doi:10.6041/j.issn.1000-1298.2022.12.032

融合字词语义信息的猕猴桃种植领域命名实体识别研究

李书琴　张明美　刘　斌
(西北农林科技大学信息工程学院, 陕西杨凌 712100)

摘要: 针对猕猴桃种植领域命名实体识别任务中实体词复杂度较高,识别精确率较低的问题,提出一种融合字词语义信息的猕猴桃种植实体识别方法。以 BiGRU－CRF 为基本模型,融合词级别和字符级别的信息。在词级别上,通过引入词集信息,并使用多头自注意力(Multiple self-attention mechanisms,MHA)调整词集中不同词的权重;同时使用注意力机制忽略不可靠的词集,将注意力集中在重要的词集上,从而提高实体识别效果;在字符级别上,引入无监督的基于转换器的双向编码表征(Bidirectional encoder representations form transformers,BERT)预训练模型增强字的语义表示。在包含 12 477 条标注样本和 7 个类别实体的猕猴桃种植领域自制语料上进行了实验,结果表明,本文模型与 SoftLexicon 模型相比,F1 值提高 1.58 个百分点。此外,本文模型在公开数据集 ResumeNER 上与 Lattice－LSTM、WC－LSTM 等模型进行实验对比取得了最佳效果,F1 值达到 96.17%,表明本文模型具有一定的泛化能力。

关键词: 猕猴桃种植; 命名实体识别; 字词融合; 语义增强; 自注意力机制; 预训练语言模型
中图分类号: TP391.1　　**文献标识码**: A　　**文章编号**: 1000-1298(2022)12-0323-09　　**OSID**:

Kiwifruit Planting Entity Recognition Based on Character and Word Information Fusion

LI Shuqin　ZHANG Mingmei　LIU Bin
(*College of Information Engineering*, *Northwest A&F University*, *Yangling*, *Shaanxi* 712100, *China*)

Abstract: Aiming at the problem of high complexity of real words and low recognition accuracy in the named entity recognition task of kiwifruit planting field, a entity recognition method of kiwifruit planting integrating character and word information was proposed. Based on BiGRU－CRF model, word level and character level information were fused. At the word level, by introducing word set information and using multiple self-attention mechanisms (MHA) to adjust the weights of different words in the word set. At the same time, attention mechanism was used to ignore the unreliable word sets and focus on the important word sets to improve the entity recognition effect. At the character level, the unsupervised bidirectional encoder representations form transformers (BERT) pre-training model was introduced to enhance the semantic representation of words. Experiments were conducted on a homemade corpus in the kiwifruit cultivation domain containing 12 477 annotated samples and seven categories of entities, and the results showed that the F1 value of the model was improved by 1.58 percentage points compared with the SoftLexicon model. In addition, the experimental comparison of the model ResumeNER with Lattice－LSTM, WC－LSTM and other models in the open data set ResumeNER was carried out, and the best recognition effect was achieved. The F1 value reached 96.17%, indicating that the method proposed had certain generalization ability.

Key words: kiwifruit planting; named entity recognition; word fusion; semantic enhancement; self-attention mechanisim; pre-trained language model

0　引言

受病虫害的侵害和种植人员对种植技术掌握不全面的影响,我国猕猴桃果实品质整体水平不高[1]。基于知识图谱的猕猴桃种植领域问答系统利用知识图谱可以准确快速回答猕猴桃种植人员的

收稿日期: 2021－12－19　修回日期: 2022－01－24
基金项目: 国家重点研发计划项目(2020YFD1100601)、陕西省重点研发计划项目(2021NY－138)和中央高校基本科研业务专项资金项目(2452019064)
作者简介: 李书琴(1965—),女,教授,博士生导师,主要从事农业信息化与智能信息系统研究,E-mail: lsq_cie@nwsuaf.edu.cn

专业问题，而命名实体识别（Named entity recognition，NER）是知识图谱构建任务中重要且关键的步骤[2]，因此，如何准确快速识别出猕猴桃种植领域命名实体对于确保猕猴桃种植业健康发展具有重要作用。

早期基于统计机器学习的条件随机场（Conditional random field，CRF）方法将实体识别看作序列标注问题，充分利用了内部和上下文信息，在农业领域得到广泛应用[3-4]。但该方法过于依赖人工特征，特征的设计需要很多专家知识，特征选择的好坏更是直接影响到命名实体识别系统的性能[5]。近年来，基于深度学习的方法在 NER 任务中取得了显著效果。深度学习可以自动学习文本特征，从而摆脱对人工特征的依赖，其中，卷积神经网络、循环神经网络以及注意力机制等常用的深度学习方法与机器学习联合使用的方式已经被成功地应用到农业垂直领域的命名实体识别任务中[6-7]。

但以上方法在处理猕猴桃种植领域文本时，需要先进行中文分词（Chinese word split，CWS），CWS 的准确性直接影响到中文命名实体识别效果。且猕猴桃种植领域命名实体识别任务主要关注猕猴桃种植文本中的猕猴桃品种、病虫害、危害部位、药剂、种植技术等实体，由于猕猴桃种植领域文本中涉及的病虫害、药剂及种植技术等多种实体术语专业性较强，CWS 容易产生大量的未登录词（Out-of-vocabulary，OOV），从而影响模型识别效果。

MENG 等[8]在中文自然语言处理中通过大量的实验表明，“字”的表现总是优于“词”的表现。一些研究者[5,9-11]为了避免 CWS 错误，直接使用基于 word2vec 等词向量训练模型训练的字向量作为嵌入层。但以上词向量训练模型在单个字符上语义表征不充分导致模型识别性能欠佳。

