 是 Google 提供的基础中文预训练模型,BERT - wwm - ext 和 RoBERTa - wwm - ext 是 Google 基于 WWM 发布的升级版本,语料为 EXT 数据,包括中文维基百科、其他百科新闻、问答等。最终训练结果如表 5 所示,可以看出 RoBERTa - wwm - ext 模型相对于 BERT - base - chinese 模型、BERT - wwm 模型、BERT - wwm - ext 模型的 F1 值分别提升 8.99、7.63、1.9 个百分点。由此可见,RoBERTa - wwm - ext 编码方式更符合柑橘病虫害文本的真实结构,在专业领域中词的完整语义比单字的语义更为重要,全词掩码方式可以让模型更好地理解 and 生成词级别的语义信息,捕获柑橘病虫害文本长词或专有名词的语义。因此,基于 WWM 的 RoBERTa - wwm - ext 编码对中文任务的处理效果最优。

表 5 不同预训练模型对比结果

Tab.5 Comparison results of different pre-training

模型	精确率	召回率	F1 值
BERT - base - chinese	72.47	74.76	73.06
BERT - wwm	75.79	73.12	74.42
BERT - wwm - ext	80.45	79.85	80.15
RoBERTa - wwm - ext	82.12	81.97	82.05

3.4.4 模型参数影响

为了验证不同参数取值对模型结果的影响,本文针对批尺寸(batch size)和学习率(l_r)参数的选择进行了实验,通过精确率、召回率和 F1 值来分析对模型的影响,以此来选出最优的参数组合。

表6为本文模型在不同批尺寸下的实验结果。可以看出,当批尺寸为8时,模型F1值达到最高。因此设置本文模型的批尺寸为8。

表6 批尺寸对模型的影响

Tab.6 Impact of batch size on model %

批尺寸	精确率	召回率	F1值
4	77.55	78.86	78.20
8	82.12	81.97	82.05
16	78.27	76.38	77.31
32	81.03	77.10	79.02

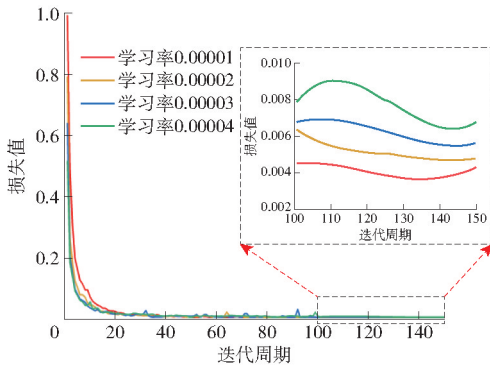


图7为本文模型在批尺寸为8、学习率分别为0.00001、0.00002、0.00003、0.00004时,模型训练过程的损失值和F1值的变化曲线。由损失值曲线可知,模型前20轮次迭代学习率为0.00004下降最快,最终收敛于0.007左右,学习率为0.00001模型收敛于0.004左右。由F1值曲线可知,在前50次迭代中不同学习率的F1值波动较大,模型经过70次迭代后趋于稳定,学习率为0.00001的F1值保持最高。因此设置本文模型学习率为0.00001。

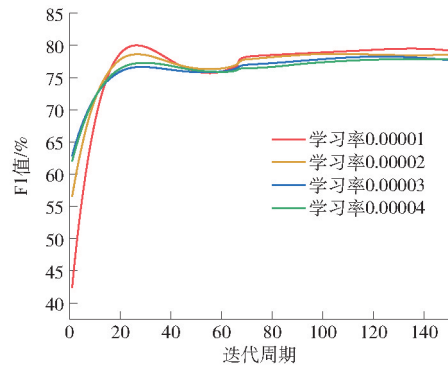


图7 学习率对模型的影响

Fig.7 Impact of learning rate on model

3.4.5 三元组抽取结果

为了验证模型在各类三元组上的抽取能力,表7列出了所提方法在柑橘病虫害数据集中对各类关系三元组抽取结果。由表7可知,本文方法对别称、病原、地区、部位、生活习性、生物防治、药剂防治7类三元组抽取的F1值均在87%以上,对症状和农

业防治措施的F1值仅在70%左右。原因可能是数据集中这2类三元组的实体特征不明显,实体跨度较大。特别是农业防治类三元组的召回率明显高于精确率,说明模型预测出的三元组中负样本较多,这是由于中文语言描述多样,近义词较多,导致模型在判断中容易混淆,造成F1值较低。

