doi:10.6041/j.issn.1000-1298.2023.11.028

词嵌入 BERT－CRF 玉米育种实体关系联合抽取方法

李书琴　庞文婷
（西北农林科技大学信息工程学院，陕西杨凌 712100）

摘要：针对玉米育种文本数据中存在重叠三元组、实体表达方式多样等问题，提出一种嵌入词汇信息的 BERT－CRF（Bidirectional encoder representations from transformers－conditional random field）玉米育种实体关系联合抽取方法。首先，分析了玉米育种语料表达特征，采用对实体边界、关系类别和实体位置信息同步标注的策略；其次，构建了嵌入词汇信息的 BERT－CRF 模型进行训练和预测，自建玉米育种知识词典，通过在 BERT 中嵌入词汇信息，融合字符特征和词汇特征，增强模型的语义能力，利用 CRF 模型输出全局最优标签序列，设计了实体关系三元组匹配算法（Entity and relation triple matching algorithm，ERTM），将标签进行匹配和映射来获取三元组；最后，为验证该方法的有效性，在玉米育种数据集上进行实验，结果表明，本文模型精确率、召回率和 F1 值分别为 91.84%、95.84%、93.80%，与现有模型相比性能均有提升。说明该方法能够有效抽取玉米育种领域知识，为构建玉米育种知识图谱及其它下游任务提供数据基础。

关键词：玉米育种；实体关系联合抽取；序列标注；BERT 语言模型；词嵌入
中图分类号：TP391.1　　**文献标识码**：A　　**文章编号**：1000-1298(2023)11-0286-09　　**OSID**：

Joint Extraction Method of Entity and Relation in Maize Breeding Based on BERT－CRF and Word Embedding

LI Shuqin　PANG Wenting
(*College of Information Engineering*, *Northwest A&F University*, *Yangling*, *Shaanxi* 712100, *China*)

Abstract: Aiming at the problems of overlapping triples and diverse entity expressions in maize breeding text data, a joint bidirectional encoder representations from transformers－conditional random field (BERT－CRF) maize breeding entity relation extraction method with embedded lexical information was proposed. Firstly, the expression characteristics of maize breeding corpus were analyzed, and a synchronous labeling strategy for entity boundary, relation type, and entity position information was adopted. Secondly, a BERT－CRF model with embedded lexical information was constructed for training and prediction, a self-built dictionary of maize breeding knowledge was designed to enhance the semantic ability of the model by embedding lexical information in BERT, integrating character features and lexical features, and using CRF model to output the globally optimal label sequence, and an entity and relation triple matching algorithm (ERTM) was designed to obtain triples by mapping and matching labels. Finally, in order to verify the effectiveness of the proposed method, experiments were carried out on maize breeding data set. The results showed that the precision, recall and F1 value were 91.84%, 95.84% and 93.80%, respectively, which improved the performance compared with the existing models. This method can extract maize breeding knowledge effectively and provide data basis for constructing maize breeding knowledge graph and other downstream tasks.

Key words: maize breeding; joint extraction of entity and relation; sequence labeling; BERT language model; word embedding

收稿日期：2023－04－28　修回日期：2023－05－25
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

0　引言

玉米育种对保障国家粮食安全作出了重要贡献。近年来，我国玉米品种审定数量急剧增加，产生了海量的信息资源[1]，这些信息多以半结构化和非结构化形式呈现在网页和书籍中，不同品种之间的联系没有得到有效梳理，限制了育种效率的提升[2]。知识图谱是一种以节点和边的形式表示知识的数据结构[3]，提供了一套解决数据之间关联性的方案，可以高效整合海量数据的有效信息[4－5]。从原始的领域数据到形成高质量知识图谱，需要经过知识抽取、知识融合与知识加工等多个步骤，其中知识抽取是关键环节[6]。因此，如何有效抽取实体关系对于玉米育种业信息化发展具有重要作用。

目前，实体关系抽取方法分为流水线法和联合抽取法两大类[7]。其中，流水线法将实体关系抽取分解为实体识别[8]和关系抽取[9]2 个独立的子任务。CHAN 等[10]采用流水线方法，识别出语料中的潜在实体，然后组合其中任意 2 个实体进行关系抽取。黄梅根等[11]使用流水线方法，提出一种多关系抽取模型，能够有效提高中文多关系三元组的抽取效果。虽然上述流水线方法易于实现且灵活性高，但是将实体识别和关系抽取分开处理容易忽视这 2 个任务之间的联系。此外，实体识别产生的错误会传递到关系分类模型中，进而影响关系抽取的性能[12－13]。

与流水线法不同，联合抽取法可以有效整合实体和关系信息，从而提高知识抽取性能[14]。一些神经网络模型[15－16]通过参数共享机制[17]联合建模，但存在无法剔除冗余实体的问题[18]。ZHENG 等[19]提出了一种序列标注方式来解决联合抽取问题，通过该标注方式可以同时识别出实体和关系，在通用领域取得了不错的效果，但不能有效处理重叠三元组（同一句子中多个三元组共享相同实体）的问题[20]。而在玉米育种语料中，同一个实体常常参与多个关系，若使用该标注方式可能会损失大量的关系。

曹明宇等[21]改进了标注方式，提出一种基于 BiLSTM－CRF 的实体关系联合抽取方法，在药物数据集上取得了较好的效果，但该方法并未完全涵盖所有重叠三元组。胡滨等[22]提出一种基于双重指针标注策略的实体关系联合抽取模型，用来解决家禽诊疗文本中重叠三元组难以抽取的问题。沈利言等[23]采用双向长短期记忆网络与注意力机制相结合的方法，有效提高了水稻病虫草害防治语料中实体关系联合抽取的准确率。但是农业各领域文本中涉及的实体术语专业性较强，在识别没有明确边界特征的实体时仍存在一定的难度。

