Journal of Computer Applications
计算机应用, 2020, 40(6): 1580 - 1586
ISSN 1001-9081
CODEN JYIIDU
2020-06-10
http://www.joca.cn

文章编号:1001-9081(2020)06-1580-07
DOI:10.11772/j.issn.1001-9081.2019111951

基于BERT的常见作物病害问答系统问句分类

杨国峰[1,2],杨　勇[1,2*]
(1. 中国农业科学院 农业信息研究所,北京 100081; 2. 农业农村部农业大数据重点实验室,北京 100081)
(*通信作者电子邮箱 wheatblue@163.com)

摘　要:问句分类作为问答系统的关键模块,也是制约问答系统检索效率的关键性因素。针对农业问答系统中用户问句语义信息复杂、差异大的问题,为了满足用户快速、准确地获取常见作物病害问句的分类结果的需求,构建了基于BERT的常见作物病害问答系统的问句分类模型。首先,对问句数据集进行预处理;然后,分别构建双向长短期记忆(Bi-LSTM)自注意力网络分类模型、Transformer分类模型和基于BERT的微调分类模型,并利用三种模型提取问句的信息,进行问句分类模型的训练;最后,对基于BERT的微调分类模型进行测试,同时探究数据集规模对分类结果的影响。实验结果表明,基于BERT的微调常见作物病害问句分类模型的分类准确率、精确率、召回率、精确率和召回率的加权调和平均值分别高于双向长短期记忆自注意力网络模型和Transformer分类模型2~5个百分点,在常见作物病害问句数据集(CCDQD)上能获得最高准确率92.46%,精确率92.59%,召回率91.26%,精确率和召回率的加权调和平均值91.92%。基于BERT的微调分类模型具有结构简单、训练参数少、训练速度快等特点,并能够高效地对常见作物病害问句准确分类,可以作为常见作物病害问答系统的问句分类模型。

关键词:自然语言处理;BERT;作物病害;问答系统;问句分类
中图分类号:S24;TP18;TP391　　**文献标志码**:A

Question classification of common crop disease question answering system based on BERT

YANG Guofeng[1,2], YANG Yong[1,2*]
(1. *Agricultural Information Institute, Chinese Academy of Agricultural Sciences, Beijing* 100081, *China*;
2. *Key Laboratory of Agricultural Big Agri-data, Ministry of Agriculture and Rural Areas, Beijing* 100081, *China*)

Abstract: As a key module of the question answering system, question classification is also a key factor that restricts the retrieval efficiency of the question answering system. Aiming at the problems of complicated semantic information and large differences of user questions in agricultural question answering system, in order to meet the needs of users to quickly and accurately obtain classification results of common crop disease questions, the question classification model of common crop disease question answering system based on Bidirectional Encoder Representations from Transformers (BERT) was constructed. Firstly, the question dataset was preprocessed. Then, Bidirectional-Long Short Term Memory (Bi-LSTM) self-attention network classification model, Transformer classification model and BERT-based fine-tuning classification model were constructed respectively, and the three models were used to extract information of questions and train question classification model. Finally, the BERT-based fine-tuning classification model was tested and the impact of dataset size on classification results was explored. The experimental results show that, the BERT-based fine-tuning common crop disease question classification model has the classification accuracy, precision, recall, weighted harmonic mean of accuracy and recall higher than those of the Bi-LSTM self-attention network classification model and the Transformer classification model by 2-5 percentage points respectively. On Common Crop Disease Question Dataset (CCDQD), it can obtain the highest accuracy of 92.46%, precision of 92.59%, recall of 91.26%, and weighted harmonic mean of accuracy and recall of 91.92%. The BERT-based fine-tuning classification model has advantages of simple structure, few parameters and fast speed, and can efficiently classify common crop disease questions accurately. So, it can be used as the question classification model for the common crop disease question answering system.

Key words: Natural Language Processing (NLP); Bidirectional Encoder Representations from Transformers (BERT); crop disease; question answering system; question classification

收稿日期:2019-11-15;**修回日期**:2020-01-04;**录用日期**:2020-01-06。
基金项目:中国农业科学院科技创新工程项目(CAAS-ASTIP-2016-AII)。
作者简介:杨国峰(1994—),男,重庆人,硕士研究生,CCF会员,主要研究方向:文本分类、情感计算; 杨勇(1975—),男,江苏海门人,副研究员,博士,主要研究方向:智慧农业、农业信息技术。

0　引言

农业科技是现在和未来中国农业增长的第一驱动力,农业技术推广是实现科技进步及农业和农村现代化的重要措施[1]。中国拥有国际上最大的农业技术推广人员队伍,然而研究[2]表明,中国农业技术推广体系未能为农民提供有效的技术服务。当前,一大批农业技术服务平台都在不断地用人工的方式解决农业生产者在作物种植过程中所遇到的病害问题,而依赖人工解决问题将消耗大量的人力、物力,并且很难及时地解决农业生产者的问题。随着人工智能技术的迅猛发展,构建专业领域的智能问答系统将能够为人们提供准确的诊断结果和个性化的信息服务[3]。因此,构建并应用作物病害智能问答系统将为解决以上问题提供解决方案,同时为中国作物病害识别的智能化研究与应用提供重要支撑。问答系统主要包括问题分析(问句分类)、信息检索和答案抽取三个部分,其中问句分类作为问答系统的关键模块,也是制约问答系统检索效率的关键性因素[4]。

