农业机械学报
Transactions of the Chinese Society for Agricultural Machinery
ISSN 1000-1298,CN 11-1964/S

《农业机械学报》网络首发论文

题目：基于 BERT 的水稻表型知识图谱中关系抽取研究
作者：袁培森，李润隆，王翀，徐焕良
收稿日期：2020-06-12
网络首发日期：2021-03-16
引用格式：袁培森，李润隆，王翀，徐焕良．基于 BERT 的水稻表型知识图谱中关系抽取研究．农业机械学报.
https://kns.cnki.net/kcms/detail/11.1964.S.20210315.1809.015.html

网络首发时间：2021-03-16 11:32:14
网络首发地址：https://kns.cnki.net/kcms/detail/11.1964.S.20210315.1809.015.html

基于 BERT 的水稻表型知识图谱中关系抽取研究

袁培森[1] 李润隆[1] 王翀[2] 徐焕良[1]
(1. 南京农业大学人 工智能学院，南京 210095；
2.国网江苏省电力有限公司信息通信分公司，南京 210024)

摘 要：水稻表型组学通过对生物的遗传信息以及内外各种表型数据进行分析和研究，对水稻的生产以及研究有着重要的指导作用。知识图谱技术通过结构化描述数据中的概念、实体和关系等信息，已经在知识存储、搜索引擎等方面获得了广泛应用。关系抽取任务作为知识图谱中的关键任务和环节，可以抽取文本中的两个实体词之间的联系。针对水稻表型知识图谱中的实体关系抽取问题，本文首先对水稻表型组学数据进行获取、标注和分类，根据植物本体论方法提出了一种对水稻的基因、环境、表型等表型组学实体进行关系分类的方法。随后提取关系数据集中的词向量、位置向量及句子向量，基于 BERT 实现水稻表型组学关系抽取模型。最后，将 BERT 模型与卷积神经网络以及分段卷积网络模型进行结果比较，在 3 种关系抽取模型对比中，BERT 获得了更好的表现，精确率达到了 95.10%、F1 值为 95.85%。

关键词：水稻表型；知识图谱；关系抽取；双向转换编码表示
中图分类号：TP391　　**文献标志码：**A　　OSID：

Relationship Extraction from Rice Phenotype Knowledge Graph Based on BERT

YUAN Peisen[1]　LI Runlong[1]　WANG Chong[2]　XU Huanliang[1]
(1.*College of Artificial Intelligence, Nanjing Agricultural University, Nanjing 210095, China*；
2. *Information and Communication Branch of State Grid Jiangsu Electric Power Co., Ltd.*，*Nanjing* 210024, *China*)

Abstract: Rice phenotype has an important guiding role in rice research by analyzing genetic information of various phenotype data. Knowledge graph technology has been widely used in knowledge storage and search engines by structurally describing the information, concepts, entities and relationships in data. As a key task in the knowledge graph, the relation extraction task can extract the connection between two entity words in the text. Within this research, rice phenotypic data was collected from the National Rice Data Center, and the data were preprocessed and annotates. The rice phenotype relationship was proposed based on the plant ontology, then method of bidirectional encoder representation from transformers is applied for classifying relation between rice genomics, environment, and phenotype data based on plant ontology. Then the word vector, position vector and sentence vector were extracted in the relation dataset, and rice phenotype relation extraction model was realized based on BERT. Finally, the results of BERT model was compared with the convolutional neural network and the piecewise convolutional network model. In the comparison of the three relationship extraction models, BERT achieved the best performance, and reached an accuracy of 95.10% and F1 value of 95.85%. Deep learning methods were used to improve the performance of relation extraction of knowledge graphs, and provide technical support for the efficient construction of a rice phenotype knowledge graph system.

Key words: rice phenotype; knowledge graph; relationship extraction; bidirectional encoder representation

收稿日期：2020-06-12　修回日期：2020-12-02
基金项目：国家自然科学基金项目(61502236、61806097)和大学生创新创业训练专项计划项目(S20190025)
作者简介：袁培森（1980－），男，讲师，博士，主要从事智能信息处理和表型数据分析研究，E-mail：peiseny@njau.edu.cn
通信作者：徐焕良（1963－），男，教授，博士生导师，主要从事农业信息化与大数据技术研究，Email：huanliangxu@njau.edu.cn.

from transformers

0 引言

植物表型数据和分析研究是植物学领域和信息科学领域近年来研究的一个交叉热点，其本质是对于植物的基因数据的三维时序表达结果，以及其地域分布特征和代际演进规律[1]。表型组学指利用生物的遗传基因组信息来对于生物的外部以及内部的表型数据进行研究的一门具有综合性的学科[2]。植物表型组学不仅研究植物的外在形状，也研究其内部结构、物理和生化性质以及遗传信息。植物表型组学数据亟需研究对其建立完整知识库的智能计算方法[3]。

中国作为世界上水稻产量最大以及消费最多的水稻种植国[4]，水稻的培育以及研究也是中国粮食安全战略的重要内容[5]。水稻表型组学研究是植物生物学的研究热点，水稻表型数据的高通量以及其高维且海量的数据特征对于数据的快速检索和知识的有效提取提出了更高的技术要求[6]。