BERT[12]等预训练语言模型采用双向的 Transformer 编码器对大规模语料进行训练，可以得到表征能力更强的字向量[13]。已有研究人员将预训练语言模型引入农林业领域命名实体识别任务中[14-17]。但实体识别任务与其它自然语言处理任务不同的是，大部分实体属于词，词中蕴含着丰富的实体信息，而字符向量却缺少该类信息。ZHANG 等[18]提出的 Lattice-LSTM 模型，将每个字符匹配到的单词通过注意力机制进行加权求和作为字符表示，但由于每个字符对应的词数目不同，无法分批处理，导致识别速度较慢，且由于模型结构复杂，无法迁移到其它网络结构中。针对该问题，LIU 等[19]提出了 4 种不同的策略将词进行固定数目的编码，使其可以分批处理从而适应各种网络结构。MA 等[20]提出了一个更为简单高效的 SoftLexicon 模型，利用 4 个词集来表示每个字符在词中的位置，同时采用词的频率作为权重对词集进行压缩，简化了序列建模结构，提高了模型计算效率。但词集内的词语义信息往往是相似的，上述研究忽略了不同词对于当前字符的重要程度，词集中包含的词信息没有得到充分利用。

基于以上问题和研究，本文提出一种融合字词语义信息的猕猴桃种植实体识别方法。首先采用多头自注意力（Multiple self-attention mechanisms，MHA）[21]来调整 SoftLexicon 词集中每个词语的权重，缓解静态词频作为权重无法学习到更为重要的词特征问题；然后采用注意力机制自动获取每个词集的重要程度，增强重要词集信息的同时抑制不重要词集信息；最后融合词集表示和 BERT 的字符表示作为命名实体识别任务的嵌入层。同时使用双向门控循环网络（Bi-directional gated recurrent unit，BiGRU）进一步提取字符之间的关系特征，最终使用 CRF 得到全局最优标签序列。

1 融合字词语义信息的实体识别模型

模型主要由 3 部分构成，嵌入层、BiGRU 编码层以及 CRF 层。嵌入层使用融合字词语义信息的表示，字符语义信息使用 BERT 预训练模型生成的字符表示，词语义信息使用注意力加权得到的词集向量表示。编码层采用 BiGRU 网络，最后通过 CRF 进行标签推理，获取全局最优标签序列。模型整体结构如图 1 所示。

1.1 融合字词语义信息的嵌入层

模型嵌入层融合了基于改进的 SoftLexicon 模型生成的词向量信息和采用 BERT 预训练语言模型生成的字符向量信息。

词向量由 4 个词集组成，对于输入文本序列 $\boldsymbol{S}=(c_1,c_2,\cdots,c_T)$，将序列中相邻的字符在词典中匹配词组，并按照每个字符 c_i 在词组中的不同位置，分别用标签为 $\boldsymbol{B}$、$\boldsymbol{M}$、$\boldsymbol{E}$、$\boldsymbol{S}$ 的 4 个集合来记录，集合 $\boldsymbol{B}(c_i)$ 表示字符 c_i 在开头且长度大于 1 的词集合，集合 $\boldsymbol{M}(c_i)$ 表示字符 c_i 在中间位置且长度大于 1 的词集合，集合 $\boldsymbol{E}(c_i)$ 表示字符 c_i 在结尾且长度大于 1 的词集合，集合 $\boldsymbol{S}(c_i)$ 表示单个字符 c_i，如果集合为空，则用“None”来填补。如图 2 所示，输入句子“猕猴桃根腐病危害软枣猕猴桃根部”，以字符 c_4 为例，因为该字符出现在“根腐病”的开头，“猕猴桃根腐病”的中间，“猕猴桃根”的结尾，所以 $\boldsymbol{B}(c_4)$ 为{“根腐病”}，$\boldsymbol{M}(c_4)$ 为{“猕猴桃根腐病”}，$\boldsymbol{E}(c_4)$ 为{“猕猴桃根”}，$\boldsymbol{S}(c_4)$ 则为{“根”}。

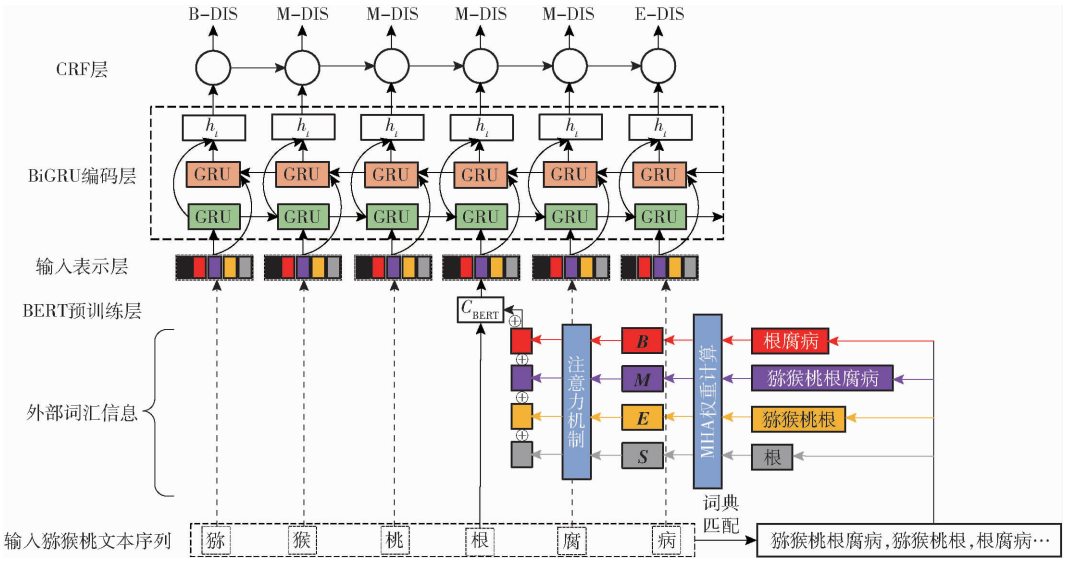

图 1　融合字词语义信息的猕猴桃种植实体识别模型

Fig. 1　Kiwifruit planting entity recognition model integrating character and word information fusion