基于以上研究和问题，本文提出一种面向玉米育种领域的实体关系联合抽取方法。充分考虑玉米育种语料特征，采用对实体边界、关系类别和实体位置信息同步标注策略，将实体关系联合抽取任务转换为序列标注任务。构建嵌入词汇信息的 BERT－CRF（Bidirectional encoder representations from transformers－conditional random field）玉米育种实体关系抽取模型，将字符特征和词汇特征在 BERT 中进行融合，增强模型的语义能力，利用 CRF 解码预测最佳序列，设计实体关系三元组匹配算法（Entity and relation triple matching algorithm，ERTM）来获取三元组，以提升知识抽取性能。

1　玉米育种语料特征与标注策略

1.1　玉米育种语料特征

玉米育种语料主要包括玉米品种生育期、品质、农艺性状和病虫害等信息，通过分析语料表达形式，有如下 2 个特征：

（1）存在重叠三元组。玉米育种语料中，同一实体常常参与多个关系。例如，语料“金禧甜玉米抗玉米纹枯病，适宜在广东种植”，按照玉米育种知识图谱结构特征，包括（金禧甜玉米，DiseaseResistance，玉米纹枯病）和（金禧甜玉米，PlantingArea，广东）2 个三元组。其中，“金禧甜玉米”是重叠实体，处于 2 种不同关系中，2 个三元组发生了重叠。

（2）实体边界不清晰。玉米育种语料中包括大量专有名词以及由数字、字母、汉字和特殊字符组成的实体，边界特征不明显，例如玉米虫害实体“二点委夜蛾”、“弯刺黑蝽”，玉米品种实体“瑞禾白糯 80”、“K 玉 788”等。

1.2　标注策略

为了解决上述玉米育种语料中的重叠三元组问题，本文将实体关系联合抽取转换为序列标注任务，设计了对实体边界、关系类别和实体位置信息同步标注的策略，具体描述如下：

（1）对于实体边界，采用“BIO”标注方式，B（Begin）－Entity 表示实体中的第 1 个字符，I（I-inside）－Entity 表示实体的中间和结尾字符，O（Other）表示非实体字符。

（2）对于关系类别，使用玉米育种知识图谱模式层定义的关系类别表示，此外，增加一个“OVE”标签表示重叠实体的关系。

（3）对于实体位置，使用“H”和“T”来表示，其

中,“H”表示该实体为三元组中的头实体,“T”表示该实体为三元组中的尾实体。

如图 1 所示,“金禧甜玉米”是重叠实体,关系类别标签表示为“OVE”,并且“金禧甜玉米”实体在 2 个三元组中都处于头实体的位置,所以实体位置标签使用“H”表示。“玉米纹枯病”实体与“金禧甜玉米”实体之间存在“DiseaseResistance”关系,则将“玉米纹枯病”的关系标签标注为“RDIS”;“广东”实体与“金禧甜玉米”实体存在“PlantingArea”关系,则将“广东”的关系标签标注为“PARE”,并且这 2 个实体均是尾实体,所以实体位置标签使用“T”表示。非实体用 O 来标注。

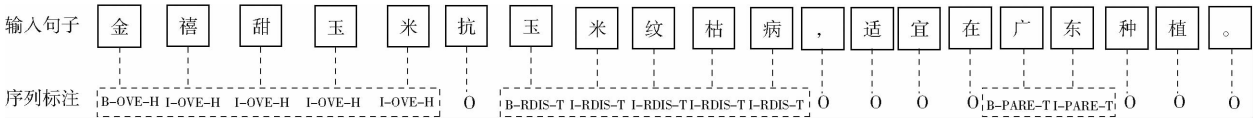

图 1 标注方式示例

Fig. 1 Example of tagging method

2 嵌入词汇信息的 BERT - CRF 抽取模型

由字符组成的词汇可以携带更多的语义,从而提供更多的实体边界信息[24-25]。针对玉米育种语料中存在实体边界不清晰的问题,本文构建了嵌入词汇信息的 BERT - CRF 玉米育种实体关系抽取模型。首先构建玉米育种知识词典,并采用 AC(Aho - corasick)算法[26]为输入句子中每个字符匹配所有可能的词汇;其次将字符特征和词汇特征在 BERT 模型中进行融合;然后通过 CRF 层进行解码;最后在三元组抽取层,设计了一种 ERTM 算法将有关系的玉米育种实体对抽取为三元组。模型整体结构如图 2 所示。