关于问答系统中问句分类的研究,从传统的支持向量机、集成学习等[5]到基于词嵌入[6-8],再到神经网络[9-11]分类算法,均在文本分类任务的各项性能评价指标上获得了极大提升,而预训练的神经网络语言模型ELMo(Embeddings from Language Models)、OpenAI GPT(Generative Pre-trained Transformer)、BERT(Bidirectional Encoder Representations from Transformers)取得了显著进展[12]。由于农业领域一直缺乏大规模可用的语料数据库,因此关于农业问答系统问句分类的研究还较少。针对特定农业领域,少数研究者开展了语言模型在农业问答系统应用的相关研究,但仍处于起步阶段。段青玲等[13]重点研究了基于支持向量机的文本分类,实现了92.5%的资讯分类准确率。为了对饮食文本信息进行二分类,赵明等[14]建立了一种基于word2vec和长短期记忆(Long Short-Term Memory, LSTM)网络的分类模型,其分类准确率为98.08%。赵明等[15]还构建了基于word2vec和双向门控循环单元(Bi-Directional Gated Recurrent Unit, BIGRU)神经网络的番茄病虫害问句分类模型,对2 000条番茄病虫害用户问句进行病害和虫害的二分类,结果表明,基于BIGRU的问句分类模型优于卷积神经网络和K最近邻等分类算法。针对传统的句子相似度算法准确率较低的问题,梁敬东等[16]通过构建基于word2vec和LSTM的神经网络计算问句相似度,并在水稻常问问题集中的问句上进行验证,测试集准确率为93.1%。为解决互联网农技推广社区问答数据增长过快的问题,张明岳等[17]构建了一种基于卷积神经网络的农业问答情感极性特征抽取分析模型,针对测试集的语性特征抽取准确率仅为82.7%。

上述研究为神经网络应用于常见作物病害问答系统问句分类提供了参考和依据,但是以上研究主要存在以下两个方面的问题:1)基于word2vec等词嵌入的文本编码方式的文本表征模型还存在局限,无法准确编码同一词在不同语境下的词义;2)尽管构建的LSTM、GRU等神经网络语言模型以及在其基础上改进的双向长短期记忆(Bidirectional-Long Short Term Memory, Bi-LSTM)网络、Transformer等模型能够利用语境信息进行训练,但是相对于自然语言处理任务使用的大型语料,以上研究用来学习的监督数据相对较少,难以学到复杂的语境表示。当前,关于基于语境化的词嵌入[18]利用海量的无监督数据学习神经网络语言模型,即神经网络语言模型的预训练,如ELMo、OpenAI GPT、BERT等模型相继出现,其中BERT以及基于BERT的改进预训练语言模型在多种自然语言任务上取得了最佳结果。

针对常见作物病害问答系统的特点与上述问题,本文研究构建基于word2vec的双向长短期记忆自注意力(Bi-LSTM Self-Attention, Bi-LSTM Self-Attention)网络分类模型、Transformer分类模型和基于BERT的微调分类模型,分别进行常见作物病害问句分类实验,选取能够高效地对常见作物病害问句进行准确分类的问句分类模型作为问答系统最终采用模型。

1　材料与方法

1.1　实验数据与预处理

本研究使用Scrapy爬虫框架,在多个百科类与农业类网站爬取44种常见作物病害相关的农业生产用户的问句,44种常见作物病害如表1所示。

表1　四十四种常见作物病害

Tab. 1　Forty-four common crop diseases

编号	病害名称	编号	病害名称	编号	病害名称
1	苹果黑星病	16	番茄斑枯病	31	黄瓜绿粉病
2	苹果灰斑病	17	番茄花叶病毒病	32	南瓜白粉病
3	苹果锈病	18	番茄黄化曲叶病毒病	33	西瓜果腐病
4	葡萄黑腐病	19	核桃炭疽病	34	柑橘黄龙病
5	葡萄褐斑病	20	枣黑腐病	35	香蕉煤纹病
6	葡萄轮斑病	21	花生褐斑病	36	胡萝卜黑斑病
7	桃树疮痂病	22	芝麻青枯病	37	甘蔗凤梨病
8	辣椒疮痂病	23	玉米灰斑病	38	甜菜立枯病
9	香蕉黄叶病	23	玉米叶斑病	39	大豆紫斑病
10	番茄白粉病	25	玉米锈病	40	绿豆猝倒病
11	番茄疮痂病	26	玉米矮花叶病毒病	41	茄子病毒病
12	番茄早疫病	27	樱桃白粉病	42	茄子褐纹病
13	番茄晚疫病	28	马铃薯早疫病	43	白菜软腐病
14	番茄斑点病	29	马铃薯晚疫病	44	草莓叶枯病
15	番茄叶霉病	30	红薯根腐病		