知识图谱将人的知识转化为图，利用计算机进行推理分析，实现从感知智能到认知智能的飞跃，是人工智能领域的一项重要技术[7]。知识图谱是一个具有结构化特征的语义知识库，使用符号的形式来描述数据中的实体以及之间的关系[8]，它利用对于语义的抽取和分析，结合了数据科学、人工智能等学科的前沿技术和方法，在学科知识库的构建领域获得了研究人员的广泛关注。

知识图谱系统的构建包括 2 个核心步骤：实体抽取；实体间关系的构建，其中实体关系的构建需要关系的抽取技术。关系抽取任务的研究目标是自动地对于两个实体和之间的联系所构成的 3 元组进行关系识别[9]。关系抽取能够将文本数据中的特征进行提取，提升到更高的层面[10]。

实体关系的抽取方法可以分为 3 类：基于模板、基于传统机器学习以及基于深度学习的方法[11]。基于模板的关系抽取方法是早期基于语料学知识以及语料的特点，由相应领域的专家和研究人员进行手工编写的模板，这种方法需要消耗大量专业人力，可移植性较差。基于传统机器学习的关系抽取方法主要包括使用核函数[12]、逻辑回归[13]以及条件随机场[14]等，是一种依赖特征工程的方法。HASEGAWA 等[15]使用聚类方法来计算上下文的相似性。赵明等[16]采用本体学习，使用有监督的基于依存句法分析的词汇-语法模式来对于百度百科植物语料库进行关系抽取，在非分类的关系抽取任务中表现较好，为构建植物领域知识图谱奠定了基础。

随着深度学习模型快速发展，深度学习在实体关系中的应用研究发展迅速。基于深度学习的关系抽取方法包括递归神经网络[17]、卷积神经网络模型[18]、双向转换编码表示模型（Bidirectional encoder representation from transformers, BERT）[19]等模型。深度学习能够实现语义特征的自动提取，从而使得模型能够对不同抽象层次上的语义进行分析[20]。BERT[19]为典型的深度学习模型，通过自动学习句中特征信息，获取句子向量表示，能够在水稻表型组学关系抽取任务中得到应用。在水稻知识图谱构建中，对水稻表型组学实体之间的复杂关系进行区分关系到水稻表型组学知识库的构建。因此，水稻表型组学的关系抽取研究十分重要。

本文使用爬虫框架获取水稻表型组学数据，根据植物本体论提出一种对水稻的基因、环境、表型等表型组学数据进行关系分类的方法。使用词向量、位置向量等算法对句中的特征进行提取，在获取的水稻表型组学实体关系数据集上构建基于双向转换编码表示 BERT 的关系抽取模型，实现了句子级别的关系抽取任务。最后将本文方法与卷积神经网络(Convolutional neural network ， CNN) [21]和分段卷积神经网络模型(Piece wise CNN，PCNN) [18]进行对比。

1 水稻表型组学关系数据集的获取

1.1 关系数据获取

本文关系数据集主要来自国家水稻数据中心（http://www.ricedata.cn/）以及维基数据中文语料库。数据获使用可对网页的结构性数据进行获取以及保存的框架 Scrapy[22]，实现水稻数据中心本体系统以及维基关系数据集的爬取。对爬取的水稻表型数据进行清洗处理，获得了用于关系分类处理的水稻表型组学关系数据集，数据集详情如表 1 所示。

表 1 数据集来源分布

Tab.1 Distribution of dataset origin

来源	数量	数据格式
水稻表型本体系统	2812	术语名称+术语 ID+术语定义+父术语+链接
维基百科语料库	3135	实体 1+实体 2+关系+句子

1. 2 关系数据分类

在水稻表型组学关系数据的分类问题上，本文参照了植物本体论(Plant ontology) [23]对植物表型组学的分类，通过关系分类将水稻的解剖结构，形态、

生长发育与植物基因数据联系起来，从而对于水稻表型组学数据进行分类。

本体[24]指的是在某一领域内的实体与其相互间关系的形式化表达，本体论是概念化的详细说明，它的核心作用是定义某一个领域内的专业词汇以及他们之间的关系[25]。

植物本体论[23]是一种结构化的数据库资源，是用来描述植物解剖学、形态学等植物学的结构性术语集合，它将植物的内部解剖结构、外表形态结构等表型组学数据与植物基因组学数据联系起来，使用关系来描述基因、环境、表型之间的联系。如今植物本体论的描述范围从最开始的水稻单个物种扩大到了 22 种植物，对这些植物的基因或基因模型、蛋白质、RNA、种质等表型和基因数据进行描述。本文依据其分类规则，将水稻表型组学数据分为 7 类：①is a，用来表示父术语以及子术语之间的关系，表示对象 O_1 是 O_2 的子类型或亚型。②has part，用来表示对象 O_1 的每个实例都有一部分 O_2 的实例。③has a morphology trait，表示 O_1 通过 O_2 的形态特征表现出来。④develop from ，表示 O_1 从 O_2 发育而来，O_2 的世系可以追溯到 O_1。⑤participate，表示实体 O_1 的每个实例都参与开发 O_2 的某些实例。⑥regulate，O_1 对 O_2 有调节或调控作用。⑦other，表示其他关系。

分类完成后的关系抽取数据集示例如表 2 所示。表 2 中，ddu1（Dwarf and disproportionate uppermost-internode1）为使用甲基磺酸乙酯诱变粳稻品种兰胜而成的矮化突变体的品种名称；SPL5（Spotted leaf 5）为经 γ 射线辐射诱导粳稻品种 Norin 8 而成的水稻类病变突变体的品种名称；FLW1(Flag Leaf Width NAL1)的中文名为剑叶宽度基因。最后，将数据集按 8:2 分为训练集和测试集。

