doi:10.6041/j.issn.1000-1298.2022.05.034

基于多语义特征的农业短文本匹配技术

金　宁[1,2]　赵春江[3,4]　吴华瑞[3,4]　缪祎晟[3,4]　王海琛[3,5]　杨宝祝[3,4]

(1.沈阳农业大学信息与电气工程学院,沈阳 110866;2.沈阳建筑大学研究生院,沈阳 110168;
3.国家农业信息化工程技术研究中心,北京 100097;4.北京农业信息技术研究中心,北京 100097;
5.沈阳建筑大学信息与控制工程学院,沈阳 110168)

摘要:“中国农技推广 APP”农业问答社区存在提问数据量大、规范性差、涉及面广、噪声多、特征稀疏等影响文本语义匹配的问题,为了改善农业提问数据相似性判断的性能,提出了融合多语义特征的文本匹配模型 Co_BiLSTM_CNN,从深度语义、词语共现、最大匹配度 3 个层面提取短文本特征,并利用共享参数的孪生网络结构,分别运用双向长短期记忆网络、卷积神经网络和密集连接网络构建文本匹配模型。试验结果表明,该模型可以更全面提取文本特征,文本相似性判断的正确率达 94.15%,与其他 6 种模型相比,文本匹配效果优势明显。

关键词:农业短文本匹配;多语义特征;词语共现;双向长短期记忆网络;卷积神经网络

中图分类号:TP183　　**文献标识码**:A　　**文章编号**:1000-1298(2022)05-0325-07　　**OSID**:

Agricultural Short Text Matching Technology Based on Multi-semantic Features

JIN Ning[1,2]　ZHAO Chunjiang[3,4]　WU Huarui[3,4]　MIAO Yisheng[3,4]　WANG Haichen[3,5]　YANG Baozhu[3,4]

(1. *School of Information and Electrical Engineering, Shenyang Agricultural University, Shenyang* 110866, *China*
2. *Graduate School, Shenyang Jianzhu University, Shenyang* 110168, *China*
3. *National Engineering Research Center for Information Technology in Agriculture, Beijing* 100097, *China*
4. *Beijing Research Center for Information Technology in Agriculture, Beijing* 100097, *China*
5. *School of Information and Control Engineering, Shenyang Jianzhu University, Shenyang* 110168, *China*)

Abstract: With the development of information technology, agricultural information consultant service based on mobile Internet has become an important part of agro-technical extension system. More than ten million questions in all have been collected by agro-technical extension Q&A community. With the continuous popularization of Q&A community, answering questions manually only by agricultural experts and technicians can neither follow the rapid growth of the questions nor meet the needs of farmers who want to be answered quickly and accurately. Agricultural intelligent Q&A is one of the effective ways to solve the problem. High quality text matching for new questions is the key technology. The accuracy of text matching is limited by the characteristics of agricultural text, such as large amount of data, poor standardization, wide range, much noise, and sparse features. In order to improve the accuracy, the deep semantics, word co-occurrence and maximum matching degree of agricultural short text were extracted and Co_BiLSTM_CNN model composed of bi-long short-term memory, convolutional neural networks, dense networks and Siamese network of shared parameters, was proposed to extract multi-semantic features. The precision, recall, F1, accuracy and time complexity were selected as evaluation indexes to comprehensively measure the performance of the model. The experimental results showed that the model could extract text features more comprehensively, with an accuracy of 94.15%. Compared with the other six text matching models, the experimental results showed obvious advantages.