表7 三元组抽取结果

Tab.7 Triplet extraction result %

类型	别称	病原	地区	症状	部位	规律	生活习性	形态特征	生物防治	农业防治	药剂防治
精确率	93.33	94.12	97.92	78.11	91.62	81.48	93.93	80.65	88.23	64.45	91.08
召回率	84.84	84.21	88.68	75.00	83.24	85.71	81.57	82.64	88.23	76.42	91.67
F1值	88.89	88.89	93.07	76.52	87.23	83.54	87.32	81.63	88.23	69.93	91.37

4 结束语

针对柑橘病虫害文本数据中存在三元组重叠、实体嵌套和复杂实体抽取困难的问题,提出实体关系联合抽取模型DPNA-CASREL,自动抽取实体关系三元组,为其他领域实体关系抽取任务提供思路。引入RoBERTa-wwm-ext预训练模型编码文本序列,与其他预训练模型对比,RoBERTa-wwm-ext可以增

强柑橘病虫害文本语义表征能力,通过BiLSTM神经网络实现长序列记忆功能。设计复杂实体标注与多级指针网络标注相结合的双重指针网络标注方法,增强模型对嵌套实体识别能力与边界模糊的复杂实体信息的捕捉能力,通过与其他模型实验对比,本文模型的精确率达到82.12%,召回率达到81.97%,F1值达到82.05%,均优于基线模型,证明模型对柑橘病虫害领域的文本三元组抽取具备一定优势。

参 考 文 献

[1] 赵瑞雪,杨晨雪,郑建华,等. 农业智能知识服务研究现状及展望[J]. 智慧农业(中英文),2022,4(4):105-125. ZHAO Ruixue, YANG Chenxue, ZHENG Jianhua, et al. Agricultural intelligent knowledge service: overview and future perspectives[J]. Smart Agriculture,2022,4(4):105-125. (in Chinese)

[2] 赵春江. 农业知识智能服务技术综述[J]. 智慧农业(中英文),2023,5(2):126-148.