表 2 关系抽取数据集示例

Tab.2 Examples of relation extraction datasets

实体 1	实体 2	关系	句子
产量性状	性状	is a	产量性状表示和植物的可收获产物相关的性状
内保护皮层	果皮	has part	大米具有内保护皮层，包括了果皮、种皮以及珠心层
ddu1	矮化	has a morphology trait	突变体植株 ddu1 表现出明显的矮化
胚芽	花药	develop from	花药与子房中间的胚珠相互结合，发育形成胚芽
SPL5	RNA 的剪接	participate	SPL5 参与植物 RNA 的剪接负调控细胞死亡以及抗性应答
FLW1	剑叶宽度	regulate	剑叶宽度受到两个显性基因的控制，即 FLW1 以及 FLW2
稻穗	穗伸出度	other	穗伸出度是用来衡量稻穗从旗叶叶鞘中伸出长度的指标

1.3 关系数据存储

水稻实体及关系采用图方式来进行建模以及数据存储，本文中使用图数据库 Neo4j[26]存放实体和关系数据。Neo4j 的核心概念是节点和边，节点用来存储实体，使用圆形图例表示，边用来存储关系数据结构中的实体之间的关系，使用带箭头的线表示。不同实体以及关系的相互连接形成复杂的数据结构，实现对于某个实体进行关系的完整增删改查等功能。

对收集的数据集进行预处理，提取 2021 个实体和 2689 条关系，通过 Cypher 语言[27]可以进行快速的查询工作。图 1 为 Neo4j 数据库存储的水稻表型组学关系示例。由于实体名称较长，图 1 中的“12 号染…”为 12 号染色体；“等位基因…”为等位基因 STV11-S。

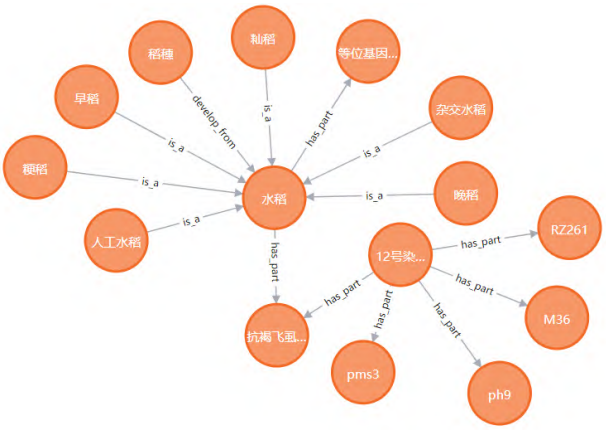

图 1 Neo4j 水稻表型关系示例

Fig.1 An example in Neo4j of rice phenotypic relationship

2 水稻表型组学关系抽取模型构建

2.1 向量化表示

本文中BERT关系抽取模型使用词向量、位置

向量以及句子向量相结合的输入向量序列，不仅能简单获取词语语义上的特征，而且能够对深层次语义进行表示和抽取。

2.1.1 词向量

本文使用 BERT 模型中的词嵌入方式来动态产生词向量，即将词转化为稠密的向量。通过这种词嵌入方式，该模型能够根据上下文预测中心词的方式来获得动态的语义特征，以解决传统词嵌入模型产生的多义词局限性，可以产生更精确的特征表示，从而提高模型性能。

BERT 的词向量生成方法如下：给定语句序列 $s = w_0, w_1, \cdots, w_n$。其中 w_0 =[CLS]， w_n =[SEP] 用来表示句子的开始以及结束。模型将原有的序列映射为具有固定长度的向量化来表示语义关系。

2.1.2 位置向量

设句子为 $s = w_0, w_1, \cdots, w_n$ ，实体为 i_1 与 i_2，则对于每一个单词 w_i，计算其与 i_1、i_2 的相对距离，即 $i - i_1$ 和 $i - i_2$，使得该句子可以根据两个实体生成两部分的位置向量，并且能体现距离和实体的关系。本文使用的位置向量维度为 50。

2.1.3 句子向量

句子向量按照句子的数目进行标记，对于第一条句子的每个单词添加 $\mathbf{v}_1$ 向量，给第二条句子中的每个单词添加一个 $\mathbf{v}_2$ 向量。

2.1.4 输入表示

BERT 模型的输入示例如图 2 所示。图 2 中的 BERT 模型输入的句子为“稻是谷类，原产地中国与印度”，模型生成每个词的词向量，根据每个词与实体之间的距离生成句向量，根据句子的条数生成对应的句向量，将此作为 BERT 模型的输入。

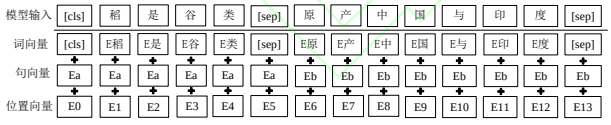

图 2 BERT 模型的输入表示示例

Fig. 2 Example of the input of BERT model

2.2 BERT 关系抽取模型构建

BERT 是以 Transformer 的编码器为基础的双向自注意力机制表示模型，能够对所有层基于上下文进行双向表示。BERT 模型使用双向的自注意力机制来进行构建，使用 Transformer 的编码器来进行编码，并且使用遮挡语言模型以及下一句预测两个方法来更有效地训练模型。