Key words: agricultural short text matching; multi-semantic features; word co-occurrence; bi-long short-term memory network; convolutional neural network

收稿日期:2021-05-27　修回日期:2021-06-28

基金项目:国家重点研发计划项目(2019YFD1101105)、北京市科技计划项目(Z191100004019007)、国家自然科学基金项目(61871041)和国家大宗蔬菜产业技术体系岗位专家项目(CARS-23-C06)

作者简介:金宁(1989—),男,博士生,沈阳建筑大学助理研究员,主要从事农业智能系统研究,E-mail:jinning21@126.com

通信作者:赵春江(1964—),男,研究员,中国工程院院士,主要从事农业人工智能与智能系统研究,E-mail:zhaocj@nercita.org.cn

0 引言

农业信息咨询服务是农业技术推广的途径之一。“中国农技推广 APP”搭建了高效、便捷的手机移动端农业信息咨询服务平台,至今已完成农业技术问答超过了千万次,涵盖了蔬菜、粮食作物、牲畜等 10 个品种,涉及病虫草害、栽培管理、动物疫病等 18 个种植、养殖方面问题[1]。面对平台问答社区飞速增长的提问数量和大量重复、相似提问的存在,仅依靠农业专家及农业技术人员的人工回答已无法满足农户快速、准确获得答案的需求。利用大数据和文本挖掘技术的文本匹配计算可有效解决上述问题。文本匹配作为文本挖掘技术的重要组成部分,是实现智能问答最关键、最基础的技术之一[2],是农业大数据智能研究领域的重点研究方向。

文本匹配即构建一个模型来学习和计算 2 个文本的相似度,判定 2 个句子是否相似[3]。文本匹配方法包括基于字面相似匹配、基于语法结构相似匹配和基于语义相似匹配等方法[4]。基于字面相似匹配一般利用重复词、同义词、公共字符串等词型表面特征,采用 N-gram[5] 和 Jaccard[6] 算法进行相似度判断。此方法仅提取了文本相同词型或词数量等浅层特征,忽视了文本语言结构、语义等深层特征的提取。基于语法结构相似匹配通常利用标注的词语属性计算文本间相似性[7],但是由于在特定领域内缺少带有词语属性标注的文本,并且专业领域内涵盖的专属词语较多,影响了此方法相似度计算的准确性。近年来基于语义相似匹配方法在文本匹配研究中应用广泛[8-10]。随着深度学习技术的发展,将文本转换为 Word2vec[11-12]、GloVe[13] 等由神经网络生成的词向量,运用深度学习文本匹配模型[14-16] 计算文本相似度已成为当前主要研究方法。深度学习方法能够从文本中自动学习、抽取特征,可有效解决人工设计特征泛化能力差,提取特征数量少的问题。卷积神经网络是常用的深度学习方法,其卷积核结构能够有效提取文本局部的信息,通过堆叠卷积层的方法提取文本不同粒度的特征[17-19],但在文本序列化特征提取,特别是长距离依存关系提取表现不佳。相比卷积神经网络,循环神经网络更适用于文本序列化建模[20-22]。循环神经网络可将历史信息保存并作用于后续的计算进而提高匹配的精确度。但此方法也存在对关键信息不敏感,训练时间长,内存消耗大等问题。注意力机制[23] 可有效解决上述问题,其通过赋予关键词语更高的权重,帮助模型快速获得最有效的信息,从而节省计算资源。在农业领域,深度学习方法在文本分类[24-26]、文本信息提取[27-28]、命名实体识别[29-31]、实体关系抽取等方面取得了一定研究成果。但现有研究仍存在文本特征提取模型结构简单,短文本特征提取不丰富、不全面等不足,对于农业领域中大量存在的专有名词、词语搭配方面的短文本挖掘仍不够深入。

本文针对农业问答社区产生的提问文本词汇总量少、专有名词多,语句规范性差,特征稀疏性强的特点,从深度语义、词语共现、最大匹配度 3 个层面提取短文本特征。根据不同层面文本特征的特点,利用孪生长短期记忆网络、卷积神经网络和密集连接神经网络,搭建基于多语义的农业短文本匹配模型 Co_BiLSTM_CNN,以实现提问文本相似性判断。

1 多语义特征文本匹配模型

本文提出的基于多语义特征文本 Co_BiLSTM_CNN 模型如图 1 所示。该模型主要由文本预处理层和文本匹配层 2 部分组成。与一般文本匹配模型相比,本文提出的模型在文本预处理层构建了词语共现关系网,可计算每个词语共现对的权重,进一步丰富了短文本特征;再根据不同文本特征的特点,利用双向长短期记忆网络、卷积神经网络和密集连接网络搭建文本匹配模型,可实现文本特征的多角度提取。