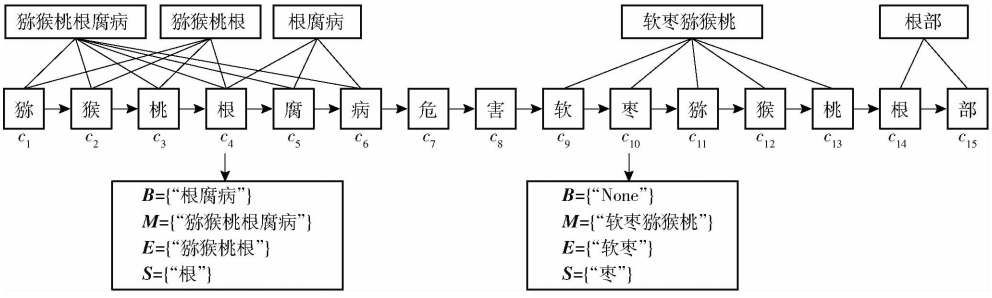

图 2　词组匹配分类

Fig. 2　Word matching classification

获得每个字符对应的 $\boldsymbol{B}$、$\boldsymbol{M}$、$\boldsymbol{E}$、$\boldsymbol{S}$ 标签词集后，需要对每个词集进行压缩得到 4 个标签的词向量。原始的 SoftLexicon 模型只使用词频 $z(w)$，即每个词 w 在词典中出现的次数作为权重进行压缩，同时对于所有词集中不满足最大长度 $l_{\max}$ 的词集用 0 进行填充，并记录该词的 $z(w)$ 为 1。词集向量 $\boldsymbol{B}$ 的具体计算方式（$\boldsymbol{M}$、$\boldsymbol{E}$、$\boldsymbol{S}$ 同理）为

$$\boldsymbol{v}_i(\boldsymbol{B}) = \frac{4}{Z}\sum_{w\in\boldsymbol{B}} z(w)\boldsymbol{e}^w(w) \tag{1}$$

其中

$$Z = \sum_{w\in\boldsymbol{B}\cup\boldsymbol{M}\cup\boldsymbol{E}\cup\boldsymbol{S}} z(w) \tag{2}$$

式中　$\boldsymbol{e}^w(w)$——词 w 对应的词向量

$\boldsymbol{v}_i$——词集向量

对于要识别的猕猴桃种植领域文本中的实体，仅使用词频作为权重时，容易出现准确率较低但召回率较高的情况，例如针对“果实软腐病”的“果”字，“果实软腐病”和“果实”均属于 $\boldsymbol{B}$ 集合，“果实软腐病”中字符“果”的正确标签为“B-DIS”，但由于“果实”在词典中出现的频率较高且没有使用注意力等方式计算权重，“果实软腐病”中的字符“果”被标记为“B-PART”，导致精确率较低但召回率较高。针对以上问题，本文采用 MHA 机制动态地调整每个词语权重，学习到更为重要的特征后再进行压缩，词集 $\boldsymbol{B}$ 的具体计算公式（$\boldsymbol{M}$、$\boldsymbol{E}$、$\boldsymbol{S}$ 同理）为

$$\boldsymbol{v}_i(\boldsymbol{B}) = \frac{4}{Z}\sum_{w\in\boldsymbol{B}} \mathrm{MHA}(z(w)\boldsymbol{e}^w(w)) \tag{3}$$

$$\mathrm{MHA}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = \mathrm{Concat}(head_1, head_2, \cdots, head_n)\boldsymbol{W}^o \tag{4}$$

$$head_j = \mathrm{Attention}(\boldsymbol{Q}_j,\boldsymbol{K}_j,\boldsymbol{V}_j) \tag{5}$$

$$\mathrm{Attention}(\boldsymbol{Q}_j,\boldsymbol{K}_j,\boldsymbol{V}_j) = \mathrm{softmax}(\boldsymbol{Q}_j\boldsymbol{K}_j^{\mathrm{T}})\boldsymbol{V}_j \tag{6}$$

式中　$\mathrm{MHA}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})$——多头注意力输出

$head$——注意力头

Concat——合并操作

$\boldsymbol{W}^o$——多头自注意力权重矩阵

$\mathrm{Attention}(\boldsymbol{Q}_j,\boldsymbol{K}_j,\boldsymbol{V}_j)$——当前词在自注意力层的输出

$\boldsymbol{Q}_j$、$\boldsymbol{K}_j$、$\boldsymbol{V}_j$—— 查询向量、键向量、值向量

上述方式对每个词集中不同词进行了权重调整，但不同词集之间的重要程度没有被区分，使用 4 个词集的目的是区分字符在词组中的不同位置，但当部分词集压缩之后的结果非常相似时，容易导致后续步骤不能明显区分字符 c_i 在所有词中所处的 4

种位置,使用4个词集的优势也相对被削弱。因此,为了进一步考虑各个词集的不同重要程度,本文采用注意力机制自动获取每个词集的重要程度,根据不同的重要程度增强重要的词集信息并抑制用处不大的词集信息,充分发挥4个词集的优势。注意力权重 $\boldsymbol{a}_i$ 的计算公式为

$$\boldsymbol{a}_i = \mathrm{sigmoid}(\boldsymbol{U}\tanh(\boldsymbol{W}\boldsymbol{V}_i^{\mathrm{T}})) \tag{7}$$

其中

$$\boldsymbol{V}_i = \begin{bmatrix} \boldsymbol{v}_i(\boldsymbol{B}) \\ \boldsymbol{v}_i(\boldsymbol{M}) \\ \boldsymbol{v}_i(\boldsymbol{E}) \\ \boldsymbol{v}_i(\boldsymbol{S}) \end{bmatrix} \tag{8}$$

式中 $\boldsymbol{V}_i$——4个词集合并后的矩阵,维度为 $4 \times d^w$

$\boldsymbol{W}$——权重矩阵,维度 $d \times d^w$

$\boldsymbol{U}$——权重矩阵,维度 $1 \times d^w$

d^w——词向量维度

最终得到重要度矩阵 $\boldsymbol{a}_i$ 的维度为 1×4,4个值分别代表4个词集的重要程度,使用该向量对4个词集进行重要度加权后可以得到更有说服力的词集表示。