参考文献[19]中对作物病害的描述信息，对收集的语料进行预处理（去除重复数据，问句转换为陈述句等），从而构建常见作物病害问句数据集（Common Crop Disease Question Dataset，CCDQD）。预处理后部分样本如表2所示。

表 2 部分预处理作物病害问句样本

Tab. 2 Some preprocessed samples of crop disease question

问句	病害名称
葡萄叶片上有红褐色不规则或圆形的斑点。	葡萄轮斑病
玉米叶片上有平行边缘不透明的灰黑色病斑。	玉米灰斑病
在干旱时候或高温高湿等环境条件下花生容易发病，且叶面出现白色的粉末。	番茄白粉病
胡萝卜病斑上出现黑霉。	胡萝卜黑斑病
柑橘黄梢，外围部分枝条或树顶新梢叶片黄化，黄化的叶片极易脱落。	柑橘黄龙病
西瓜皮上有水浸状的各种病斑，有的西瓜皮裂开，腐烂。	西瓜果腐病
湿度大时，花生病斑上可见灰褐色粉状霉层，叶柄和茎秆染病，病斑长椭圆形，暗褐色。	花生褐斑病
茄子从苗期到成株期，地上各部位均可发病，以果实受害最重。	茄子褐纹病
茎、果上的病斑近圆形或椭圆形，褐色，略凹陷，斑点上散生小黑点。	番茄斑枯病
苹果树叶子出现了圆形黄色斑点，边缘为红色。	苹果锈病

1.2 双向长短期记忆自注意力网络分类模型

本文使用自注意力技术来生成句子词嵌入分类模型[20]，由 Bi-LSTM 和一层全连接 Softmax 层构成，如图 1 所示。在 Bi-LSTM 的基础上，模型将通过自注意力机制得到句子的表示输出到隐藏层，然后通过全连接层进行分类。

如图1所示，向模型输入一个含有 n 个词的句子进行词嵌入，得到 $\boldsymbol{S}=(\boldsymbol{w}_1,\boldsymbol{w}_2,\cdots,\boldsymbol{w}_n)$，$\boldsymbol{w}_i$ 表示序列中第 i 个标记（Token）对应的词嵌入。$\boldsymbol{h}_1,\boldsymbol{h}_2,\cdots,\boldsymbol{h}_n$ 为隐藏层的对应输出，Bi-LSTM 将句子嵌入为 $\boldsymbol{M}$，$(\boldsymbol{m}_1,\boldsymbol{m}_2,\cdots,\boldsymbol{m}_r)$ 表示聚焦句子不同的部分，其中注意力权重为 $\boldsymbol{A}_{i1},\boldsymbol{A}_{i2},\cdots,\boldsymbol{A}_{in}$。

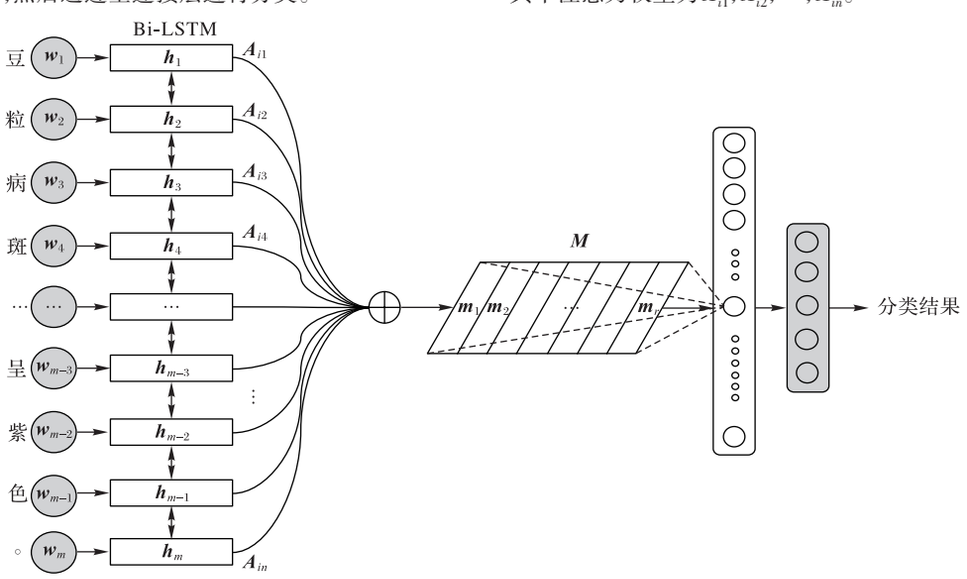

图 1 双向长短期记忆自注意力网络分类模型

Fig. 1 Bi-LSTM self-attention network classification model

给定训练集 $\{(\boldsymbol{S}^{(i)},y^{(i)})|i=1,2,\cdots,N\}$，其中类别标签 $y^{(i)}\in\{1,2,\cdots,K\}$（$K$ 为可能的类别数目）。为了削弱末级分类器的复杂度，强迫模型学习到更有效的表示，有助于下游问句分类任务，把从文本 $\boldsymbol{S}^{(i)}$ 生成的矩阵级别表示 $\boldsymbol{M}^{(i)}$ 输入 Softmax 层便得到离散类别标签的预测概率分布为：