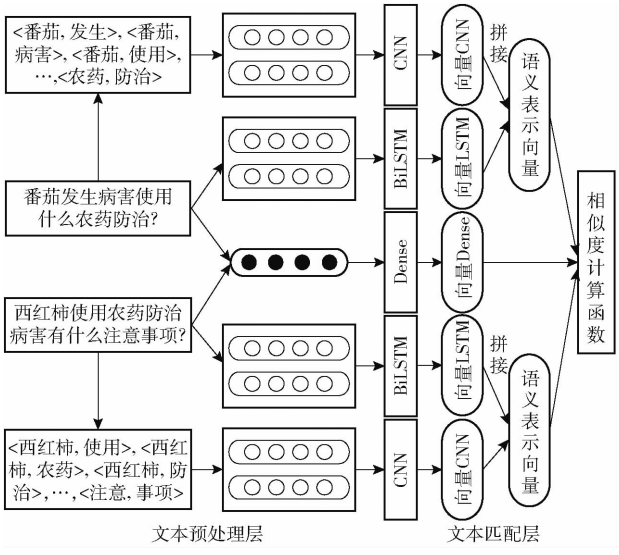

图 1 Co_BiLSTM_CNN 模型结构图

Fig. 1 Schematic of Co_BiLSTM_CNN

1.1 文本预处理

(1)文本分词及词性标注

为了进一步提高文本特征提取的精确度,提升文本处理效率,更加全面、完整地保留文本语义,首先运用加载停用词表方法,删除文本包含的停用词、特殊字符、标点符号等冗余信息,减少噪声对文本语义的干扰;然后利用 jieba 中文分词库对文本进行分词,针对文本中农业领域专有词语识别难的问题,加

载搜狗农业词汇字典作为分词字典，以提高专有名词的识别度；为了更加深入地提取文本特征，将分词结果的词性作为文本特征，按照北京大学词性标注规范[32]，对分词结果的词性逐一进行标注，利于提取更能体现文本语义的名词、动词。对于具有语义代表性的农业领域专有词语，在词性标注时专门自定义了词性，更加方便模型学习、提取文本特征。

(2)词向量转换

由于文本分词结果无法被模型直接读取，需将中文分词结果转换为数字向量形式表示。Word2vec[33]是近年来常用的基于神经网络的文本向量化表示方法，包括根据上下文预测当前单词和根据目标词语预测上下文信息 2 种模式，通过文本数据特征可学习得到连续、低维词向量，有效解决了传统词向量模型带来的语义孤立和维度爆炸问题。

(3)文本特征增强

文本词语共现关系是指在同一文本内，几个词语频繁共同出现构成的关系，是一种基于统计学的重要文本特征。在一个问句内，几个词语在固定窗口范围内多次共同出现，寓意其很可能表述相同的主题，在语义上具有一定的关联性。对于 2 个不同的问句，相同词语共现关系的数量越多，代表着两者之间语义相似度越高。问句一般属于短文本，具有词语数量少、语义特征不明显的特点，利用文本词语共现关系可有效增强文本语义特征。为了进一步突出文本核心语义，减少噪声干扰，在构建词语的共现关系时，先对词语的词性进行初步筛选，仅统计更能代表文本语义的名词、动词以及农业字典中包含的专有名词；再对词频进行限定，仅统计高于 3 次的词语，在保留了原有问句核心词语的基础上，使文本语义特征更加丰富。

1.2　多特征文本匹配模型

(1)孪生网络模型

孪生网络(Siamese network)是一种双边的网络模型架构，多用于判断 2 个输入的相似性，具体结构如图 2 所示。左右 2 个输入首先经过 LSTM 或者 CNN 等神经网络模型训练而映射到新的空间，形成输入在新空间中的表示；其次通过距离度量的方式可计算左右 2 个神经网络输出的距离。训练参数包括左右参数共享及左右参数不等 2 种模式，在判断 2 个问句相似性问题上，由于 2 个问句结构基本相同，因此选择左右参数共享模式，即利用相同的参数提取 2 个问句的特征，该模式具有训练参数少，训练效率高，避免过拟合等优势。本文在构建文本匹配模型时均使用左右参数共享的孪生网络模型。