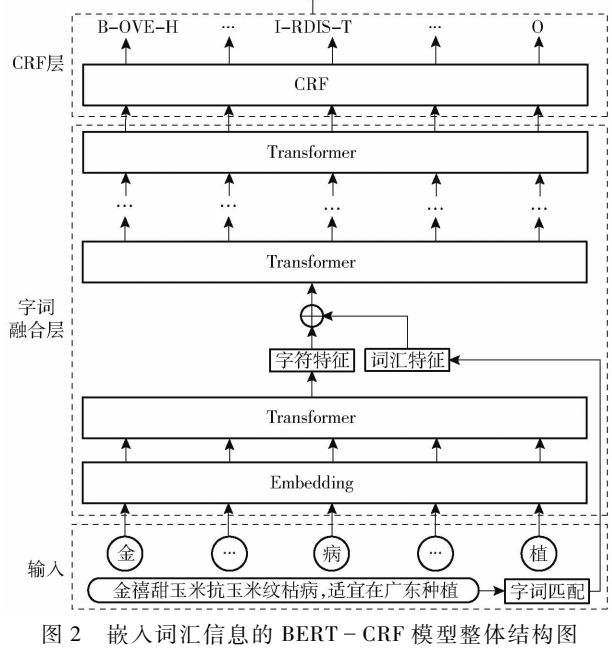

图 2 嵌入词汇信息的 BERT - CRF 模型整体结构图

Fig. 2 Overall framework of BERT - CRF model with lexical information

2.1 字词匹配

为了利用词汇特征,本文从中国玉米品种系谱数据库、种业商务网、玉米育种书籍和文献中提取相应词汇,构建了玉米育种知识词典,并通过 Glove 模型转换为词向量,共计 7 583 个词汇。相比于常用的 KMP 和前后向最大匹配等匹配算法,AC 算法可以在 $O(n)$ 的时间复杂度以内仅扫描一次输入,就能匹配到词典中所有出现过的子字符串,匹配效率较高,因此本文采用 AC 算法进行字词匹配。将输入句子表示为 $s=\{c_1,c_2,\cdots,c_n\}$,c_i 表示第 i 个字符,n 表示字符个数,根据 AC 算法得到句子中的潜在词汇,AC 算法步骤如下:

(1)构建 Trie 树:将待匹配的词汇构建成 Trie 树,通过 Trie 树的分支和叶子节点,可以快速定位到词汇中的每个字符。

(2)设置 fail 指针:对 Trie 树进行广度优先遍历,为每个节点设置 fail 指针,在匹配失败时能够进行状态转移,将 Trie 树扩展成 AC 自动机。

(3)词汇匹配:根据形成的自动机进行词典匹配,得到句子中的潜在词汇。

例如,以{“金禧甜玉米”,“玉米纹枯病”,“玉米螟”}为例构建 Trie 树,AC 自动机的状态如图 3 所示。然后将 s 中的每个字符与包含该字符的所有词汇转换为字词对 $s_{cw}=\{(c_1,ws_1),(c_2,ws_2),\cdots,(c_n,ws_n)\}$ 的形式,其中 ws_i 是由包含 c_i 的词汇组成的集合。

2.2 字词融合层

与传统语言模型相比,BERT[27]拥有强大的语言表征能力和特征提取能力,可以深层次地理解上下文信息。因此,本文采用 BERT 模型来获取丰富的字符特征。BERT 模型由 Embedding 层和 Transformer 层组成。首先将包含 n 个字符的句子 s 输入 Embedding 层进行编码,包括字嵌入(Token embeddings)、句子嵌入(Segment embeddings)和位置嵌入(Position embeddings),分别得到字向量、句子向量和位置向量,其中,字向量是输入句子中每个字符本身的向量表示,句子向量用来表示句子的全

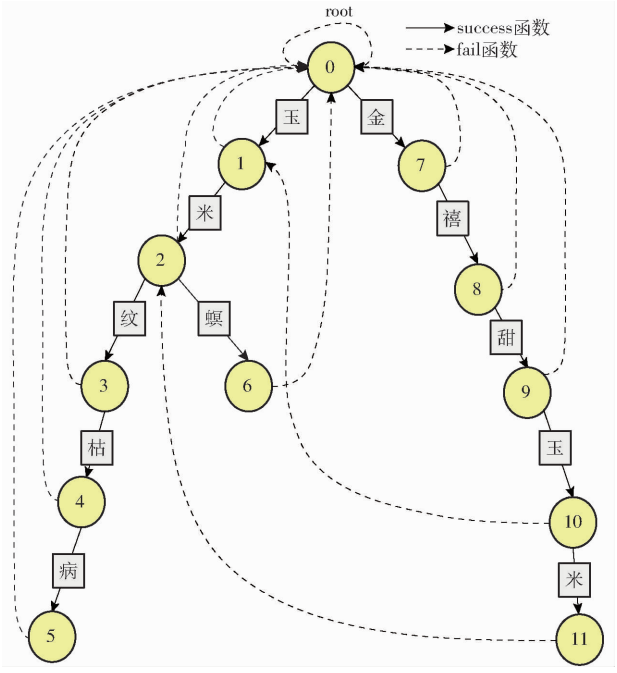

图 3　AC 自动机状态图

Fig. 3　AC automata state diagram

局语义信息，位置向量用来表示字符在句子中对应的位置信息，并添加[CLS]和[SEP]符号来表示句子的开始位置与结束位置。输入向量如图 4 所示。

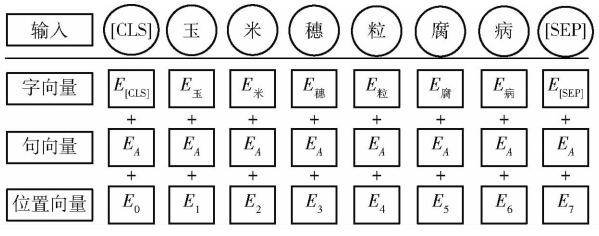

图 4　BERT 的输入向量

Fig. 4　Input vector of BERT

实体边界一般对应于词汇边界，而 BERT 模型以字符为最小粒度进行切分，仅能获取字符级嵌入，所包含的语义信息有限。传统引入词汇信息的方法采用中文分词工具对输入文本进行分词，但是分词的效果会严重影响模型的效果，若分词不正确，会导致实体边界的错误识别。为了更充分地利用词汇信息，本文引入了上文构建的玉米育种词典特征信息，将字符特征和词汇特征在 BERT 中进行融合，以实现对字符和词汇级信息的同时捕捉，解决玉米育种实体边界不清晰的问题。融合方法如下：