2.2.1 双向自注意力机制

BERT 使用双向自注意力机制[28]来进行构建。双向自注意力机制是注意力机制中的一种，注意力机制在自然语言处理领域的多个任务获得了实际应用。注意力机制可以描述为一个查询 Q 到相应键值对 $< K, V >$ 的一个映射的过程[29]，可描述为

$$A_t(Q, K, V) = S_f(S_m(Q, K)) \cdot V \tag{1}$$

式中 A_t ——注意力机制

S_f ——Softmax 函数

S_m ——相似度函数

注意力值的计算过程可分为 3 部分：①首先计算查询 Q 和每个键 K 之间的相似度 S ，获得权重，使用的相似度计算函数有点积、拼接以及感知机。②使用 Softmax 函数进行权重归一化。③将权重以及键值对中值 V 进行加权，获得最终的注意力的值。自注意力机制即检索自身的键值对进行加权处理， $Q = K = V$ ，将序列进行重新编码，获得更具整体性的特征序列[30]。自注意力机制的结构图如图 3 所示。

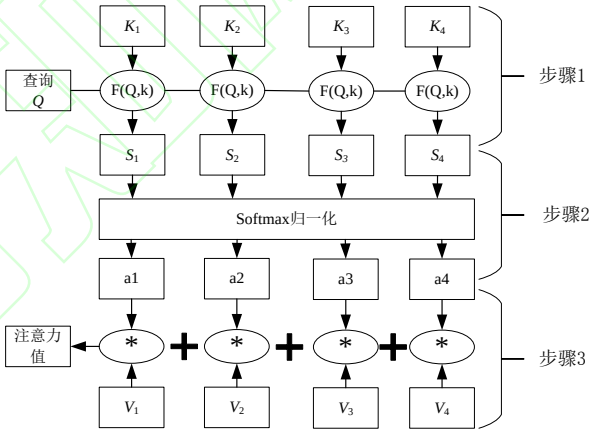

图 3 自注意力机制结构图

Fig. 3 Self-attention mechanism structure diagram

自注意力机制将输入序列通过向量映射的方式输入到嵌入层，注意力层进行查询向量和值向量的相似度计算，Softmax 层使用函数加权后将序列输出。BERT 所用的多头自注意力机制在输入到注意力层之前对查询 Q 、键 K 以及值 V 进行多次线性变换，线性变换的次数即为多头，多头自注意力机制可以获得多种序列的子特征，进而获得较长序列中的相隔较远的向量特征[31]。

2.2.2 Transformer 编码器

BERT 使用 Transformer 编码器进行编码，Transformer[32]通过对于语义信息以及位置信息的分析来完成自然语言处理任务，其框架为编码器加解码器结构。其中，编码器框架使用了层叠结构，每一层有 2 个部分：进行加权处理的多头注意力机制和进行前馈化网络的全连接层，在 2 个部分之间

使用残差进行连接然后进行标准化。解码器的层数与编码器相同，同时在每一层之内还添加了一个进行计算翻译效果的部分。Transformer 编码器结构图如图 4 所示，图中的 $N\times$ 表示编码器或解码器包含的层数：

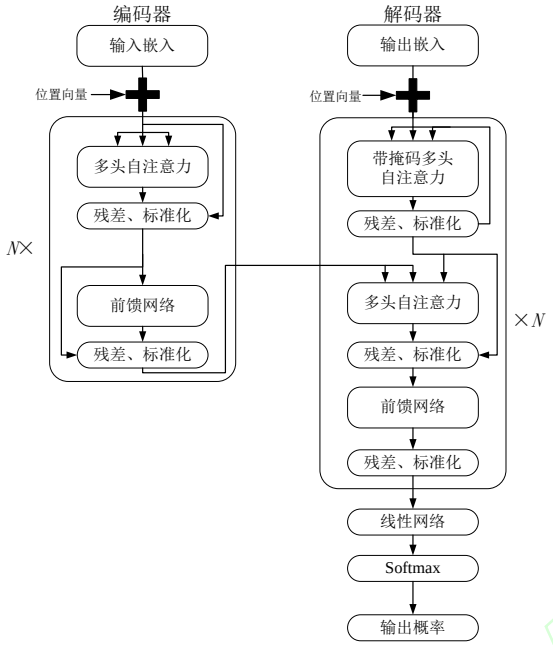

图 4Transformer 结构图

Fig. 4 Transformer structure diagram

Transformer 编码器部分由 3 部分组成：①先对输入句子进行向量化，将词嵌入到编码器中。②编码器接受向量序列，随后使用自注意力机制对序列进行处理，通过对序列中所有单词之间建立联系来进行序列编码，处理后的序列通过残差网络进行求和与归一化。③自注意力机制结束以后，输入到全连接的前馈网络中，输出标准化后的向量。

BERT 模型使用了多个 Transformer 编码器进行编码，编码器输出后进入到一个全连接层与激活函数构成的分类层并输出相应的概率[33]。图 5 是对水稻表型进行编码示例，输入的句子为“产量性状是与植物可收获产物相关的性状”。