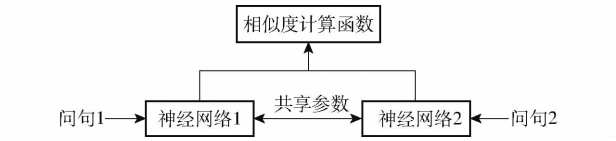

图 2　孪生网络模型结构图

Fig. 2　Schematic of siamese network

(2)BiLSTM 模型文本特征提取

长短期记忆网络(Long short-term memory, LSTM)在标准循环神经网络(Recurrent neural network, RNN)结构基础上，增加了遗忘门、输入门、输出门和细胞状态，能够在有效保存长序列的历史信息同时，规避了标准 RNN 梯度消失和梯度爆炸等问题，具体结构见图 3。

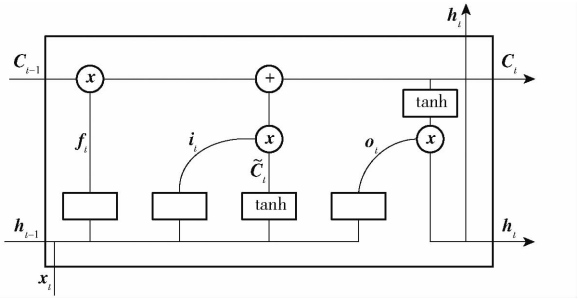

图 3　LSTM 结构图

Fig. 3　Schematic of LSTM

首先遗忘门 $\boldsymbol{f}_t$ 读取上一时刻细胞状态 $\boldsymbol{C}_{t-1}$ 输出 $\boldsymbol{h}_{t-1}$ 和当前细胞状态输入 $\boldsymbol{x}_t$，为 $\boldsymbol{C}_{t-1}$ 输出一个值，用于决定 $\boldsymbol{C}_{t-1}$ 丢弃信息程度。输入门 $\boldsymbol{i}_t$ 根据读取的输入信息，利用 Sigmod 函数确定需要更新的信息，然后利用 tanh 函数对输入进行计算，生成备用更新向量 $\widetilde{\boldsymbol{C}}_t$。综合 $\boldsymbol{f}_t$ 和 $\boldsymbol{i}_t$ 的输出，将上一时刻细胞状态 $\boldsymbol{C}_{t-1}$ 更新为此刻的细胞状态 $\boldsymbol{C}_t$。输出门 $\boldsymbol{o}_t$ 利用 Sigmod 函数确定细胞状态的输出，接着 tanh 函数处理此刻的细胞状态 $\boldsymbol{C}_t$，并将它与 Sigmod 函数的输出相乘，得到此刻细胞状态的输出 $\boldsymbol{h}_t$，具体计算公式为

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_f[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_f) \tag{1}$$

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_i) \tag{2}$$

$$\widetilde{\boldsymbol{C}}_t = \tanh(\boldsymbol{W}_C[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_C) \tag{3}$$

$$\boldsymbol{C}_t = \boldsymbol{f}_t \boldsymbol{C}_{t-1} + \boldsymbol{i}_t \widetilde{\boldsymbol{C}}_t \tag{4}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o[\boldsymbol{h}_{t-1}, \boldsymbol{x}_t] + \boldsymbol{b}_o) \tag{5}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \tanh(\boldsymbol{C}_t) \tag{6}$$

式中　σ——Sigmod 函数

tanh——双曲正切函数

$\boldsymbol{W}_f$、$\boldsymbol{W}_i$、$\boldsymbol{W}_C$、$\boldsymbol{W}_o$——遗忘门、输入门、细胞状态及输出门权重矩阵

$\boldsymbol{b}_f$、$\boldsymbol{b}_i$、$\boldsymbol{b}_C$、$\boldsymbol{b}_o$——遗忘门、输入门、细胞状态及输出门偏置