$$p^{(i)}=\mathrm{softmax}(\boldsymbol{W}_S\boldsymbol{M}^{(i)}+b_S) \tag{1}$$

其中：$\boldsymbol{W}_S$ 为 $\boldsymbol{M}^{(i)}$ 相应的权重；b_S 是偏差。

本文将 Cross Entropy Loss 损失函数加上 Softmax 层权值矩阵的 Frobenius 范数约束作为训练整个网络的损失函数：

$$Loss(\theta)=-\sum_{i=1}^{N}\ln p_{y^{(i)}}^{(i)}+\alpha\left\|\boldsymbol{W}_S\right\|_{\mathrm{F}}^2 \tag{2}$$

其中：θ 表示网络中所有参数；$p_{y^{(i)}}^{(i)}$ 为 $p^{(i)}$ 的第 $y^{(i)}$ 个分量；α 是惩罚系数，用来调节惩罚项 $\left\|\boldsymbol{W}_S\right\|_{\mathrm{F}}^2$ 的比重，下标 F 表示 Frobenius 范数。

1.3 Transformer 分类模型

本文构建的 Transformer 分类模型基于编码器（Encoder）结构[21]，由于 Transformer 模型训练参数较多且模型较为复杂，因此仅使用了两个编码器，每个编码器包含一个多头注意力子层和一个前馈网络子层。模型中的所有子层以及嵌入层的输出尺寸为200。图2为两个编码器的 Transformer 分类模型。

句子的词嵌入（Sentence Embedding）与对应位置词嵌入（Position Embedding）相加后作为输入，编码器第一个子层是多头自注意力（Multi-head Attention），输出表示为 $sublayer(x)$，经过残差连接和层规范（Add & Layer Norm，LN）输出为：

$$output=\mathrm{LN}(x+sublayer(x)) \tag{3}$$

编码器第二个子层是逐项前馈网络（Feed Forward Network，FFN），由两个线性变换组成，其中每一层的参数都

不同,输入和输出的维度为200,内部层的维度为2 400。

$$FFN(output) = \mathrm{Max}(0, output*\boldsymbol{W}_1 + b_1)\boldsymbol{W}_2 + b_2 \quad (4)$$

其中,$\boldsymbol{W}_1$、$\boldsymbol{W}_2$和b_1、b_2分别为*output*相应线性变换的权重和偏差。

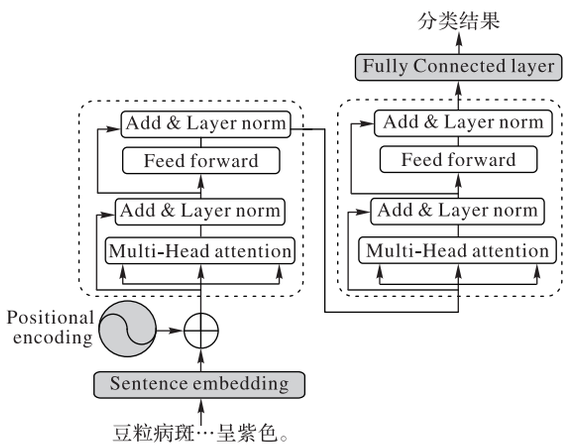

图 2　Transformer分类模型

Fig. 2　Transformer classification model

线性变换之后再经过残差连接和层规范,将最后一个编码器第二子层的最终输出传递进全连接层并输出分类结果。训练过程使用Cross Entropy Loss损失函数。

1.4　基于BERT的微调分类模型

借助语言模型来辅助自然语言处理任务已经得到了学术界较为广泛的探讨[22],通常有两种方式:1)基于特征,指利用语言模型的中间结果(语言模型词嵌入),将其作为额外的特征,引入到原任务的模型中,如ELMo模型[23];2)基于微调,指利用大量语料训练语言模型,并在语言模型基础上增加少量神经网络层来完成具体任务,采用有标记的语料来有监督地训练新模型,这个过程中语言模型的参数并不固定,如OpenAI GPT[24]。上述模型的输入为从左向右输入一个文本序列,或将从左向右输入和从右向左输入的训练结合起来。然而,BERT是一种新的预训练语言模型,即双向编码表征Transformer的模型。相关研究[12]表明:双向训练的语言模型对语境的理解会比单向的语言模型更深刻,提取语料特征更高效。