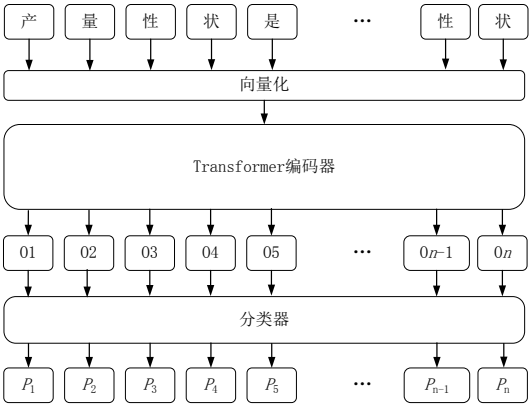

图 5BERT 模型结构图

Fig.5 BERT model structure diagram

BERT 模型在使用过程中，仅仅需要在编码器后面加上一层全连接层就能够完成关系抽取任务。在后期的微调部分中，设之前遮盖处理后的输出向量为 $\mathbf{C}$，使用 Softmax 分类器来完成关系分类的概率 P_r 的计算：

$$P_r = S_f(\mathbf{C}\mathbf{W}^T) \tag{2}$$

对于本文的关系多分类问题，类别标签 $y \in \{1, 2, .M$，有 M 个取值。给定测试样本 x，Softmax 函数预测属于类别 $c \in M$ 的条件概率为

$$p(y = c \mid x) = S_f(\mathbf{w}_c^T x) = \frac{\exp(\mathbf{w}_c^T x)}{\sum_{i=1}^{M} \exp(\mathbf{w}_i^T x)} \tag{3}$$

式中 $\mathbf{w}_c$ ——权重

$\mathbf{w}_i$ ——第 i 的类权重

p ——概率值

BERT 模型输出关系类别以及其对应的概率。另外，BERT 模型在预训练部分使用了遮挡语言模型以及下一句预测两个方式来训练模型。

2.2.3 遮挡语言模型

遮挡语言模型(Masked language model) [19]指的是在进行 BERT 模型训练时，由于进行的注意力机制是多头而不是单向的，如果按照 CNN 等模型的训练方式进行训练，则 BERT 模型的训练将成为一个先获得后文再进行预测的任务，无法正确获取语义特征，因此进行双向注意力机制训练时，BERT 使用了遮挡语言模型，将输入的词进行随机的遮盖，从而使得双向编码器能够真正的对于前后文进行预测[19]。本文对于 15%的词进行遮挡，并且遵循以下规律：①被遮挡的词 80%可能性被替换成屏蔽符号 $[mask]$。②10%的概率被换成随机的词。③10%的概率保持原有的单词不变。这样后期的微调部分的向量输入不会与遮盖处理中的向量差距太大。

2.2.4 下一句预测

下一句预测(Next sentence prediction) [19]使 BERT 模型能够学习下一句和上一句的内在联系，BERT 模型在数据集中随机选取句子 S_1，对于其下一句 S_2，有 50%的概率将 S_2 替换为无关的句子 S_3，以此来学习句子间的关系。

3 关系分类结果与分析

3.1 试验环境

本文选择 Intel Corei5-8250u 处理器@1.6GHz，8GB 内存，1TGB 硬盘，Windows 10 操作系统。

3.2 参数设置

BERT 模型的参数设置如表 3 所示。为防止模型训练后期的波动，学习率衰减采用了文献[34]中的推荐值，设置为 2×10^{-5}。

表 3 BERT 模型参数设置

Tab.3 BERT parameter settings

参数	数值
隐藏层大小	1536
学习率衰减	2×10^{-5}
最大序列长度	80
迭代次数	5
批尺寸	8
梯度下降算法	ADAMW

梯度下降算法(Gradient descent optimizer) [35]能够帮助模型进行目标函数的最大化或最小化计算，一个优秀的梯度下降算法能够减少损失函数的值。常用的梯度下降算法有随机梯度下降 (Stochastic gradient descent，SGD) [35]、自适应力矩估计 (Adaptive moment estimation，ADAM) [36]、解耦权重衰减的自适应矩估计 (Adaptive moment estimation with decoupled weight decay，ADAMW) [37]等，本文选择 ADAMW 算法。

3.3 数据集

根据植物本体论进行实体关系数据的分类，共获得 7 大类，2689 条关系数据，类型有： is a、has part 、 has a morphology trait 、 develop from 、participate、regulate、other。各个关系类型的数量及分布如表 4 所示。

表 4 水稻表型组学关系数据集的数量分布

Tab.4 Distribution of the relationship quantity of the rice phenotypic relationship in dataset

关系	数量
is a	1088
has part	456
has a morphology trait	152
develop from	127
participate	160
regulate	235
other	471

3.4 算法性能评估指标

使用精确率(Precision，P)、召回率(Recall，R)、F1 值作为评价指标，将 BERT 与传统的卷积神经网络模型[21]与分段卷积神经网络模型[18]进行对比。

3.5 BERT 模型关系分类结果

本部分对梯度下降算法[35]、批尺寸[38]2 个参数和表 2 中的不同关系 3 个方面进行了试验分析测试。

3.5.1 梯度下降算法

对于 BERT 关系抽取模型，本文进行了梯度下降算法的对比，选择批(Batch)尺寸为 8，3 种梯度下降算法在 BERT 模型上的结果如图 6 所示。

由图 6 可以看出，对比的精度、召回率和 F13 个指标，ADAMW 比 SGD 和 ADAM 高，SGD 最低，3 个指标在 0.6 左右。ADAM 和 ADAMW 都在 0.94 以上。