输入句子转换为字词对序列后，将其对应的字词向量对表示为$(\boldsymbol{u}_i^c, \boldsymbol{v}^{ws})$，其中 $\boldsymbol{u}_i^c$ 为 Transformer 层输出的第 i 个字符的字符向量，$\boldsymbol{v}^{ws} = \{\boldsymbol{v}_{i1}^w, \boldsymbol{v}_{i2}^w, \cdots, \boldsymbol{v}_{im}^w\}$ 是第 i 个字符所对应的词向量集合，m 为词向量的个数，$\boldsymbol{v}^{ws}$ 中的第 j 个词向量表示为 $\boldsymbol{v}_{ij}^w$。

此时字符向量 $\boldsymbol{u}_i^c$ 和词向量 $\boldsymbol{v}_{ij}^w$ 的维度是不一样的，变换词向量的维度进行对齐操作，对齐后的词向量表示为 $\boldsymbol{p}_{ij}^w$，计算公式为

$$\boldsymbol{p}_{ij}^w = \boldsymbol{W}_2(\tanh(\boldsymbol{W}_1\boldsymbol{v}_{ij}^w + \boldsymbol{b}_1)) + \boldsymbol{b}_2 \tag{1}$$

式中　$\boldsymbol{W}_1$——参数矩阵，维度 $d_c \times d_w$

$\boldsymbol{W}_2$——参数矩阵，维度 $d_c \times d_c$

d_c——字符特征维度

d_w——词嵌入维度

$\boldsymbol{b}_1$、$\boldsymbol{b}_2$——标量偏置

通过维度变换后，第 i 个位置的词向量集合表示为 $\boldsymbol{P} = \{\boldsymbol{p}_{i1}^w, \boldsymbol{p}_{i2}^w, \cdots, \boldsymbol{p}_{im}^w\}$。每个字符都匹配了多个词汇，但是这些词汇的重要性是不同的，因此采用注意力机制计算每个词向量的权重 $r = \{r_{i1}, r_{i2}, \cdots, r_{in}\}$，$r_{ij}$ 表示第 i 个字符不同词向量的权重，计算公式为

$$r = \text{softmax}(\boldsymbol{u}_i^c \boldsymbol{W}_{\text{attn}} \boldsymbol{P}^{\mathrm{T}}) \tag{2}$$

式中　$\boldsymbol{W}_{\text{attn}}$——注意力权重矩阵

将注意力权重和词向量相乘累加得到该字符的词汇向量 $\boldsymbol{h}_i^w$。计算公式为

$$\boldsymbol{h}_i^w = \sum_{j=1}^{m} r_{ij} \boldsymbol{p}_{ij}^w \tag{3}$$

将字符向量和加权词汇向量进行归一化，得到特征融合向量 $\boldsymbol{g}_i$ 为

$$\boldsymbol{g}_i = \lambda \boldsymbol{u}_i^c + (1 - \lambda)\boldsymbol{h}_i^w \tag{4}$$

式中　λ——权重

最后，将 $\boldsymbol{g}_i$ 输入到剩余的 Transformer 层，得到包含丰富语义特征的向量作为字词融合层的输出。

2.3　CRF 层

通过字词融合层得到的特征向量是相互独立的，直接将其输入到全连接层中，对每个字的标签进行预测时，无法考虑标签之间的依赖关系，为了解决这一问题，本文使用 CRF 模型来获取全局最优标签序列。CRF 的约束主要包括以下 3 点：①语料中第 1 个词的标签应该表示为“B－”或“O”。②首个标签是以“B－”开始的，而不是“I－”。例如“O I－label”是错误的。③“B－label”为标签的开头，则后面的标签为“I－label”，且 label 为相同的类型。例如“B－NAME－T I－NAME－T”是合理的标签序列，而“B－NAME－T I－DIS－T”是不合理的标签序列。对于给定的观测序列 x，输出标签序列 $y = (y_1, y_2, \cdots, y_n)$ 的评估分数计算公式为

$$s(x, y) = \sum_{i=0}^{n} A_{y_i, y_{i+1}} + \sum_{i=1}^{n} P_{i, y_i} \tag{5}$$

式中　$A_{y_i, y_{i+1}}$——标签转移概率

P_{i, y_i}——第 i 个字输出标签为 y_i 的概率

在得到序列 y 的评分后，标签序列 y 对应的预测概率表示为

$$P(y|x) = \frac{\mathrm{e}^{s(x,y)}}{\sum_{\tilde{y} \in Y_x} \mathrm{e}^{s(x,\tilde{y})}} \tag{6}$$

式中 $\hat{y}$——真实的标注序列

Y_x——所有可能的标签序列集合

最后,利用 Viterbi 动态规划算法将得分最高的序列作为最优序列,表示为 $L=\{l_1,l_2,\cdots,l_n\}$。

2.4 三元组抽取层

在玉米育种知识图谱中,将所有的节点集合表示为 $E=\{e_1,e_2,\cdots,e_n\}$,所有的关系集合表示为 $R=\{r_1,r_2,\cdots,r_n\}$。本文提出一种玉米育种实体关系三元组匹配算法 ERTM,对 CRF 层输出的最优序列 $L=\{l_1,l_2,\cdots,l_n\}$ 进行推理和预测,以此来获得玉米育种实体关系三元组集 $T=\{(h,r,t)\mid(h,t\in E,r\in R)\}$,解决玉米育种数据中的重叠三元组抽取问题。ERTM 算法描述如下:

(1)通过最优序列 L 得到完整实体,并得到其对应的关系类型和位置信息。

L 中的每个标签 l_i 都包含实体边界、关系类别和实体位置信息。遍历每个标签 l_i,通过实体边界标签得到完整实体 e_i,如果实体边界为“B”,则该标签对应的文本为实体中的第 1 个字符,如果实体边界为“I”,则该标签对应的文本为实体的中间和结尾字符,如果实体边界为“O”,则该标签对应的文本为非实体字符;通过关系类别标签得到关系 r_i;通过位置标签得到 e_i 在三元组中的位置。

(2)若实体的关系类别标签为“OVE”,且实体位置标签为“H”时,将其作为头实体,向后查找位置标签为“T”的实体,依次与其进行匹配形成三元组。

(3)若实体的关系类别标签为非“OVE”,且实体位置标签为“T”时,将其作为尾实体,向前查找位置标签为“H”的实体,并进行匹配形成三元组。

(4)关系类别标签为“OVE”的实体持续参与匹配,直到遇到下一个关系类别标签为“OVE”的实体时,该条数据实例的三元组抽取任务结束。

(5)关系类别标签为非“OVE”的实体匹配完成后不参与后续匹配。

通过上述步骤得到玉米育种知识图谱的三元组集合 T,对所有的实体 e_i 进行整合得到节点集合 E,将所有的关系 r_i 进行整合得到关系集合 R。

ERTM 算法的流程为

```
Input:标签序列 L
Onput:三元组集 T,实体集 E
获取实体边界、关系类别和实体位置信息
i←0
list_E[ ]
While i < len(l) do        //遍历所有数据
    list_T[h][r][t]        //初始化当前重叠实体三元组
    If getOve(l) not in list_T[h]  //判断是否为新的重叠实体
        h←h + 1
        r←0
        t←0
    r′,t′ = getRT (l)      //得到关系和尾实体
    list_E. add(h,t′)      //得到不同类别实体
    list_T[h][r][t]←Triple(r′,t′)  //构造三元组
    r←r + len(r′)
    t←t + len(t′)
End while
Return list_T[h][r][t];list_E[ ]
```

3 实验

3.1 数据集

本文采用 Scrapy 爬虫框架,从中国种业大数据平台、种业商务网、中国玉米品种系谱数据库等网站爬取相关数据,并且从玉米育种纸质版书籍中收集相关数据。对爬取的数据进行清洗,过滤掉数据中存在的无关文本以及对可标注的文本进行筛选。本文梳理了玉米育种领域知识,针对数据特点和领域专家指导将需要进行知识抽取的实体类别定义为 9 类,包括品种、天数、农艺性状、品质、病害、虫害、非生物逆境、密度、地区。通过对实体之间的关系梳理,得到玉米育种实体之间的 8 种关系,关系类别信息如表 1 所示。

表 1 关系类别及实例

Tab.1 Categories and examples of relation

关系类别	示例
GrowthPeriod	陕单 625→生育期 117 d
AgronomicCharacter	农科大 8 号→株型紧凑
Quality	陕单 653→籽粒容重 797 g/L
DiseaseResistance	榆玉 9 号→瘤黑粉病
PestResistance	内单 901→玉米螟
StressResistance	A3678→抗倒伏
PlantingDensity	诚信 707→亩留苗 4 500 株
PlantingArea	瑞禾白糯 80→北京

本文共标注 2 120 篇玉米育种数据,将这些数据以比例 7∶2∶1 分为训练集、测试集和验证集。数据规模如表 2 和表 3 所示。

表 2 实验数据规模

Tab.2 Experimental data scale

项目	品种数	总字符数	句子数
训练集	1 484	546 524	13 306
测试集	424	190 885	4 325
验证集	212	93 408	2 029

表 3　不同关系类别数据规模

Tab. 3　Different types of relation data scale

关系类别	训练集			测试集			验证集		
	品种	句子	关系	品种	句子	关系	品种	句子	关系
GrowthPeriod	1 047	1 225	1 412	296	322	447	134	201	208
AgronomicCharacter	1 132	1 452	4 224	346	388	965	176	225	665
Quality	987	1 128	4 700	264	356	1 584	160	206	802
DiseaseResistance	1 189	1 425	7 290	357	401	2 486	180	213	1 283
PestResistance	323	455	657	122	153	247	75	82	92
StressResistance	327	476	550	154	186	571	66	73	83
PlantingDensity	1 022	1 332	1 471	268	300	362	98	121	215
PlantingArea	1 058	1 293	2 480	287	312	668	101	132	460

3.2　实验环境及模型参数

本实验的软硬件环境：操作系统 Ubuntu 16.04，GPU NVIDIA GeForce RTX 3090，CUDA 11.1，使用 Python 3.8 作为开发语言，使用 Pytorch 1.8.1 完成模型的构建和训练。本文预训练语言模型采用 BERT-base 版本，包括 12 个 Transformer 层，768 维隐藏层和 12 头多头注意力机制，批尺寸设置为 32，学习率设置为 0.001，最大序列长度设置为 150，Dropout 设置为 0.5。

3.3　评价指标

为了验证本文模型的效果，使用精确率 P(Precision)、召回率 R(Recall)以及 F1 值(F-measure)3 个指标对模型进行评估。

3.4　结果分析

3.4.1　不同模型抽取性能对比

为了验证本文模型在玉米育种语料上的实体关系抽取性能，分别使用 BiLSTM－CRF[28]、WC－LSTM[25]、LR－CNN[29]、LGN[30]和 BERT－CRF[31]模型与本文模型进行对比实验，实验结果如表 4 所示。