由于BERT可以用于各种自然语言处理的任务(如分类任务、问答任务),且仅需在核心模型的基础上进行简单修改,因此本文将构建基于BERT的微调分类模型,如图3所示。

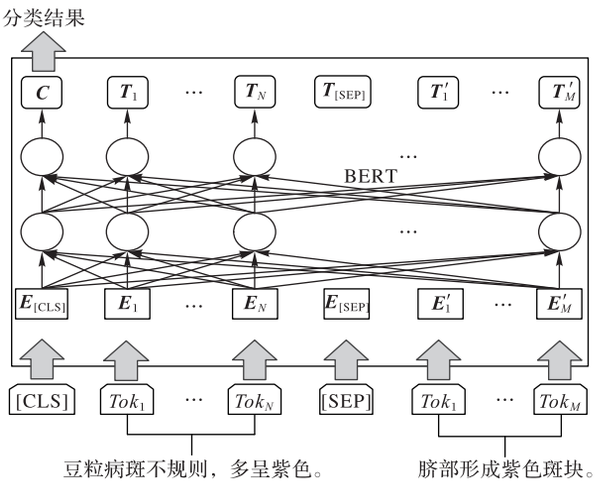

图 3　基于BERT的微调分类模型

Fig. 3　BERT-based fine-tuning classification model

对于问答系统的问句分类任务,基于BERT的微调分类模型在预训练BERT模型的输出结果后增加一个分类层(全连接层)进行微调。

把作物病害问句输入模型后,将被传递到词嵌入层,包括进行标记词嵌入、句子词嵌入和位置词嵌入。在图3和图4中:Tok_i表示第i个Token,随机遮挡部分字符;$\boldsymbol{E}_i$表示第i个Token的嵌入向量;$\boldsymbol{T}_i$表示第i个Token在经过BERT处理之后得到的特征向量。

BERT和可学习的权值矩阵($\boldsymbol{W}$)所有参数都经过微调,以最大化正确分类的概率。BERT根据[CLS]标志生成一组特征向量$\boldsymbol{C}$,并将其与$\boldsymbol{W}$相乘,再经过Softmax预测各个类别的概率,其中概率最大的类别为最后输出的分类类别。

利用预训练语言模型的参数权重对模型初始化,使用Cross Entropy Loss损失函数对基于BERT的微调分类模型进行有监督的训练。

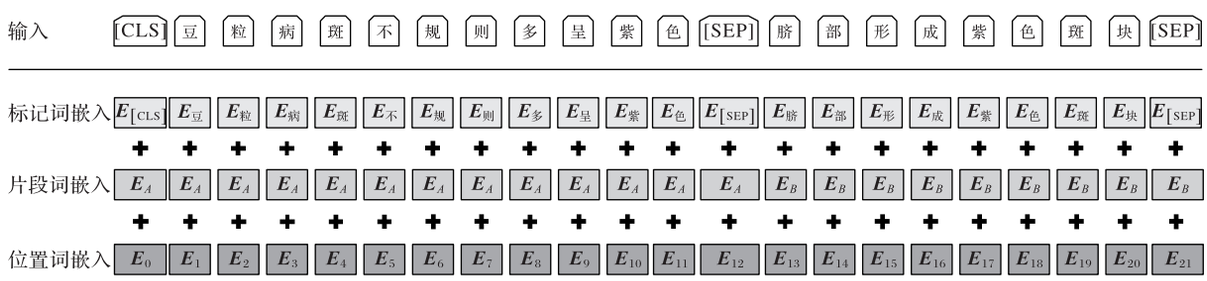

图 4　BERT的输入表示

Fig. 4　Input representation of BERT

2　问句分类实验

2.1　实验设计

本研究收集44种常见作物病害的问句,为防止数据集类别不平衡的影响,使每种作物病害样本数相近(60~80条),同时将每条病害问句标注为对应所属类别,共得到有3 300条样本。利用常用的优化学习Adam梯度下降算法[25]训练优化双向长短期记忆自注意力网络分类模型、Transformer分类模型和基于BERT的微调分类模型,再根据损失函数动态调整每个参数的学习率。

将超参数初始化并调优,最终设置如下:训练批量(batch)大小为16,问句词嵌入长度为200,多注意力头数(multi-head)为12,初始学习率为0.001,最小学习率为

0.000 01，迭代次数（epochs）为 100。为减轻三种模型在训练过程中过度参数化、过拟合，以及避免偶然出现的不良局部最小值现象，设置 Dropout 参数[26]为 0.1。同时，为得到可靠稳定的模型，采用十折交叉验证（10-fold cross-validation）的方法[27]进行训练。