为了避免本文模型受太多分词影响,在最终的嵌入层表示中融入了特征向量 $\boldsymbol{x}_{BERT}$,该向量是BERT预训练语言模型在大规模语料下通过学习上下文语义信息得到的,能够表征字的多义性,增强句子的语义表示,更好地挖掘结构复杂的猕猴桃种植领域命名实体特征信息。

将4个词集表示和字符向量连接后,得到字符的最终表示为

$$x_i = [\boldsymbol{x}_{BERT}^{c}(c_i); a_{i1}\boldsymbol{v}_i(\boldsymbol{B}); a_{i2}\boldsymbol{v}_i(\boldsymbol{M}); a_{i3}\boldsymbol{v}_i(\boldsymbol{E}); a_{i4}\boldsymbol{v}_i(\boldsymbol{S})] \tag{9}$$

式中 a_{i1}、a_{i2}、a_{i3}、a_{i4}——对应字符 c_i 的4个词集 $\boldsymbol{B}$、$\boldsymbol{M}$、$\boldsymbol{E}$、$\boldsymbol{S}$ 的重要程度

1.2 BiGRU 编码层

编码层将融合字词语义信息的嵌入层最终表示序列作为输入,对序列中的字与字之间的关系进行特征提取。采用GRU作为特征提取层,该网络与长短期记忆网络(Long short-term memory, LSTM)类似,与LSTM的区别是不再采用单元状态记录或传输信息,将遗忘门和输入门合并为一个单一的更新门,用隐藏状态控制信息传输和记录,用更新门和重置门控制隐藏层状态的最终输出,用隐藏状态控制信息传输和记录。

但单向的GRU只能获取目标词的前文信息。例如,针对猕猴桃病害实体“猕猴桃叶斑病”,目标词为“斑”,GRU只能提取到“斑”的前一个字“叶”的特征,提取不到后面“病”的特征。而目标词的上下文信息均会影响到对目标词的预测,进而影响命名实体的识别性能。因此,为了精确识别猕猴桃种植领域命名实体,本文采用双向GRU(BiGRU)网络模型。

BiGRU的输出由正向GRU和反向GRU组成,对于输入文本序列 $X = (\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_t, \cdots, \boldsymbol{x}_n)$,$\boldsymbol{x}_t$ 是 t 时刻的输入向量,正向GRU输出计算公式为

$$z_t = \sigma(\boldsymbol{W}_Z[h_{t-1}, \boldsymbol{x}_t]) \tag{10}$$

$$r_t = \sigma(\boldsymbol{W}_r[h_{t-1}, \boldsymbol{x}_t]) \tag{11}$$

$$\tilde{h}_t = \tanh(\boldsymbol{W}[r_t h_{t-1}, \boldsymbol{x}_t]) \tag{12}$$

$$h_t = (1 - z_t)h_{t-1} + z_t \tilde{h}_t \tag{13}$$

式中 z_t——更新门 r_t——重置门

σ——sigmoid 函数

$\tilde{h}_t$——t 时刻的候选状态

$\boldsymbol{W}_Z$、$\boldsymbol{W}_r$——权重矩阵

h_t、h_{t-1}——t 和 $t-1$ 时刻的输出

反向GRU与正向GRU的区别是,反向GRU从输入序列的最后一个字向前运算。$\boldsymbol{x}_t$ 经过正向和反向GRU编码后得到的输出为 $h_t = [\overrightarrow{h_t}; \overleftarrow{h_t}]$。

1.3 CRF 层

猕猴桃种植领域命名实体标签推理的任务是对序列文本中每个字符进行分类,类别包括B-VAR、M-DIS、O等。通过BiGRU编码层得到的特征向量是相互独立的,直接输入到全连接层中判定每个字符的标签时无法学习到文本标签间的约束关系,如B-VAR后面不可能是M-DIS。采用CRF全局优化来学习猕猴桃种植领域文本序列标签间的约束关系。

考虑到标签之间的约束关系,CRF引入一个转移矩阵 $\boldsymbol{A}$。对于输入句子 X 来说,输出标签序列 $y = \{y_1, y_2, \cdots, y_n\}$ 的得分定义为

$$\mathrm{score}(X, y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=0}^{n} N_{i, y_i} \tag{14}$$

式中 N_{i,y_i}——第 i 个字是标签 y_i 的概率

$A_{y_i, y_{i+1}}$——标签转移概率

$\mathrm{score}(X, y)$——输出序列得分

对所有输出序列 y 计算得分,采用Viterb动态规划算法得到猕猴桃种植领域文本序列标签的最优序列,进而对命名实体标签进行推理和预测。

2 实验设置

2.1 数据集

本文实验数据主要是通过爬虫框架,抓取百度百科和360百科网站有关猕猴桃种植的语料,少量

数据摘取自与猕猴桃种植领域相关的知网文献和书籍。对获取的句子进行清洗和去重后做人工标注，得到 12 477 个猕猴桃种植领域相关的句子作为最终实验数据集。

猕猴桃种植领域实体共 7 类，类别定义如表 1 所示。

表 1　猕猴桃种植领域实体类别

Tab. 1　Kiwifruit planting field entity category

类别符号	实体类别	类别定义	示例
VAR	品种	不同品种的猕猴桃名称	软枣猕猴桃、中华猕猴桃
DIS	病害	猕猴桃可能遭受的病害	猕猴桃软腐病、叶枯病
PEST	虫害	猕猴桃可能遭受的虫害	叶蝉、蝙蝠蛾、桑白蚧
PART	部位	各种病虫害危害猕猴桃部位	叶片、果实、根部、枝干
MED	药剂	处理各种病虫害的药剂	多菌灵、代森胺、敌百虫液
PLAC	区域	不同品种猕猴桃分布区域	陕西南部、云南北部、湖北
CLI	气候	不同区域的气候	温带大陆性气候、亚热带季风性气候