- ZHAO Chunjiang. Agricultural knowledge intelligent service technology: a review[J]. *Smart Agriculture*, 2023, 5(2): 126 - 148. (in Chinese)
- [3] 赵春江. 智慧农业的发展现状与未来展望[J]. *华南农业大学学报*, 2021, 42(6): 1 - 7.
ZHAO Chunjiang. Current situations and prospects of smart agriculture[J]. *Journal of South China Agricultural University*, 2021, 42(6): 1 - 7. (in Chinese)
- [4] 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)
- [5] 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)
- [6] 鄂海红, 张文静, 肖思琪, 等. 深度学习实体关系抽取研究综述[J]. *软件学报*, 2019, 30(6): 1793 - 1818.
E Haihong, ZHANG Wenjing, XIAO Siqi, et al. Survey of entity relationship extraction based on deep learning[J]. *Journal of Software*, 2019, 30(6): 1793 - 1818. (in Chinese)
- [7] 夏龙. 基于知识图谱的柑橘病虫害问答系统设计及实现[D]. 长沙: 湖南农业大学, 2022.
XIA Long. Design and implementation of citrus disease and pest answering system based on knowledge graph[D]. Changsha: Hunan Agricultural University, 2022. (in Chinese)
- [8] 谭迦瀚. 柑橘病虫害知识图谱构建研究[D]. 长沙: 湖南农业大学, 2022.
TAN Jiahao. Construction of knowledge map of citrus diseases and insect pests[D]. Changsha: Hunan Agricultural University, 2022. (in Chinese)
- [9] LIU Yafei, WEI Siqi, HUANG Haijun, et al. Naming entity recognition of citrus diseases and pests based on the BERT - BiLSTM - CRF model[J]. *Expert Systems With Applications*, 2023, 234: 121103.
- [10] MIWA M, BANSAL M. End-to-end relation extraction using LSTMs on sequences and tree structures[C]// *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2016: 11050 - 1116.
- [11] 张少伟, 王鑫, 陈子睿, 等. 有监督实体关系联合抽取方法研究综述[J]. *计算机科学与探索*, 2022, 16(4): 713 - 733.
ZHANG Shaowei, WANG Xin, CHEN Zirui, et al. Survey of supervised joint entity relation extraction methods[J]. *Journal of Frontier of Computer Science and Technology*, 2022, 16(4): 713 - 733. (in Chinese)
- [12] REN X, WU Z Q, HE W Q, et al. CoType: joint extraction of typed entities and relations with knowledge bases[C]// *Proceedings of the 26th International Conference on World Wide Web*. Perth: International World Wide Web Conferences Steering Committee, 2017: 1015 - 1024.
- [13] FU T, LI P, MA W. GraphRel: modeling text as relational graphs for joint entity and relation extraction[C]// *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.
- [14] ZENG X R, ZENG D J, HE S Z, et al. Extracting relational facts by an end-to-end neural model with copy mechanism[C]// *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, 2018: 506 - 514.
- [15] WANG Y, SUN C, WU Y, et al. UniRE: a unified label space for entity relation extraction[C]// *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, 2021: 220 - 231.
- [16] LI X Y, YIN F, SUN Z J, et al. Entity-relation extraction as multi-turn question answering[C]// *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019: 1340 - 1350.
- [17] ZHENG S, WANG F, BAO H, et al. Joint extraction of entities and relations based on a novel tagging scheme[C]// *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, 2017: 1227 - 1236.
- [18] DAI D, XIAO X Y, LYU Y J, et al. Joint extraction of entities and overlapping relations using position-attentive sequence labeling[C]// *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019: 6300 - 6308.
- [19] 李书琴, 庞文婷. 词嵌入 BERT - CRF 玉米育种实体关系联合抽取方法[J]. *农业机械学报*, 2023, 54(11): 286 - 294.
LI Shuqin, PANG Wenting. Joint extraction method of entity and relation in maize breeding based on BERT - CRF and word embedding[J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2023, 54(11): 286 - 294. (in Chinese)
- [20] WEI Z P, SU J L, WANG Y, et al. A novel cascade binary tagging framework for relational triple extraction[C]// *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020: 1476 - 1488.
- [21] 胡滨, 汤保虎, 姜海燕, 等. 家禽诊疗文本多实体关系联合抽取模型研究[J]. *农业机械学报*, 2021, 52(6): 268 - 276.
HU Bin, TANG Baohu, JIANG Haiyan, et al. Joint extraction model of multi-entity relations for poultry diagnosis and treatment text[J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2021, 52(6): 268 - 276. (in Chinese)
- [22] 周俊, 郑彭元, 袁立存, 等. 基于改进 CASREL 的水稻施肥知识图谱信息抽取研究[J]. *农业机械学报*, 2022, 53(11): 314 - 322.
ZHOU Jun, ZHENG Pengyuan, YUAN Licun, et al. Knowledge graph information extraction for rice fertilization based on improved CASREL[J]. *Transactions of the Chinese Society for Agricultural Machinery*, 2022, 53(11): 314 - 322. (in Chinese)
- [23] 夏声广. 柑橘病虫害诊断与原色图谱[M]. 北京: 中国农业出版社, 2021.
- [24] 蔡明段, 彭成绩. 新编柑橘病虫害诊断与防治图鉴[M]. 广东: 广东科技出版社, 2020.
- [25] EBERTS M, ULGES A. Span-based joint entity and relation extraction with transformer pre-training[C]// *Proceedings of the 24th European Conference on Artificial Intelligence*. Amsterdam: IOS Press, 2020: 2006 - 2013.
- [26] BEKOULIS G, DELEU J, DEMEESTER T, et al. Joint entity recognition and relation extraction as a multi-head selection problem[J]. *Expert Systems with Applications*, 2018, 114: 34 - 45.
- [27] WANG Y, YU B, ZHANG Y, et al. TPLinker: single-stage joint extraction of entities and relations through token pair linking[C]// *Proceedings of the 28th International Conference on Computational Linguistics*, 2020: 1572 - 1582.
- [28] SHANG Y M, HUANG H Y, MAO X L. OneRel: joint entity and relation extraction with one module in one step[C]// *Proceedings of the 36th AAAI Conference on Artificial Intelligence*. AAAI Press, 2022: 11285 - 11293.
- [29] LUO W Q, WANG J B, YAN X W, et al. Unveiling the railway traffic knowledge in tibet: an advanced model for relational triple extraction[J]. *Sustainability*, 2023, 15: 14942.