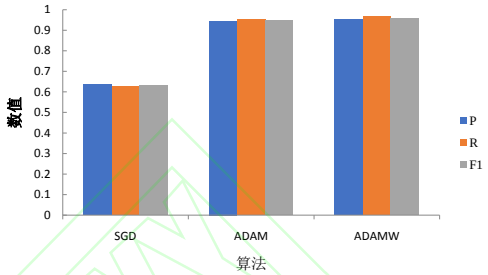

图 6 不同梯度下降算法在 BERT 模型上的对比

Fig.6 Comparison of gradient descent algorithm on BERT model

3.5.2 批尺寸

选择批尺寸分别为 8、16、32、64 来进行试验，选择 ADAMW 作为梯度下降算法，其在 BERT 模型上的结果如表 5 所示。

表 5 不同批尺寸在 BERT 上的对比

Tab.5 Comparison of batch on BERT

批尺寸	*P*%	*R*%	F1%
8	9511	9661	9585
16	9459	9502	9481
32	9448	9661	9563
64	9423	9502	9462

据表 5 可知，批尺寸为 8 时，使用 ADAMW 算法的精确度 *P* 达到了 0.9511，其召回率为 0.9661，F1 值为 0.9585。相比批大小为 16、32 和 64，精确度 *P* 分别提高 0.55%，0.67%和 0.94%；F1 值分别提高 1.10%，0.23%和 1.30%。

3.5.3 不同关系上的结果

本试验批尺寸为 8，BERT 模型使用 ADAMW 时对本文数据集上的不同关系抽取结果进行对比，结果如表 6 所示。

表 6BERT 模型对不同关系的处理结果

Tab.6 Results of different relations in the BERT relation extraction model

关系类型	*P*%	*R*%	F1%
develop from	9607	4364	6002
has a morphology trait	9667	4833	6444

has part	8614	8286	8447
is a	9857	8771	9283
participate	9667	4833	6445
regulate	9167	6111	7333
other	9131	7778	8399

由表 6 结果可知，BERT 模型对于不同关系的 F1 值都不低于 0.6002，但是对于不同关系的处理效果也有不同。其中，对于 has part、is a、other、regulate 关系分类效果较好，其 F1 值都不小于 0.7333，而对于 develop from、participate、has a morphology trait 的分类效果相对较差，其 F1 值均小于等于 0.6445。在 7 种关系中，is a 关系类型的测试结果最佳，其 F1 值达到了 0.9283，是表现最差的 develop from 类型的 1.5472 倍。develop from、has a morphology 和 participate 3 个关系分类效果较差的原因是这 3 个关系类别的数据库中关系数较少，且数据集中各个类别的分布不均衡。

其解决方法有：①可以通过增加这 3 个类别实体关系数据使 BERT 模型提取更多有效的语义和词汇特征。②可以将各个关系数据的条数进行调整，保持各个类别实体关系数据的数量均衡。

3.6 3 种模型对比

本文将 CNN[21]、PCNN[18]与 BERT 模型进行对比，CNN 和 PCNN 模型的参数设置如表 7 所示。

表 7 CNN 和 PCNN 模型参数设置

Tab.7 CNN and PCNN model parameter settings

参数	数值/方法
卷积核尺寸	3
隐藏层数	230
填充大小	1
丢弃率	0
词向量维度	100
权重衰减	1×10^{-5}
学习率衰减	0.1
卷积激活函数	ReLU
池化方法	最大池化
迭代次数	5

表 7 中的 ReLU [39]线性整流函数是一种常用的激活函数。

CNN[21]、PCNN 以及 BERT 模型上的测试结果如表 8 所示。

表 8 3 种关系抽取模型结果对比

Tab.8 Comparison of three relation extraction models

关系抽取模型	*P*%	*R*%	F1%
CNN	8179	8235	8207
PCNN	8595	8167	8366
BERT	9511	9661	9585

从表 8 试验结果可知：卷积神经网络 CNN 在批尺寸为 16 时，使用 SGD 算法时获得最高精确率、召回率与 F1 值，精确度为 81.79%，召回率为 82.35%，F1 值为 82.07%。分段卷积神经网络 PCNN 的批尺寸为 16，使用 SGD 算法时，获得最高精确率、召回率与 F1 值，精确度为 85.95%，召回率为 81.67%，F1 值为 83.66%。BERT 模型在：隐藏层大小为 1536，最大序列长度为 80，学习率衰减为 2×10^{-5}，训练轮数为 5，批为 8，梯度下降算法为 ADAMW 时，关系抽取的精确率、召回率与 F1 值达到最优，精确率为 95.11%，召回率为 96.61%，F1 值为 95.85%。

如表 8 所示，BERT 在精确率、召回率以及 F1 值上都明显高于其他两种模型。其 F1 值是 CNN 的 1.17 倍、PCNN 的 1.15 倍。

综上所述，在使用 BERT 模型进行水稻表型组学数据关系抽取时，BERT 模型能够根据上下文预测中心词的方式来获得动态的词向量，使用自注意力机制获得双向的语义特征，大幅度提高了关系抽取的质量。