表 4　不同模型抽取性能对比

Tab. 4　Comparison of extraction performance between different models　%

模型	P	R	F1 值
BiLSTM－CRF	88.72	90.14	89.42
WC－LSTM	89.78	91.42	90.60
LR－CNN	88.93	92.55	90.70
LGN	89.91	92.82	91.34
BERT－CRF	90.94	94.04	92.47
本文模型	91.84	95.84	93.80

从表 4 中可以看出，BiLSTM－CRF 模型的 F1 值只有 89.42%，识别效果相对较差，这是因为 BiLSTM－CRF 使用随机初始化的字向量作为嵌入层，包含的语义信息有限。WC－LSTM 模型利用 4 种不同的策略将单词信息编码为固定的矢量，将字和词级别的特征进行融合，F1 值为 90.60%，比 BiLSTM－CRF 模型提高 1.18 个百分点。LR－CNN 模型用 CNN 对字符特征编码，并且引入了反思机制使模型看到全局的信息，把高层特征作为输入，使用注意力调整每一层的词汇特征，F1 值为 90.7%，识别效果略优于 WC－LSTM。LGN 模型把词汇信息集成到图神经网络中，采用图结构聚合局部信息，并且融入了全局信息，F1 值达到 91.34%，相比 BiLSTM－CRF 模型提高 1.92 个百分点，比 LR－CNN 模型提高 0.64 个百分点。验证了嵌入词汇信息结合先验知识和边界信息的有效性。

BERT－CRF 模型的 F1 值为 92.47%，相比于 BiLSTM－CRF 模型，F1 值提高 3.05 个百分点，说明 BERT 模型能充分学习到字符特征以及上下文特征，可以获取更丰富的语义信息。

与上述传统的模型相比，本文构建的嵌入玉米育种领域词汇信息的 BERT－CRF 模型，将词汇信息嵌入 BERT 层，可以更加充分地利用词汇特征，并且使用 CRF 得到全局最优标签序列。实验结果表明，本文所构建的模型在玉米育种领域识别效果优于其他模型，精确率、召回率和 F1 值分别为 91.84%、95.84% 和 93.80%。相较于 BiLSTM－CRF 和 BERT－CRF 模型，F1 值分别提高 4.38、1.33 个百分点。

3.4.2　模型对各类实体间关系的预测结果分析

本文模型在玉米育种领域文本中各类关系的精确率、召回率和 F1 值如表 5 所示。从表 5 中可以看出，整体效果较好，在“PlantingDensity”关系的识别结果相对欠佳，F1 值为 71.07%，通过分析这类关系对应的语料文本，发现其灵活多样，包括“种植密度每亩”、“亩留苗”、“公顷保苗”、“种植密度中等水肥地”等多种描述，且文本较长。所以在预测过程中会出现预测错误的情况。

3.4.3　消融实验

为了评估本文模型每部分对整体模型的影响，

表 5 关系类型预测结果

Tab. 5 Prediction results of relation type %

关系	P	R	F1 值
OVE	97.51	97.95	97.73
GrowthPeriod	87.69	92.60	90.08
AgronomicCharacter	98.26	99.38	98.82
Quality	95.11	98.02	96.54
DiseaseResistance	95.17	98.22	96.67
PestResistance	81.72	91.63	86.39
StressResistance	77.28	85.34	81.11
PlantingDensity	67.93	74.52	71.07
PlantingArea	92.88	99.55	96.10

对添加词汇信息和 CRF 层进行消融实验,实验结果如表 6 所示。

表 6 消融实验结果

Tab. 6 Ablation experiment result %

模型	P	R	F1 值
本文模型	91.84	95.84	93.80
无嵌入词汇信息	90.94	94.04	92.47
无 CRF 层	91.24	94.46	92.93

从表 6 中可以看出,通过将玉米育种词汇信息嵌入 BERT 层能明显提升模型性能,F1 值提高 1.33 个百分点。在实验中相比较单纯使用 BERT,增加了 CRF 层后 F1 值提升 0.87 个百分点,说明 CRF 层加入约束是有必要的,减少了错误的预测序列。因此添加词汇信息和 CRF 层有利于模型性能的提升。

3.4.4 词典规模对性能的影响

为了验证玉米育种知识词典规模对模型的影响,从词典中随机抽取 60%、70%、80% 和 90% 的词汇,构造 4 个大小不同的词典,并进行对比实验,实验结果如表 7 所示。

表 7 词典规模对性能的影响

Tab. 7 Impact of dictionary size on performance %

词典抽取比例/%	P	R	F1 值
60	91.12	94.51	92.78
70	91.33	94.44	92.86
80	91.32	94.99	93.12
90	91.66	95.61	93.59
100	91.84	95.84	93.80

从表 7 中可以看出,随着玉米育种知识词典规模的增加,模型的性能也随之提升,进一步说明了嵌入玉米育种领域词汇的有效性。

3.4.5 模型参数对性能的影响

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(batchsize)和学习率 l_r 2 个参数的选择进行了实验,通过 P、R 和 F1 值来分析对模型的影响,以此来选出最优的参数组合。