在中文文本中,词语的前后组合顺序会直接影响文本语义。为了更全面地提取文本特征,深入挖掘词语相关联的上下文信息,利用 BiLSTM 模型提取文本的特征,解决了 LSTM 单向传输问题。BiLSTM 由前向 LSTML 和后向 LSTMR 组成,双向 LSTM 共享权值,所有神经元内的权重参数均相同,具体结构如图 4 所示。

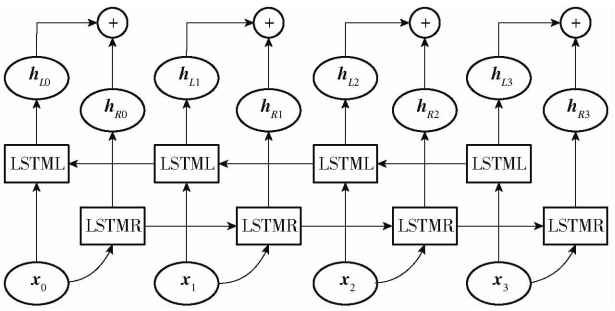

图 4 BiLSTM 结构图

Fig. 4 Schematic of BiLSTM

当输入文本词向量为$\{\boldsymbol{x}_0,\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3\}$时,LSTML 训练得到向量$\{\boldsymbol{h}_{L0},\boldsymbol{h}_{L1},\boldsymbol{h}_{L2},\boldsymbol{h}_{L3}\}$,LSTMR 训练得到向量$\{\boldsymbol{h}_{R0},\boldsymbol{h}_{R1},\boldsymbol{h}_{R2},\boldsymbol{h}_{R3}\}$,最后将生成的向量拼接得到$\{[\boldsymbol{h}_{L0},\boldsymbol{h}_{R0}],[\boldsymbol{h}_{L1},\boldsymbol{h}_{R1}],[\boldsymbol{h}_{L2},\boldsymbol{h}_{R2}],[\boldsymbol{h}_{L3},\boldsymbol{h}_{R3}]\}$。在计算文本相似度任务中,由于$[\boldsymbol{h}_{L3},\boldsymbol{h}_{R3}]$包括了前向与后向的所有特征信息,一般将其作为文本的特征表示。

(3)CNN 模型文本特征提取

由于在文本预处理阶段选择了文本词语共现关系作为文本的特征,原文本已由分词结果表示更改为词语共现对表示,改变了原来的文本语序特征,因此 CNN 模型更适合提取以词语共现对表示文本的语义特征。CNN 模型主要由卷积层和池化层组成,卷积层按照设定尺寸的卷积核对文本向量进行卷积计算得到更高级的特征表示。

池化层包括平均池化和最大池化两类,可将卷积层得到的高阶特征进行稀疏处理以及下采样,在降低维度的同时,保留了关键的特征信息,舍弃了不显著的信息,有效减少了模型参数及数据计算量,降低了模型过拟合。为了遴选辨识度更高的特征,本文选择最大池化方法,即在每个池化窗口内,取每一维的最大特征。

(4)DNN 模型文本特征提取

根据统计的文本词语共现情况,本文构建了词语共现网络,网络中的点是数据集中词频数大于 3 的词语,网络中的边是词语和词语之间的共现关系,边的权重计算公式为

$$f_{i,j}=\frac{n_{i,j}}{\sum_k n_{i,k}} \tag{7}$$

式中 $f_{i,j}$——词语共现对$\langle i,j\rangle$边的权重

$n_{i,j}$——词语共现对$\langle i,j\rangle$的词频

k——与词语 i 组成词语共现对的词语数量

$n_{i,k}$——词语共现对$\langle i,k\rangle$的词频

为了进一步丰富文本特征,使模型更好地适应小规模数据集,本文将 2 个问句词语共现对的最大匹配度和权重作为文本特征。文本特征表示为$\{l_1,l_2,g,w\}$,其中 l_1和 l_2表示 2 个问句包含的词语共现对数量,g 表示 2 个问句的词语共现对最大匹配度,w 表示词语共现对的权重之和。将文本特征输入到 DNN 网络模型中,映射成新空间向量。