4 结论

本研究基于植物本体论提出了基于水稻表型组学的关系分类的方法，将水稻表型的实体关系分为 7 类；使用词向量、位置向量以及句子向量进行句子特征抽取，构建 BERT 模型；将 BERT 模型与 CNN、PCNN 模型进行对比分析。试验结果表明，在参数满足上述一致条件下，对数据集的 BERT 模型，其精确率、召回率与 F1 的处理结果分别为 95.11%、 96.61%和 95.85%，达到了预期的分类效果。

参考文献

[1] 赵春江. 植物表型组学大数据及其研究进展[J]. 农业大数据学报, 2019, 1(2): 5-18.
ZHAO Chunjiang. Big data of plant phenomics and its research progress[J]. Journal of Agricultrual Big Data, 2019, 1(2): 5-18.（in Chinese）

[2] 周济，TARDIEU F,PRIDMORE T，等. 植物表型组学:发展、现状与挑战[J]. 南京农业大学学报, 2018, 41(4): 580-588.

ZHOU Ji, TARDIEU F,PRIDMORE T, et al. Plant phenomics: history, present status and challenges[J]. Journal of Nanjing Agricultural University, 2018, 41(4): 580-588（in Chinese）
[3] 潘映红. 论植物表型组和植物表型组学的概念与范畴[J]. 作物学报, 2015, 41(2): 175-186.
PAN Yinghong. Analysis of concepts and categories of plant phenome and phenomics [J]. Acta Agronomica Sinica, 2015, 41(2): 175-186（in Chinese）
[4] 凌霄霞，张作林，翟景秋，等. 气候变化对中国水稻生产的影响研究进展[J]. 作物学报, 2019, 45(3): 323-334.
LING Xiaoxia, ZHANG Zuolin, ZHAI Jingqiu, et al. A review for impacts of climate change on rice production in China[J]. Acta Agronomica Sinica ,2019, 45(3): 323-334.(in Chinese)
[5] 黎志康. 我国水稻分子育种计划的策略[J]. 分子植物育种, 2005(5): 603-608.
LI Zhikang. Strategies for molecular rice breeding in China[J].Molecular Plant Breeding, 2005(5): 603-608.（in Chinese）
[6] 陈凯文，俞双恩，李倩倩，等. 不同水文年型下水稻节水灌溉技术方案模拟与评价[J/OL]. 农业机械学报, 2019, 50(12):268-277.
CHEN Kaiwen,YU Shuang'en, LI Qianqian, et al. Simulation and evaluation of technical schemes for water-saving irrigation of rice in different hydrological years[J/OL].Transactions of the Chinese Society for Agricultural Machinery,2019,50(12):268-277.
http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20191231&journal_id=jcsam. DOI：10.6041/j.issn.1000-1298.2019.12.031. （in Chinese）
[7] 漆桂林，高桓，吴天星. 知识图谱研究进展[J]. 情报工程, 2017, 3(1): 4-25.
QI Guilin, GAO Huan, WU Tianxing. The research advances of knowledge graph [J]. Technology Intelligence Engineering ,2017, 3(1): 4-25.（in Chinese）
[8] 刘知远，孙茂松，林衍凯，等. 知识表示学习研究进展[J]. 计算机研究与发展, 2016, 53(2): 247-261.
LIU Zhiyuan, SUN Maosong, LIN Yankai, et al. Knowledge representation learning: a review[J]. Journal of Computer Research and Development, 2016, 53(2): 247-261.（in Chinese）
[9] SCHUTZ A, BUITELAAR P. RelExt : a tool for relation extraction from text in ontology extension[C] // International Semantic Web Conference, Galway, Ireland,2005: 593-606.
[10] 郭喜跃，何婷婷. 信息抽取研究综述[J]. 计算机科学, 2015, 42(2): 14-17.
GUO Xiyue, HE Tingting. Survey about research on information extraction[J]. Computer Science, 2015, 42(2): 14-17.（in Chinese）
[11] 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）
[12] LIN Y, LIU Z, SUN M. Neural relation extraction with multi-lingual attention[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vancouver, Canada, 2017:34-43.
[13] WANG M. A re-examination of dependency path kernels for relation extraction[C]// Proceedings of the Third International Joint Conference on Natural Language Processing, Hyderabad, India, 2008:841-846.
[14] CULOTTA L, GIANGUZZA A, MANNINO M R, et al. Polycyclic aromatic hydrocarbons (pah) in Vulcano Island (aeolian archipelago) mud utilized for therapeutic purpose[J]. Polycyclic Aromatic Compounds, 2007, 27(4): 281-294.
[15] HASEGAWA T, SEKINE S, GRISHMAN R. Discovering relations among named entities from large corpora[C]//Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics: Association for Computational Linguistics, Sapporo, Japan , 2004: 415-423.
[16] 赵明，杜亚茹，杜会芳，等. 植物领域知识图谱构建中本体非分类关系提取方法[J/OL]. 农业机械学报, 2016, 47(9): 278-284.
ZHAO Ming, DU Yaru, DU Huifang, et al. Research on ontology non-taxonomic relations extraction in plant domain knowledge graph construction [J/OL]. Transactions of the Chinese Society for Agricultural Machinery, 2016, 47(9):278-284.
http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20160938&journal_id=jcsam. DOI：10.6041/j.issn.1000-1298.2016.09.038. （in Chinese）
[17] KATIYAR A, CARDIE C. Going out on a limb: joint extraction of entity mentions and relations without dependency trees[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Honolulu, Hawaii, USA, 2017: 917-928.
[18] ZENG D, LIU K, LAI S, et al. Relation classification via convolutional deep neural network[C]// Proceedings of the 25th International Conference on Computational Linguistics, Dublin, Ireland, 2014: 2335–2344.