图 5 为本文模型在不同批尺寸下的实验结果。从图 5 可以看出,当批尺寸为 32 时,模型 F1 值达到最高。因此设置本文模型的批尺寸为 32。

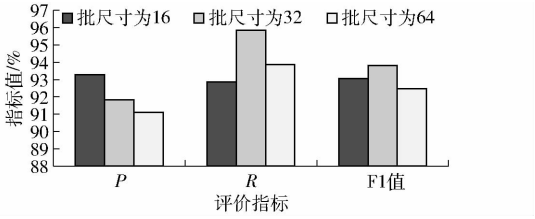

图 5 批尺寸对模型的影响

Fig. 5 Impact of batchsize on model

图 6 为本文模型在批尺寸为 32 时,学习率分别为 0.01、0.001 和 0.0001 时的结果。从图 6 中可以看出,当学习率为 0.001 时,F1 值达到最高。因此设置本文模型学习率为 0.001。

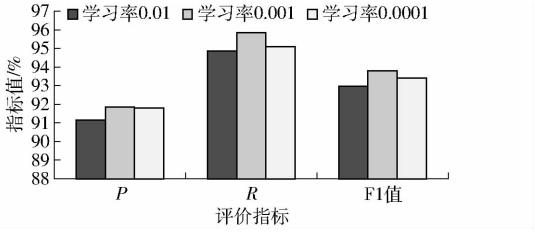

图 6 学习率对模型的影响

Fig. 6 Impact of learning rate on model

3.4.6 三元组抽取实验

为了测试玉米育种实体关系三元组匹配算法 ERTM 提取三元组的能力,通过随机抽取共包含 500 个玉米育种三元组的文本进行验证,评估其准确率。各类三元组抽取准确率如表 8 所示。

表 8 三元组抽取结果

Tab. 8 Triplet extraction result %

三元组类型	准确率
品种—GrowthPeriod—生育期	98.60
品种—AgronomicCharacter—农艺性状	92.52
品种—Quality—品质	90.26
品种—DiseaseResistance—病害	93.45
品种—PestResistance—虫害	95.44
品种—StressResistance—非生物逆境	98.21
品种—PlantingDensity—密度	99.23
品种—PlantingArea—地区	99.38
平均值	95.88

从表 8 中可以看出,整体结果较为均衡,玉米育种三元组数据准确率达到 95.88%,证明了 ERTM 可有效抽取玉米育种领域中的三元组知识,其中,“品种—AgronomicCharacter—农艺性状”、“品种—Quality—品质”和“品种—DiseaseResistance—病害”三类三元组的准确率偏低,通过对语料进行分析,

发现少量文本在一条语料中出现多个玉米品种实体时，会导致实体匹配时容易混淆，从而出现评估指标偏低的情况。例如“富友 987，春播出苗至成熟 128 天，比对照沈玉 21 早 2 天，幼苗叶鞘紫色，叶缘紫色，……。经鉴定抗大斑病，……。”，此条语料围绕玉米品种实体“富友 987”展开描述，当“大斑病”实体在向前与玉米品种实体进行匹配时，匹配正确的三元组应为（富友 987，DiseaseResistance，大斑病），但是由于存在玉米品种实体“沈玉 21”的干扰，得到匹配错误的三元组（沈玉 21，DiseaseResistance，大斑病）。

4　结论

（1）提出了一种面向玉米育种领域的实体关系联合抽取方法。针对玉米育种文本数据中存在大量重叠三元组的问题，设计了对实体边界、关系类别和实体位置信息同步标注策略，将实体关系联合抽取任务转换为序列标注任务。相比于传统的流水线方法，该方法能够有效提高标注效率，为其他领域实体关系抽取任务提供思路。

（2）构建了嵌入词汇信息的 BERT－CRF 玉米育种实体关系抽取模型，该模型将字符特征和词汇特征在 BERT 中进行融合，增强模型的语义能力，利用 CRF 解码预测最佳序列，通过三元组匹配算法 ERTM 获取玉米育种实体关系三元组。实验结果表明，本文模型精确率达到 91.84%，召回率达到 95.84%，F1 值达到 93.80%，相较于 BiLSTM－CRF 和 BERT－CRF 模型，F1 值分别提高 4.38 个百分点和 1.33 个百分点，此外，随着玉米育种知识词典规模的增加，模型性能也随之提升，验证了嵌入玉米育种领域词汇可以提升模型的抽取准确率。玉米育种三元组抽取准确率达到 95.88%，验证了 ERTM 算法在玉米育种领域抽取重叠三元组的有效性与实用性。

参　考　文　献

[1] 刘志铭，张晓龙，兰进好，等. 1979—2020 年我国玉米品种审定情况回顾与展望[J]. 玉米科学，2021，29(2)：1－7，15.
LIU Zhiming，ZHANG Xiaolong，LAN Jinhao，et al. Review and prospect of approved maize varieties in China from 1979 to 2020[J]. Journal of Maize Sciences，2021，29(2)：1－7，15.（in Chinese）

[2] 李建新，席蒙慧，张嘉玮，等. 中国玉米品种及其亲本系谱数据库的创建与利用[J]. 中国农业科学，2020，53(16)：3404－3411.
LI Jianxin，XI Menghui，ZHANG Jiawei，et al. Construction and utilization of database for Chinese maize varieties and their genealogy[J]. Scientia Agricultura Sinica，2020，53(16)：3404－3411.（in Chinese）

[3] 江旭晖，沈英汉，李子健，等. 社交知识图谱研究综述[J]. 计算机学报，2023，46(2)：304－330.
JIANG Xuhui，SHEN Yinghan，LI Zijian，et al. A survey of social knowledge graph[J]. Chinese Journal of Computers，2023，46(2)：304－330.（in Chinese）

[4] 张吉祥，张祥森，武长旭，等. 知识图谱构建技术综述[J]. 计算机工程，2022，48(3)：23－37.
ZHANG Jixiang，ZHANG Xiangsen，WU Changxu，et al. Survey of knowledge graph construction techniques[J]. Computer Engineering，2022，48(3)：23－37.（in Chinese）

[5] 陈明，朱珏樟，席晓桃. 基于知识图谱的花卉病虫害知识管理方法[J]. 农业机械学报，2023，54(3)：291－300.
CHEN Ming，ZHU Juezhang，XI Xiaotao. Knowledge management method of flower diseases and pests based on knowledge graph[J]. Transactions of the Chinese Society for Agricultural Machinery，2023，54(3)：291－300.（in Chinese）

[6] 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）

[7] 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）

[8] LI J，SUN A，HAN J，et al. A survey on deep learning for named entity recognition[J]. IEEE Transactions on Knowledge and Data Engineering，2020，34(1)：50－70.