本文研究采用准确率、精确率和召回率以及综合指标 F_1 值作为问句分类模型的测评指标[28]。其中：准确率是指分类正确的问句数除以整个数据集的问句总数；精确率是指分类器正确判断为该类的问句数与分类器判断属于该类的问句总数之比；召回率是指分类器正确判断为该类的问句数与属于该类的问句总数之比。F_1 值是精确率和召回率的调和平均值，最大值为 1，最小值为 0：

$$F_1 = \frac{2PR}{P+R} \times 100\% \tag{5}$$

其中：P 为精确率；R 为召回率。

2.2 分类实验结果及分析

对构建的双向长短期记忆自注意力（Bi-LSTM Self-Attention）网络分类模型、Transformer 分类模型和基于 BERT 的微调分类模型分别使用预处理后的同一数据集进行分类实验。

由表 3 可知，对于不同的迭代次数，基于 BERT 的微调分类模型与另外两个模型相比，其准确率、精确率和召回率均高几个百分点。当三个模型同时训练 100 迭代次数时，对于 F_1 值指标，基于 BERT 的微调分类模型 F_1 值为 91.92%，比 Transformer 分类模型高 2.62 个百分点，比利用双向长短期记忆自注意力网络分类模型高 4.46 个百分点，表明：基于 BERT 的微调分类模型的问句分类效果最优，Transformer 分类模型居中，双向长短期记忆自注意力网络分类模型结果稍差。

本实验也证实了双向长短期记忆自注意力网络分类模型提取问句语言特征的能力弱于 Transformer，尽管双向长短期记忆自注意力网络分类模型增加了自注意力机制，但是 Transformer 分类模型的准确率、精确率和召回率以及 F_1 值均超过了双向长短期记忆自注意力网络分类模型。相较于双向长短期记忆自注意力网络分类模型和 Transformer 分类模型，虽然基于 BERT 的微调分类模型与 Transformer 分类模型的网络结构有相似的结构，但基于 BERT 的微调分类模型在预训练阶段通过无监督的方法学习具有上下文语义的词嵌入特征，能更好地表达语义，在微调阶段再用监督的方法训练 BERT 模型和全连接层的参数。基于 BERT 的微调模型的优势在于充分利用了具有上下文语义的信息，可基于少量监督学习样本，针对不同下游任务改造模型实现目标。实验结果表明，基于 BERT 的微调分类模型具有结构简单、训练参数少、训练速度快等特点，同时能够高效地对常见作物病害问句准确分类，可以作为问答系统问句分类模型。

表 3 问句分类模型的分类结果

Tab. 3 Classification results of question classification models

模型	迭代次数	准确率/%	精确率/%	召回率/%	F_1/%
双向长短期记忆自注意力网络分类模型	30	80.89	83.23	82.49	82.86
	50	83.26	84.57	85.18	84.87
	100	87.72	88.38	86.57	87.46
Transformer 分类模型	30	83.54	81.14	82.84	81.98
	50	85.24	87.67	86.29	86.97
	100	89.67	90.58	88.05	89.30
基于 BERT 的微调分类模型	30	85.82	87.26	85.48	86.36
	50	90.53	89.94	88.48	89.20
	100	92.46	92.59	91.26	91.92

使用基于 BERT 的微调分类模型进行作物病害问句分类，正确分类的部分实例与对应的分类结果如表 4 所示。

表 4 问句分类结果的部分实例

Tab. 4 Some examples of question classification results

问句	分类结果
叶片红褐色，而且有不规则形的病斑和同心轮纹；潮湿的时候，背面还有灰褐色的霉层。	葡萄轮斑病
玉米叶片变黄枯死，病斑后期在叶片两面出现灰黑色霉层。	玉米灰斑病
在番茄叶正面有放射状的粉斑和圆形粉斑，叶片表面有白色粉状物。	番茄白粉病
胡萝卜有褐色的小病斑，病斑边缘有黄色圈。	胡萝卜黑斑病
柑橘叶片叶脉及附近的组织变绿色；叶肉变黄。	柑橘黄龙病
西瓜沿叶片中脉出现不规则褐色病斑，有的扩展到叶缘，叶背面呈水浸状。	西瓜果腐病
花生的叶柄和茎秆有椭圆形褐色的病斑。	花生褐斑病
茄子出现褐色病斑、轮纹排列的小黑点。	茄子褐纹病
番茄叶柄和茎上长有椭圆形褐色病斑，其上长有黑色小粒点。	番茄斑枯病
苹果叶子像生锈了似的。	苹果锈病

基于上述实验，选择作物病害问答系统问句分类效果最优的基于 BERT 的微调分类模型，对影响问句分类效果的样本数量（训练数据集规模）进行研究。在保持模型结构和初始超参数不变的情况下，改变数据集规模（1 100、2 200、3 300）进行实验，迭代次数均设置为 100 次，测试结果如表 5 所示。

表 5 不同样本数量的模型分类结果

Tab. 5 Model classification results under different sample sizes

样本数	准确率/%	精确率/%	召回率/%	F_1/%
1 100	87.13	85.69	86.42	86.05
2 200	89.69	88.21	87.53	87.87
3 300	92.46	92.59	91.26	91.92