2.2　数据标注

采用 BMESO 标注策略：B（Begin）表示实体开始，M（Median）表示中间部分，E（End）表示结尾部分，S（Single）表示单个字符的实体，O（Other）表示其它非命名实体字符，并在最后加上实体类别。如“桑/B－PEST 白/M－PEST 蚧/E－PEST 的/O 主/O 要/O 危/O 害/O 部/O 位/O 是/O 叶/B－PART 片/E－PART”。在训练时，添加了保证训练长度一致的 PAD 占位符，同时用[CLS]和[SEP]标记句子的首部和尾部。

2.3　实验数据

将自建的实验数据按 7∶1∶2 划分为训练集、验证集和测试集，训练集 8 734 条，验证集 1 248 条，测试集 2 495 条，数据规模如表 2 所示。

表 2　实验数据规模

Tab. 2　Experimental data scale

项目	总字符数	句子数	句子平均字符数
训练集	1 144 154	8 734	131
验证集	158 496	1 248	127
测试集	329 340	2 495	132

数据集中包含实体 24 740 个，其中品种 5 364 个、病害 887 个、虫害 1 784 个、部位 7 985 个、药剂 1 314 个、区域 7 201 个、气候 205 个。不同类型实体在训练集、验证集和测试集中的统计如表 3 所示。

表 3　不同类别实体数据规模

Tab. 3　Data size of different types of entities

实体类别	训练集	验证集	测试集
品种	3 780	532	1 052
病害	561	105	221
虫害	1 248	132	404
部位	5 521	780	1 684
药剂	907	114	293
区域	5 098	685	1 418
气候	144	23	38
总计	17 259	2 371	5 110

2.4　实验环境及模型参数

实验环境：操作系统 Ubuntu 16.04，CUDA 9.2，cudnn 7.6.5；4 个 E5－2620 v4 @2.10 GHz 的 CPU，一个 GTX TITAN X GPU；内存 94 GB，显存 12 GB；编译环境为 Python 3.6.3 和 PyTorch 1.8.1。

本文实验使用 Glove 无监督模型在大规模猕猴桃种植语料下训练得到的词典。模型使用 Adam 优化算法进行参数调整，最大迭代次数为 50，选取其中最优结果作为最终实验结果。为了与其它方法对比，本文模型同样采用单层 BiGRU 网络。具体参数设置如表 4 所示。

表 4　模型参数设置

Tab. 4　Parameter setting of model

参数	数值
Glove 词向量维度	50
MHA 隐层维数	50
MHA 多头数目	5
MHA 归一化参数	1
GRU 层数	1
GRU 隐层维数	300
全局归一化参数	0.5
全局学习率	0.001 5

2.5　评价指标

命名实体识别的评价标准有精确率（Precision，P）、召回率（Recall，R）和 F1 值（F-measure）。

3　实验结果与分析

3.1　不同模型实体识别结果对比

为了验证本文模型在猕猴桃种植领域的命名实体识别效果，在同一实验环境下，使用不同模型进行对比实验，对比模型包括：BiLSTM－CRF 模型[22]、Lattice－LSTM 模型[18]、WC－LSTM 模型[19]、SoftLexicon 模型[20] 和 BERT－BiLSTM－CRF 模型[23]，6 组实验结果如表 5 所示。

表 5 不同模型实体识别结果

Tab. 5 Entity recognition results of different models

%

模型	P	R	F1 值
BiLSTM - CRF[22]	86.57	89.84	88.17
Lattice - LSTM[18]	87.65	90.52	89.06
WC - LSTM[19]	88.41	91.63	89.99
SoftLexicon[20]	88.14	92.63	90.33
BERT - BiLSTM - CRF[23]	88.89	92.29	90.56
本文模型	90.32	93.56	91.91

本文 BiLSTM - CRF 模型使用 Glove 无监督模型训练得到的字向量作为嵌入层,然后将其输入到 BiLSTM - CRF 中进行序列标注,虽然 Glove 模型得到的字向量能够在一定程度上捕捉到字的一些语义特性,但识别的 F1 值只有 88.17%,识别效果不佳。Lattice - LSTM 模型使用注意力机制对当前字符匹配到的词进行加权求和,显著提高了识别效果,F1 值提高 0.89 个百分点,说明引入外部词汇可以有效提高猕猴桃种植领域命名实体识别性能。WC - LSTM 模型对当前字符所有匹配到的词采用自注意力编码生成权重向量后,与字向量直接拼接得到最终的字符表示,序列编码依旧使用 BiLSTM - CRF 结构,F1 值高达 89.99%。SoftLexicon 模型为了简化模型结构,在嵌入层使用词频代替注意力加权的方式,同时加入了表示当前字符在词不同位置的 4 个词集标签,将压缩得到的 4 个词集向量和字符向量进行拼接得到最终字符表示,与使用注意力加权的 WC - LSTM 模型相比,F1 值提高 0.34 个百分点,说明引入 4 种词集信息可以有效提高文本命名实体的识别性能。BERT - BiLSTM - CRF 模型只使用 BERT 预训练增强的字符向量作为嵌入层,F1 值高达 90.56%,与使用 Glove 模型字向量作为嵌入层相比,F1 值提高 2.39 个百分点,说明 BERT 预训练模型可以学习到更全面的字符特征。

本文模型采用 MHA 和 Attention 对词和词集加权,将得到的词集向量与 BERT 预训练语言模型得到字符向量融合后作为字符的最终表示,并使用 BiGRU 进行序列编码,CRF 模型进行标签推理,实验结果表明,本文模型的 F1 值达 91.91%,相较于其它模型,本文模型在猕猴桃种植领域命名实体识别任务中表现更加出色。

3.2 MHA 和 Attention 影响对比实验

为了验证添加 MHA 调整词权重和 Attention 获取词集重要程度对模型的影响,本文对添加 MHA 和 Attention 进行对比实验,实验结果如表 6 所示。