2 试验与结果分析

2.1 试验数据

试验数据来源于“中国农技推广 APP”中的问答社区,共有 20 000 对问题组合,并通过人工标记的方法,标注了问题组合对中 2 个问句是否相似。表 1 中为部分样例试验数据,语义相同的 2 个问句,其标签值为 1,共有 9 233 个,占 46.2%;语义不同的,标签值为 0,共 10 767 个,占 53.8%。

表 1 试验数据样例

Tab. 1 Sample of test data

问句 1	问句 2	标签值
这个玉米是什么虫害?要怎么防治?	请问玉米上有芽虫怎么办	0
种猪寄生虫病药物防控技术有哪些?	猪传染性胃肠炎有几种类型?疫苗预防效果怎么样?	0
水稻纹枯病的防治方法有哪些?	如何防治水稻纹枯病?	1
请问葡萄白腐病的症状是什么?	请问葡萄白粉病的症状是什么?	0
请问老师马铃薯如何施肥?	土豆施肥有哪些注意事项?	1
甜瓜黑斑病的发病条件是什么?	什么环境下容易引发甜瓜黑斑病?	1
机械播种有哪些优点呢?	机械播种有哪些注意事项呢?	0
蓝莓茎溃疡病有什么症状?	蓝莓茎得了溃疡病是什么原因?	0

试验数据随机划分为训练数据、验证数据和测试数据 3 部分,其中训练数据占试验数据量的 80%,用于学习文本特征;验证数据占 10%,用于调整模型的超参数,用于对模型的能力进行初步评估;测试数据占 10%,用于评估最终模型的泛化能力。

2.2 评价指标

试验选择精确率、召回率、F1 值、正确率以及时间复杂度作为评价指标,全面衡量模型性能。精确率、召回率、F1 值以及正确率是统计学以及信息检

索相似性判断领域的标准指标。精确率计算的是模型判断相似且实际也相似的句子数量，即正样本与数据集中全部句子数量的比值，代表着模型对正样本结果的查准程度。召回率计算的是正样本与数据集中全部相似句子总数的比值，代表着模型对正样本的查全程度。F1 值同时考虑精确率和召回率，让两者同时达到最高并取得平衡。正确率表示预测正确的结果占全部句子数量的比例。时间复杂度包括对训练时间和测试时间的统计，用于衡量模型的效率，取 10 次试验耗时的平均值作为试验结果。

2.3　参数设置

问句中包含的有效词语使用 128 维的词向量表示，包含词语数量最大为 30 个，包含词语共现对数量最大为 50 个。BiLSTM 模型输出特征维度是 128 维，CNN 模型卷积核数量是 128，孪生神经网络的 2 个子网络共享参数。为了避免过拟合，模型使用 dropout 函数，随机让一些神经元失活，失活率设定为 0.5。

2.4　对比模型

本文模型与近年在文本相似度计算领域常用的 4 种深度学习模型进行比较，对比模型包括基于孪生网络的 MaLSTM[34]、ESIM[35]、ABCNN[36]、CDNN[37]，其中 MaLSTM 模型利用 2 个共享权重的 LSTM 提取文本特征，使用 Word2vec 方法将长度不同的句子转换为长度相同的向量，运用曼哈顿距离计算文本相似度；ESIM 利用 BiLSTM 提取 2 个句子的文本特征，计算 2 个句子的相似性矩阵，对文本特征进行加权，得到新的文本特征表示向量，最后利用全连接层预测文本是否相似；ABCNN 在 CNN 基础上，分别在卷积计算前和池化计算前引入了注意力机制，判断文本相似情况；CDNN 利用 CNN 提取文本特征，利用文本特征相似度矩阵对 2 个句子特征进行加权计算后，判断文本相似性。