[19] DEVLIN J, CHANG M, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018:1-15.
[20] 袁培森，杨承林，宋玉红，等. 基于Stacking集成学习的水稻表型组学实体分类研究[J/OL]. 农业机械学报, 2019, 50(11): 144-152.
YUAN Peisen,YANG Chenlin,SONG Yuhong,et al.Classification of rice phenomics entities based on stacking ensemble learning[J/OL]. Transactions of the Chinese Society for Agricultural Engineering, 2019,50(11):144-152.
http://www.j-csam.org/jcsam/ch/reader/view_abstract.aspx?flag=1&file_no=20191116&journal_id=jcsam. DOI：10.6041/j.issn.1000-1298.2019.11.016 .(in Chinese)
[21] HAFEMANN L G , SABOURIN R , OLIVEIRA L S. Learning features for offline handwritten signature verification using deep convolutional neural networks[J]. Pattern Recognition, 2017, 70: 163-176.
[22] WANG J, GUO Y. Scrapy-based crawling and user-behavior characteristics analysis on taobao[C]// 2012 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery, Sanya, China, 2012: 44-52.
[23] COOPER L, JAISWAL P. The plant ontology: a tool for plant genomics[M]. Plant Bioinformatics, Springer, 2016, 89-114.
[24] 唐晓波，翟夏普. 基于本体知识集合的知识检索研究[J]. 图书馆学研究, 2018(1): 60-66.
TANG Xiaobo, ZHAI Xiapu. A research on the knowledge retrieval based on ontology knowledge set [J]. Research on Library Science,2018(1): 60-66.（in Chinese）
[25] MAEDCHE A, STAAB S. Ontology learning for the semantic web[J]. IEEE Intelligent systems, 2001, 16(2): 72-79.
[26] WEBBER J. A programmatic introduction to neo4j[C]// Proceedings of the 3rd Annual Conference on Systems, Programming, and Applications: Software for Humanity, Chicago , IL , USA, 2012: 217-218.
[27] HOLZSCHUHER F, PEINL R. Performance of graph query languages: comparison of cypher, gremlin and native access in Neo4j[C]//Proceedings of the Joint EDBT/ICDT 2013 Workshops, Genoa, Italy, 2013: 195-204.
[28] SEO M, KEMBHAVI A, FARHADI A, et al. Bidirectional attention flow for machine comprehension[J]. arXiv preprint arXiv:1611.01603, 2016:1-17.
[29] XIAO T, XU Y, YANG K,et al. The application of two-level attention models in deep convolutional neural network for fine-grained image classification[C]// Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, Massachusetts, USA,2015: 842-850.
[30] 胡荣磊，芮璐，齐筱，等. 基于循环神经网络和注意力模型的文本情感分析[J]. 计算机应用研究, 2019, 36(11): 3282-3285.
HU Ronglei, RUI Lu, QI Xiao, et al. Text sentiment analysis based on recurrent neural networks and attention model[J]. Application Research of Computers, 2019, 36(11): 3282-3285.（in Chinese）
[31] 袁和金，张旭，牛为华，等. 融合注意力机制的多通道卷积与双向GRU模型的文本情感分析研究[J]. 中文信息学报, 2019, 33(10): 109-118.
YUAN Hejin, ZHANG Xu, NIU Weihua, et al. Sentiment analysis based on multi-channel convolution and bi-directional gru with attention mechanism[J]. Journal of Chinese Information Processing, 2019, 33(10): 109-118.（in Chinese）
[32] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in Neural Information Processing Systems, Long Beach, CA, USA,2017: 5998-6008.
[33] 沈忱林，张璐，吴良庆，等. 基于双向注意力机制的问答情感分类方法[J]. 计算机科学, 2019, 46(7): 151-156.
SHEN Chenlin, ZHANG Lu,WU Liangqing, et al .Sentiment classification towards question-answering based on bidirectional attention mechanism.[J]. Computer Science, 2019, 46(7): 151-156.（in Chinese）
[34] SMITH L N. A disciplined approach to neural network hyper-parameters: Part 1--learning rate, batch size, momentum, and weight decay[J]. arXiv preprint arXiv:1803.09820, 2018:89-132.
[35] BOTTOU L. Large-scale machine learning with stochastic gradient descent[C]. Proceedings of COMPSTAT'2010, Springer, 2010, 177-186.
[36] KINGMA D P, BA J. Adam: A method for stochastic optimization[J]. arXiv preprint arXiv:1412.6980, 2014:1-39.
[37] LOSHCHILOV I, HUTTER F. Fixing weight decay regularization in adam[J].arXiv preprint arXiv: 1711.05101, 2018:14-28.
[38] YOU Y, GITMAN I, GINSBURG B. Scaling sgd batch size to 32k for imagenet training[J]. arXiv preprint arXiv:1708.03888, 2017, 6-21.

[39] HARA K, SAITO D, SHOUNO H. Analysis of function of rectified linear unit used in deep learning[C]// Proceedings of 2015 International Joint Conference on Neural Networks (IJCNN), Budapest, Hungary, 2015:1-8.