由表 5 可得，训练数据集的规模对基于 BERT 的微调分类模型的分类效果有较大影响。鉴于真实场景下作物病害识别分类对准确率有较高的需求，因此在对基于 BERT 的微调分类模型进行训练时，大量高质量的训练数据集可以提高整个作物病害问答系统类别分类的准确率。

探究基于 BERT 的微调分类模型作为作物病害问答系统问句模型的有效性，并进行相关实验。在不同样本数量的数据集下，三种模型的有效性（F_1 值）如图 5 所示。

由图 5 可知：BERT 模型的 F_1 值在不同数量的数据集上均保持最高；而当数据集较小时，双向长短期记忆自注意力网络分类模型 F_1 值比 Transformer 分类模型的 F_1 值高；随着数据集

的增大,Transformer分类模型的F_1值反超双向长短期记忆自注意力网络分类模型F_1值。BERT模型的预训练是一个耗时的过程,通过使用预训练好的模型进行分类任务微调仍然能够实现很好的分类效果,验证了基于BERT的微调分类模型作为作物病害问答系统问句模型的有效性。

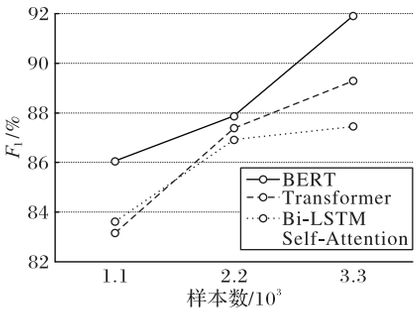

图 5　不同模型的有效性结果对比

Fig. 5　Validity comparison of different models

3　结语

本文针对农业问答系统用户问句语义信息复杂、差异大的问题,构建了基于BERT用于问句分类任务的作物病害问答系统的问句分类模型,证实了两阶段模型(超大规模预训练与具体任务微调)较强的表征学习能力对问句分类的影响。通过与双向长短期记忆自注意力网络分类模型和Transformer分类模型进行对比,由实验结果可知,基于BERT的微调分类模型的准确率、精确率和召回率以及F_1值均高于另外两种问句分类模型2~5个百分点,表明基于BERT的微调常见作物病害问句分类模型可以高效地提取问句文字的特征,用于后续的分类实验。对三种模型使用不同数量的数据集进行实验,结果表明,随着数据集规模的增加,基于BERT的微调分类模型与另外两种模型相比F_1值明显上升,表明了问句分类模型的选择和训练数据集的规模对常见作物病害问答系统问句分类效果具有较大影响。接下来将进一步扩大作物病害问句类别的覆盖范围,满足更多用户对作物病害类别识别的需求。

参考文献 (References)

[1] HUANG J, ROZELLE S. Technological change: rediscovering the engine of productivity growth in China's rural economy [J]. Journal of Development Economics, 1996, 49(2): 337-369.

[2] 孙生阳,孙艺夺,胡瑞法,等. 中国农技推广体系的现状、问题及政策研究[J]. 中国软科学,2018(6):25-34. (SUN S Y, SUN Y D, HU R F, et al. Current situation, problems and policy of agricultural extension system in China [J]. China Soft Science, 2018(6): 25-34.)

[3] 赵春江. 智慧农业发展现状及战略目标研究[J]. 智慧农业,2019,1(1):1-7. (ZHAO C J. State-of-the-art and recommended developmental strategic objectives of smart agriculture [J]. Smart Agriculture, 2019, 1(1): 1-7.)

[4] 郑实福,刘挺,秦兵,等. 自动问答综述[J]. 中文信息学报,2002,16(6):46-52. (ZHENG S F, LIU T, QIN B, et al. Overview of question-answering [J]. Journal of Chinese Information Processing, 2002, 16(6): 46-52.)

[5] 苏金树,张博锋,徐昕. 基于机器学习的文本分类技术研究进展[J]. 软件学报,2006,17(9):1848-1859. (SU J S, ZHANG B F, XU X. Advances in machine learning based text categorization [J]. Journal of Software, 2006, 17(9): 1848-1859.)

[6] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space [EB/OL]. [2019-03-12]. https://arxiv. org/pdf/1301. 3781. pdf.

[7] LE Q, MIKOLOV T. Distributed representations of sentences and documents [EB/OL]. [2019-03-12]. https://cs. stanford. edu/~quocle/paragraph_vector. pdf.

[8] PENNINGTON J, SOCHER R, MANNING C. Glove: global vectors for word representation [C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg: ACL, 2014: 1532-1543.

[9] LAI S, XU L, LIU K, et al. Recurrent convolutional neural networks for text classification [C]// Proceedings of the 29th AAAI Conference on Artificial Intelligence. Palo Alto: AAAI Press, 2015:2267-2273.

[10] ZHOU P, QI Z, ZHENG S, et al. Text classification improved by integrating bidirectional LSTM with two-dimensional max pooling [C]// Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers. Stroudsburg: ACL, 2016: 3485-3495.