2.5　结果分析

表 2 展示了 5 种试验模型针对农业问答短文本数据集的试验结果。本文提出的 Co_BiLSTM_CNN 模型的正确率、精确率、召回率及 F1 值均超过了 94%，较对比模型均有大幅提高。在对比模型中，MaLSTM 模型的 4 项指标均接近 90%。对比正确率和 F1 值可知，Co_BiLSTM_CNN 模型相对于其他模型提高了近 5 个百分点，说明该模型能够较为全面地提取文本特征，模型在相似性计算方面的总体性能较好。此外，以 LSTM 模型框架为基础的 MaLSTM 和 ESIM 的评价指标明显高于以 CNN 模型为基础的 CDNN 和 ABCNN。对比精确率可知，Co_BiLSTM_CNN 仍为最优，说明该模型对于相似性文本的判断比较准确，MaLSTM 和 ESIM 在预测文本为相似文本的表现明显优于 CDNN 和 ABCNN。对比召回率可知，4 种对比试验模型表现较为接近，均超过了 89%，ESIM 召回率达到 91%，但是与 Co_BiLSTM_CNN 模型仍有较大的差距。

表 2　5 种试验模型的试验结果对比

Tab.2　Comparison of five models　　%

试验模型	正确率	精确率	召回率	F1 值
MaLSTM	89.50	89.40	89.60	89.50
ESIM	87.85	84.00	91.00	87.36
ABCNN	82.20	76.08	89.59	82.28
CDNN	86.65	82.80	89.71	86.12
Co_BiLSTM_CNN	94.15	94.08	94.17	94.12

为了进一步验证词语共现特征对于本文相似性计算的有效性，将 4 种对比模型进行改进，融合了文本词语共现特征。图 5 展示了 4 种对比试验模型原模型与融合了词语共现特征模型在相似性计算正确率方面的对比结果。如图 5 所示，针对农业短文本数据集，本文提出的融合词语共现特征的文本表示方式在 4 种试验模型的相似性计算正确率明显高于原模型，其中 MaLSTM 正确率最高，达 92.4%。特别对于以 CNN 为基础的试验模型，正确率提升效果更加明显，最高增幅达到 10.2 个百分点。4 种对比模型在融合了词语共现特征后，正确率虽明显提高，但仍低于本文提出的 Co_BiLSTM_CNN 模型。

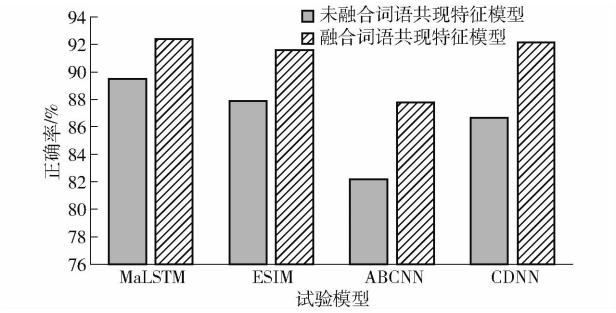

图 5　不同模型文本相似性计算正确率对比

Fig.5　Comparison of accuracy by different models

为了验证试验模型对不同规模数据集的适用性，本文将 Co_BiLSTM_CNN 与 MaLSTM、CDNN，以及对应的融入词语共现特征的 CO_MaLSTM 和 CO_CDNN 模型，分别在 10 000、15 000、20 000 组问句对数据集中进行试验。图 6 展示了 5 种试验方法在不同规模数据集下的相似度计算正确率。由图 6 可知，全部试验模型在较大规模数据集中取得的试验结果均优于小规模数据集，Co_BiLSTM_CNN 在不同的规模数据集中均取得了最优的试验结果，特别在小规模数据集中，Co_BiLSTM_CNN 优势更为突出，体现了该模型具有较强的鲁棒性。融合了词语共现