表 6 MHA 和 Attention 影响实验结果

Tab. 6 MHA and Attention affected experiment results

%

模型	P	R	F1 值
SoftLexicon[20]	88.14	92.63	90.33
SoftLexicon + MHA	88.96	92.82	90.85
SoftLexicon + Attention	88.83	92.41	90.58
SoftLexicon + MHA + Attention	89.25	93.13	91.15

从表 6 可以看出,在 SoftLexicon 模型中添加 MHA 对词集中的词进行权重调整时,F1 值提高 0.52 个百分点;当在 SoftLexicon 模型中添加 Attention 调整词集取值时,F1 值提高 0.25 个百分点,同时添加两者时,精确率、召回率和 F1 值均有显著提高,与未添加任何机制的 SoftLexicon 模型相比,F1 值总体提高 0.82 个百分点,比单独添加 MHA 或 Attention 机制效果都好。因此,使用 MHA 对词向量进行加权和使用 Attention 对词集向量进行调整可以提升模型性能,两者同时使用可以进一步提升猕猴桃种植领域命名实体识别性能。

3.3 字词融合对比实验

为了验证使用 BERT 预训练语言模型增强字符表示和引入词集向量作为外部词汇对于模型的提升效果,分别使用 Glove 字符表示和词集向量融合表示、基于 BERT 的字符增强表示以及 BERT 字符增强和词集向量融合的表示作为嵌入,其中 SoftLexicon 表示以 Glove 字符表示和词集向量融合表示作为嵌入层的模型,结果如表 7 所示。

表 7 字词融合实验结果

Tab. 7 Char and word fusion experiment results %

模型	P	R	F1 值
SoftLexicon[20]	88.14	92.63	90.33
BERT	88.89	92.29	90.56
BERT + 词集	89.52	92.58	91.02

本文词集向量均没有添加 MHA 和 Attention 机制,仅使用词频作为权重计算词集信息,编码层均为单层的 BiLSTM。SoftLexicon 模型融合 Glove 字符向量和词集向量作为嵌入,由于 Glove 模型特征提取能力有限,无法获取更全面的语义信息,得到的字词向量包含上下文信息较少,而且猕猴桃种植领域实体专业性较强,结构复杂,从而导致模型的识别性能不佳。使用 BERT 预训练语言模型得到字符向量作为嵌入层时,与使用 Glove 字词向量相比,其 F1 值提高 0.23 个百分点,原因是 BERT 预训练语言模型可以提取出序列中与领域相关的更丰富的上下文信息,增强字符表示。融合 BERT 增强的字符向量和词集信息作为嵌入层时,识别性能有了显著提高,其

F1 值高达 91.02%，与使用 Glove 字词向量和单纯使用 BERT 字符向量相比，F1 值分别提高 0.69 个百分点和 0.46 个百分点，表明使用 BERT 预训练语言模型增强的字符表示和引入外部词汇信息融合的方式确实可以提高本文猕猴桃种植领域命名实体识别效果。

3.4　编码层对比实验

为了验证 BiGRU 编码层对模型的影响，分别使用 BiLSTM、CNN、Transformer 和 BiGRU 作为编码层进行实验对比，结果如表 8 所示。

表 8　编码层实验结果

Tab.8　Experimental results of coding layer　%

模型	*P*	*R*	F1 值
BiLSTM	88.14	92.63	90.33
CNN	79.77	90.50	84.80
Transformer	85.02	90.52	87.68
BiGRU	88.77	92.48	90.59

从表 8 可以看出，BiGRU 作为编码层时，与 BiLSTM、CNN 或 Transformer 作为编码层相比，模型的识别效果最好，F1 值达到 90.59%，说明使用 BiGRU 作为编码层更适合猕猴桃种植领域命名实体识别任务，可以进一步提高命名实体识别水平。

3.5　嵌入层通用性实验

本文提出的添加 MHA 和 Attention 机制以及使用 BERT 预训练语言模型的方法仅改变了嵌入层，可以与不同的序列建模层联合使用，具有较好的通用性。为了验证不同序列建模层在本文模型中的通用性，将序列建模层的单层 BiGRU 更换为 CNN，卷积层个数为 2，卷积核大小为 1 和 3，通用性实验结果如表 9 所示。

表 9　通用性实验结果

Tab.9　Performance of commonality test　%

模型	*P*	*R*	F1 值
基于字的 CNN	78.25	84.26	81.14
SoftLexicon(CNN)	79.77	90.50	84.80
本文模型(CNN)	84.02	93.86	88.67

由表 9 可知，本文模型的识别效果最优，与基于字的 CNN 和 SoftLexicon 模型相比，精确率分别提高 5.77、4.25 个百分点。说明本文模型能够更好地利用外部词汇信息，具有更好的通用性。并且与 表 7 进行对比时，可以看出编码层使用 BiGRU 时，模型识别效果更优。

3.6　消融实验

为了验证本文模型嵌入层各个部分对整体模型的影响，对添加 MHA 机制、Attention 机制和 BERT 预训练模型增强的字符进行消融实验，实验结果如表 10 所示。

表 10　消融实验结果

Tab.10　Ablation experimental results　%

模型	*P*	*R*	F1 值
本文模型	90.32	93.56	91.91
本文模型 - MHA	90.52	93.01	91.75
本文模型 - Attention	89.32	93.61	91.42
本文模型 - BERT	89.22	93.45	91.29

从表 10 可以看出，使用 BERT 预训练语言模型增强的字符表示对模型性能提升最明显，F1 值提高 0.62 个百分点，相比之下，添加 MHA 机制的提升效果最小，但总体来看，本文提出的每个改进点，均对模型性能有一定程度的提升。

3.7　模型各类实体识别效果分析

表 11 对比了 BiLSTM - CRF 模型[22]、Lattice - LSTM 模型[18]、WC - LSTM 模型[19]、SoftLexicon 模型[20]、BERT - BiLSTM - CRF 模型[23]和本文模型在 7 类实体上的识别效果。