[11] LE T T H, KIM J, KIM H. Classification performance using gated recurrent unit recurrent neural network on energy disaggregation [C]// Proceedings of the 2016 International Conference on Machine Learning and Cybernetics. Piscataway: IEEE, 2016: 105-110.

[12] 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. Stroudsburg: ACL, 2019: 4171-4186.

[13] 段青玲,魏芳芳,张磊,等. 基于Web数据的农业网络信息自动采集与分类系统[J]. 农业工程学报,2016,32(12):172-178. (DUAN Q L, WEI F F, ZHANG L, et al. Automatic acquisition and classification system for agricultural network information based on Web data [J]. Transactions of the Chinese Society of Agricultural Engineering, 2016, 32(12): 172-178.)

[14] 赵明,杜会芳,董翠翠,等. 基于word2vec和LSTM的饮食健康文本分类研究[J]. 农业机械学报,2017,48(10):202-208. (ZHAO M, DU H F, DONG C C, et al. Diet health text classification based on word2vec and LSTM [J]. Transactions of the Chinese Society for Agricultural Machinery, 2017, 48(10): 202-208.)

[15] 赵明,董翠翠,董乔雪,等. 基于BIGRU的番茄病虫害问答系统问句分类研究[J]. 农业机械学报,2018,49(5):271-276. (ZHAO M, DONG C C, DONG Q X, et al. Question classification of tomato pests and diseases question answering system based on BIGRU [J]. Transactions of the Chinese Society for Agricultural Machinery, 2018, 49(5): 271-276.)

[16] 梁敬东,崔丙剑,姜海燕,等. 基于word2vec和LSTM的句子相似度计算及其在水稻FAQ问答系统中的应用[J]. 南京农业大学学报,2018,41(5):946-953. (LIANG J D, CUI B J, JIANG H Y, et al. Sentence similarity computing based on word2vec and LSTM and its application in rice FAQ question-answering system

[J]. Journal of Nanjing Agricultural University, 2018, 41(5): 946-953.)

[17] 张明岳,吴华瑞,朱华吉. 基于卷积模型的农业问答语性特征抽取分析[J]. 农业机械学报,2018,49(12):203-210. (ZHANG M Y, WU H R, ZHU H J. Analysis of extraction of semantic feature in agricultural question and answer based on convolutional model [J]. Transactions of the Chinese Society for Agricultural Machinery, 2018, 49(12): 203-210.)

[18] 李枫林,柯佳. 词向量语义表示研究进展[J]. 情报科学,2019,37(5):155-165. (LI F L, KE J. Research progress of word vector semantic representation [J]. Information Science, 2019, 37(5): 155-165.)

[19] 中国农业科学院植物保护研究所,中国植物保护学会. 中国农作物病虫害[M]. 3版. 北京:中国农业出版社,2015:26-58. (Institute of Plant Protection of Chinese Academy of Agricultural Sciences, China Society of Plant Protection. Chinese Crop Pests and Diseases [M]. 3rd ed. Beijing: China Agricultural Press, 2015: 26-58.)

[20] LIN Z, FENG M, DOS SANTOS C N, et al. A structured self-attentive sentence embedding [EB/OL]. [2019-03-12]. https://arxiv.org/pdf/1703.03130.pdf.

[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. Red Hook: Curran Associates Inc., 2017: 6000-6010.

[22] 郁可人,傅云斌,董启文. 基于神经网络语言模型的分布式词向量研究进展[J]. 华东师范大学学报(自然科学版),2017(5):52-65,79. (YU K R, FU Y B, DONG Q W. Survey on distributed word embeddings based on neural network language models [J]. Journal of East China Normal University (Natural Science), 2017(5): 52-65, 79.)

[23] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations [C]// Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Stroudsburg: ACL, 2018: 2227-2237.

[24] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training [EB/OL]. [2019-03-12]. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.

[25] KINGMA D P, BA J L. Adam: a method for stochastic optimization [EB/OL]. [2019-03-12]. https://arxiv.org/pdf/1412.6980.pdf.

[26] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from overfitting [J]. The Journal of Machine Learning Research, 2014, 15(1): 1929-1958.

[27] 陈鹏,冯海宽,李长春,等. 无人机影像光谱和纹理融合信息估算马铃薯叶片叶绿素含量[J]. 农业工程学报,2019,35(11):63-74. (CHEN P, FENG H K, LI C C, et al. Estimation of chlorophyll content in potato using fusion of texture and spectral features derived from UAV multispectral image [J]. Transactions of the Chinese Society of Agricultural Engineering, 2019, 35(11): 63-74.)

[28] POWERS D M W. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation [J]. Journal of Machine Learning Technologies, 2011, 2(1): 37-63.

This work is partially supported by the Science and Technology Innovation Project of Chinese Academy of Agricultural Sciences (CAAS-ASTIP-2016-AII).

YANG Guofeng, born in 1994, M. S. candidate. His research interests include text categorization, affective computing.

YANG Yong, born in 1975, Ph. D., associate research follow. His research interests include smart agriculture, agricultural information technology.