[9] 袁培森，李润隆，王翀，等. 基于 BERT 的水稻表型知识图谱实体关系抽取研究[J]. 农业机械学报，2021，52(5)：151－158.
YUAN Peisen，LI Runlong，WANG Chong，et al. Entity relationship extraction from rice phenotype knowledge graph based on BERT[J]. Transactions of the Chinese Society for Agricultural Machinery，2021，52(5)：151－158.（in Chinese）

[10] CHAN Y S，ROTH D. Exploiting syntactico-semantic structures for relation extraction[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics：Human Language Technologies，2011：551－560.

[11] 黄梅根，刘佳乐，刘川. 基于 BERT 的中文多关系抽取方法研究[J]. 计算机工程与应用，2021，57(21)：234－240.
HUANG Meigen，LIU Jiale，LIU Chuan. Research on improved BERT's Chinese multi-relation extraction method[J]. Computer Engineering and Applications，2021，57(21)：234－240.（in Chinese）

[12] 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.
[13] 李冬梅, 张扬, 李东远,等. 实体关系抽取方法研究综述[J]. 计算机研究与发展, 2020, 57(7): 1424 - 1448.
LI Dongmei, ZHANG Yang, LI Dongyuan, et al. Review of entity relation extraction methods[J]. Journal of Computer Research and Development, 2020, 57(7): 1424 - 1448. (in Chinese)
[14] TAN Z, ZHAO X, WANG W, et al. Jointly extracting multiple triplets with multilayer translation constraints[C]// Proceedings of the AAAI Conference on Artificial Intelligence,2019, 33(1): 7080 - 7087.
[15] FU T J, LI P H, MA W Y. 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: 1409 - 1418.
[16] SUN C, GONG Y, WU Y, et al. Joint type inference on entities and relations via graph convolutional networks[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019: 1361 - 1370.
[17] NAYAK T, NG H T. Effective modeling of encoder-decoder architecture for joint entity and relation extraction[C]// Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 8528 - 8535.
[18] YUAN Y, ZHOU X, PAN S, et al. A relation-specific attention network for joint entity and relation extraction[C]// IJCAI, 2020: 4054 - 4060.
[19] ZHENG S, WANG F, BAO H, et al. Joint extraction of entities and relations based on a novel tagging scheme[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, 2017: 1227 - 1236.
[20] 冯钧, 张涛, 杭婷婷. 重叠实体关系抽取综述[J]. 计算机工程与应用, 2022, 58(1): 1 - 11.
FENG Jun, ZHANG Tao, HANG Tingting. Survey of overlapping entities and relations extraction[J]. Computer Engineering and Applications, 2022, 58(1): 1 - 11. (in Chinese)
[21] 曹明宇, 杨志豪, 罗凌, 等. 基于神经网络的药物实体与关系联合抽取[J]. 计算机研究与发展, 2019, 56(7): 1432 - 1440.
CAO Mingyu, YANG Zhihao, LUO Ling, et al. Joint drug entities and relations extraction based on neural networks[J]. Journal of Computer Research and Development, 2019, 56(7): 1432 - 1440. (in Chinese)
[22] 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)
[23] 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)
[24] 赵鹏飞, 赵春江, 吴华瑞, 等. 基于 BERT 的多特征融合农业命名实体识别[J]. 农业工程学报, 2022, 38(3): 112 - 118.
ZHAO Pengfei, ZHAO Chunjiang, WU Huarui, et al. Recognition of the agricultural named entities with multi-feature fusion based on BERT[J]. Transactions of the CASE, 2022, 38(3): 112 - 118. (in Chinese)
[25] LIU W, XU T, XU Q, et al. An encoding strategy based word-character[C]// Proceedings of the 2019 Conference of the North,2019.
[26] 徐懿彬. 基于 Aho - Corasick 自动机算法的概率模型中文分词 CPACA 算法[J]. 电子科技大学学报, 2017, 46(2): 426 - 433.
XU Yibin. A probability model Chinese word segmentation algorithm based on Aho - Corasick automata algorithm[J]. Journal of University of Electronic Science and Technology of China, 2017, 46(2): 426 - 433. (in Chinese)
[27] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4171 - 4186.
[28] HUANG Z, XU W, YU K. Bidirectional LSTM - CRF models for sequence tagging[J]. arXiv Preprint, arXiv: 1508.01991, 2015.
[29] GUI T, MA R, ZHANG Q, et al. CNN - Based Chinese NER with lexicon rethinking[C]// IJCAI,2019: 4982 - 4988.
[30] GUI T, ZOU Y, ZHANG Q, et al. A lexicon-based graph neural network for Chinese NER[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 1040 - 1050.
[31] HU S, ZHANG H, HU X, et al. Chinese named entity recognition based on BERT - CRF model[C]// 2022 IEEE/ACIS 22nd International Conference on Computer and Information Science, 2022: 105 - 108.