从表 11 可以看出，本文提出的模型识别效果优于其它模型。使用 BERT 预训练的字向量作为嵌入层时，与 SoftLexicon 模型相比，除病害类别外，其它 6 种类别实体的识别效果均有所提升，说明使用 BERT 预训练的字向量可以有效提升本文命名实体识别效果。本文模型部位类别识别的 F1 值高达 96.87%，病害类别识别的 F1 值为 96.17%，对于实体结构复杂的虫害识别 F1 值高达 95.70%，与 SoftLexicon 模型相比，在 7 种类别实体上识别效果均有所提升，说明融合 BERT 预训练语言模型增强的字符表示和添加不同层次注意力机制等方法可以有效提升本文模型在猕猴桃种植领域实体识别效果。本文模型与 BERT - BiLSTM - CRF 模型相比，在 6 种实体类别上也有不同幅度的提升，进一步验证了本文方法在猕猴桃种植领域实体识别任务上的优势。与 SoftLexicon 模型相比，本文模型对虫害类别的 F1 值提升最高，提升 3.13 个百分点，原因是该类别存在虫害嵌套、歧义等干扰信息，在没有其它充足的上下文语义信息时容易预测错误，例如在识别“棉红蜘蛛”和“红蜘蛛”、“盲椿象”和“椿象”、“二点叶螨”和“叶螨”等实体时，SoftLexicon 模型只识别出“红蜘蛛”、“椿象”和“叶螨”等，从而造成实体识别效果差，而本文模型则可以识别出正确的实体。

3.8　模型泛化性分析

为了验证本文模型泛化性和稳定性，本文在 ResumeNER 公开数据集上开展了实验。实验结果

表 11 不同类别实体识别结果

Tab. 11 Entity recognition results of different types %

模型	评价指标	品种	病害	虫害	部位	药剂	区域	气候
BiLSTM – CRF[22]	*P*	88.03	96.05	91.13	91.59	66.00	81.21	50.00
	R	87.13	94.19	91.13	96.25	71.74	87.08	40.00
	F1 值	87.58	95.11	91.13	93.86	68.75	84.04	44.44
Lattice – LSTM[18]	*P*	88.40	94.85	91.67	94.18	62.68	82.95	66.67
	R	89.71	93.88	92.40	95.82	76.78	87.74	66.67
	F1 值	89.05	94.36	92.03	94.99	69.02	85.28	66.67
WC – LSTM[19]	*P*	88.61	95.49	92.32	94.62	64.82	84.43	66.67
	R	91.05	94.02	93.04	96.60	77.16	88.05	80.00
	F1 值	89.81	94.75	92.68	95.60	70.45	86.20	72.73
SoftLexicon[20]	*P*	89.17	94.87	91.64	95.33	64.00	84.81	66.67
	R	91.08	95.48	93.52	96.92	78.26	86.72	80.00
	F1 值	90.12	95.17	92.57	96.12	70.42	85.75	72.73
BERT – BiLSTM – CRF[23]	*P*	91.04	93.67	92.72	94.26	69.63	86.21	66.67
	R	91.76	95.48	95.56	97.17	72.28	85.82	80.00
	F1 值	91.40	94.57	94.12	95.69	70.93	86.01	72.73
本文模型	*P*	92.37	95.87	95.86	95.14	69.23	85.93	80.00
	R	92.99	96.48	95.54	98.67	72.20	89.08	80.00
	F1 值	92.68	96.17	95.70	96.87	70.68	87.48	80.00

如表 12 所示。结果表明,本文模型表现良好,F1 值达到 96.17%,显著高于 BiLSTM – CRF 模型[22],与 Lattice – LSTM[18]、WC – LSTM[19]、SoftLexicon[20]、BERT – BiLSTM – CRF[23]模型相比也均有提升。

表 12 各模型在公开数据集上识别效果对比

Tab. 12 Comparison of recognition effect of each model on public data set %

模型	*P*	*R*	F1 值
BiLSTM – CRF[22]	93.66	93.31	93.48
Lattice – LSTM[18]	94.81	94.11	94.46
WC – LSTM[19]	95.14	94.79	94.96
SoftLexicon[20]	95.30	95.77	95.53
BERT – BiLSTM – CRF[23]	95.75	95.28	95.51
本文模型	96.25	96.08	96.17

4 结束语

本文面向猕猴桃种植领域,提出一种融合字词语义信息的命名实体识别模型,有效解决了猕猴桃种植领域命名实体结构复杂、识别精确率较低的问题。该模型使用 MHA 调整词向量权重,并使用注意力机制进一步获取每个词集的重要程度,使模型更好地利用外部词汇信息,融入 BERT 预训练语言模型提取的字符增强表示,使嵌入层输出包含更丰富的上下文信息,编码层使用 BiGRU 模型进一步提高识别效果。通过实验证明,本文模型对 7 种猕猴桃种植领域实体的识别 F1 值高达 91.91%,在公开数据集 ResumeNER 上也有较好的效果。

参 考 文 献

[1] 姜正旺, 钟彩虹. 试论猕猴桃科普与果实品质提升的重要性[J]. 中国果树, 2020(1): 1 – 8.
JIANG Zhengwang, ZHONG Caihong. On the importance of popularizing kiwifruit and improving fruit quality[J]. China Fruits, 2020(1): 1 – 8. (in Chinese)

[2] GIZEM A, DIDEM M, SENIZ D, et al. An evaluation of recent neural sequence tagging models in Turkish named entity recognition[J]. arXiv:2005.07692, 2021.

[3] 李想, 魏小红, 贾璐, 等. 基于条件随机场的农作物病虫害及农药命名实体识别[J]. 农业机械学报, 2017, 48(增刊): 178 – 185.
LI Xiang, WEI Xiaohong, JIA Lu, et al. Recognition of crops, diseases and pesticides named entities in Chinese based on conditional random fields[J]. Transactions of the Chinese Society for Agricultural Machinery, 2017, 48(Supp.): 178 – 185. (in Chinese)

[4] 张剑, 吴青, 羊昕旖, 等. 基于条件随机场的农业命名实体识别[J]. 计算机与现代化, 2018(1): 123 – 126.
ZHANG Jian, WU Qing, YANG Xinyi, et al. Chinese agricultural named entity recognition based on conditional random fields [J]. Computers and Modernization, 2018(1): 123 – 126. (in Chinese)

[5] 张栋, 陈文亮. 基于上下文相关字向量的中文命名实体识别[J]. 计算机科学, 2021, 48(3): 233 – 238.
ZHANG Dong, CHEN Wenliang. Chinese named entity recognition based on context-dependent word vector[J]. Computer

Science, 2021, 48(3): 233 – 238. (in Chinese)
[6] 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.
SONG Linpeng, LIU Shihong, WANG Cui. Text entity extraction of agricultural technical requirements based on word vector + BiLSTM + CRF[J]. Jiangsu Agricultural Sciences,2021,49(5):186 – 193. (in Chinese)
[7] 谢聪娇,高静,陈俊杰. 面向农作物病虫害领域的命名实体识别[J]. 内蒙古农业大学学报(自然科学版),2022,43(1):86 – 90.
XIE Congjiao, GAO Jing, CHEN Junjie. Named entity recognition for the field of crop pests and diseases[J]. Journal of Inner Mongolia Agricultural University(Natural Science Edition),2022,43(1):86 – 90. (in Chinese)
[8] MENG Y, LI X, SUN X, et al. Is word segmentation necessary for deep learning of Chinese representations[J]. arXiv preprint arXiv:1905.05526, 2019.
[9] 沈利言, 姜海燕, 胡滨, 等. 水稻病虫草害与药剂实体关系联合抽取算法[J]. 南京农业大学学报, 2020, 43(6): 1151 – 1161.
SHEN Liyan, JIANG Haiyan, HU Bin,et al. A study on joint entity recognition and relation extraction for rice diseases pests weeds and drugs[J]. Journal of Nanjing Agricultural University, 2020, 43(6): 1151 – 1161. (in Chinese)
[10] 郭旭超, 唐詹, 刁磊, 等. 基于部首嵌入和注意力机制的病虫害命名实体识别[J]. 农业机械学报, 2020, 51(增刊2): 335 – 343.
GUO Xuchao, TANG Zhan, DIAO Lei,et al. Named entity recognition of pests and diseases based on radical embedding and attention mechanism[J]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(Supp.2): 335 – 343. (in Chinese)
[11] 赵鹏飞, 赵春江, 吴华瑞, 等. 基于注意力机制的农业文本命名实体识别[J]. 农业机械学报, 2021, 52(1): 185 – 192.
ZHAO Pengfei, ZHAO Chunjiang, WU Huarui, et al. Named entity recognition of agricultural text based on attention mechanism[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(1): 185 – 192. (in Chinese)
[12] DEVLIN J, CHANG M, 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. Stroudsburg, PA: Association for Computational Linguistics, 2019: 4171 – 4186.
[13] 焦凯楠,李欣,朱容辰. 中文领域命名实体识别综述[J]. 计算机工程与应用,2021,57(16):1 – 15.
JIAO Kainan, LI Xin, ZHU Rongchen. Overview of Chinese domain named entity recognition[J]. Computer Engineering and Applications,2021,57(16):1 – 15. (in Chinese)
[14] 李亮德,王秀娟,康孟珍,等. 基于语义融合与模型蒸馏的农业实体识别[J]. 智慧农业(中英文),2021,3(1):118 – 128.
LI Liangde, WANG Xiujuan, KANG Mengzhen, et al. Agricultural entity recognition based on semantic fusion and model distillation[J]. Smart Agriculture,2021,3(1):118 – 128. (in Chinese)
[15] 岳琪,李想. 基于 BERT 和双向 RNN 的中文林业知识图谱构建研究[J]. 内蒙古大学学报(自然科学版),2021,52(2):176 – 184.
YUE Qi, LI Xiang. Construction of Chinese forestry knowledge graph based on BERT and bidirectional RNN[J]. Journal of Inner Mongolia University(Natural Science Edition),2021,52(2):176 – 184. (in Chinese)
[16] 陈晓玲,唐丽玉,胡颖,等. 基于 ALBERT 模型的园林植物知识实体与关系抽取方法[J]. 地球信息科学学报,2021,23(7):1208 – 1220.
CHEN Xiaoling, TANG Liyu, HU Ying, et al. Knowledge entity and relation extraction method of garden plants based on ALBERT model[J]. Journal of Geo-information Science,2021,23(7):1208 – 1220. (in Chinese)
[17] 李林,周晗,郭旭超,等. 基于多源信息融合的中文农作物病虫害命名实体识别[J]. 农业机械学报,2021,52(12):253 – 263.
LI Lin, ZHOU Han, GUO Xuchao, et al. Research on named entity recognition of pests and diseases based on multi-source information fusion[J]. Transactions of the Chinese Society for Agricultural Machinery,2021,52(12):253 – 263. (in Chinese)
[18] ZHANG Y, YANG J. Chinese NER using lattice LSTM[J]. arXiv preprint arXiv:1805.02023, 2018.
[19] LIU W, XU T, XU Q, et al. An encoding strategy based word-character[C]// Proceedings of the 2019 Conference of the North, 2019: 2379 – 2389.
[20] MA R, PENG M, ZHANG Q, et al. Simplify the usage of Lexicon in Chinese NER[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, 2020.
[21] 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, 2017:5998 – 6008.
[22] HUANG Z, WEI X, KAI Y. Bidirectional LSTM – CRF models for sequence tagging[J]. arXiv:1508.01991.
[23] 王子牛,姜猛,高建瓴,等. 基于 BERT 的中文命名实体识别方法[J]. 计算机科学,2019,46(增刊2):138 – 142.
WANG Ziniu, JIANG Meng, GAO Jianling, et al. Chinese named entity recognition method based on BERT[J]. Computer Science,2019,46(Supp.2):138 – 142. (in Chinese)