特征的 3 个试验模型在小规模数据集中的正确率仍在 90% 左右,明显高于未融合词语共现特征的 MaLSTM 和 CDNN。

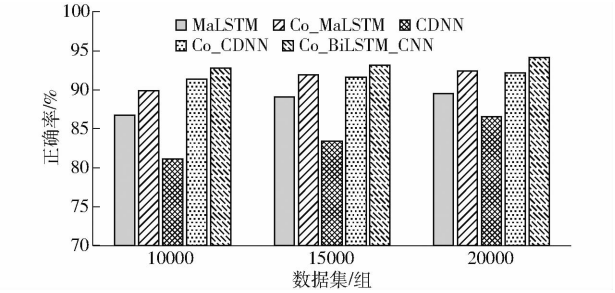

图 6 不同数据集的文本匹配正确率

Fig. 6 Text matching accuracy for different datasets

表 3 展示了 5 种试验模型在 20 000 组问答对数据集中的训练和测试所用时间。CDNN 和 Co_CDNN 在训练时间上优势明显,MaLSTM、Co_MaLSTM 训练时间较长,融合了 CNN、BiLSTM 和 DNN 多种模型的 Co_BiLSTM_CNN 由于结构组成复杂,导致其训练时间最长。在实际应用中,由于文本匹配模型的训练一般采取离线方式运行,因此模型对时间复杂度要求不高,更注重相似性判断的正确率及结果反馈时间。此外,融合了词语共现特征的 Co_CDNN 和 Co_MaLSTM 的训练时间与基础 CDNN、MaLSTM 训练时间基本相同,可见在文本相似性判断任务中,融合词共现特征对模型时间复杂度影响较小。

表 3 各试验模型的离线训练时间

Tab. 3 Offline training time comparison of models

试验模型	MaLSTM	Co_MaLSTM	CDNN	Co_CDNN	Co_BiLSTM_CNN
离线训练时间/s	4 195	4 233	415	421	5 631

3 结论

(1)Co_BiLSTM_CNN 模型满足问句相似性判断实际需要。Co_BiLSTM_CNN 能够多角度提取文本特征,在农业问答问句判断方面具有良好的试验效果,相似性判断正确率明显提高,正确率达到 94.15%,并且在小规模数据集上的表现仍然良好。

(2)利用文本词语共现关系增强文本特征,可在保证时间效率的基础上,有效提高相似性判断的正确率,有效解决特征稀疏文本相似性判断难的问题。该方法在多个试验模型中针对不同规模的数据均取得了最优的试验效果,具有广泛的适用性。

参 考 文 献

[1] 中国农技推广信息平台[DB/OL]. [2021-04-21] http://njtg.nercita.org.cn/user/index.shtml.

[2] GOMMA W H, FAHMY A A. A survey of text similarity approaches[J]. International Journal of Computer Applications, 2013, 68(13): 13-18.

[3] 庞亮,兰艳艳,徐君,等. 深度文本匹配综述[J]. 计算机学报, 2017, 40(4): 985-1003.
PANG Liang, LAN Yanyan, XU Jun, et al. A survey on deep text matching[J]. Chinese Journal of Computers, 2017, 40(4): 985-1003. (in Chinese)

[4] 王寒茹,张仰森. 文本相似度计算研究进展综述[J]. 北京信息科技大学学报(自然科学版),2019,34(1):68-74.
WANG Hanru, ZHANG Yangsen. A survey on research progress of text similarity calculation[J]. Journal of Beijing Information Science & Technology University, 2019, 34(1): 68-74. (in Chinese)

[5] MAIPRADIT R, HATA H, MATSUMOTO K. Sentiment classification using N-gram inverse document frequency and automated machine learning[J]. IEEE Software, 2019, 36(5): 65-70.

[6] AYUB M, GHAZANFAR M A, MAQSOOD M, et al. A Jaccard base similarity measure to improve performance of CF based recommender systems[C]// 2018 International Conference on Information Networking (ICOIN). IEEE Computer Society, 2018.

[7] 田堃,柯永红,穗志方. 基于语义角色标注的汉语句子相似度算法[J]. 中文信息学报,2016,30(6):126-132.
TIAN Kun, KE Yonghong, SUI Zhifang. Chinese sentence similarity computing based on semantic roles annotation[J]. Journal of Chinese Information Processing, 2016, 30(6): 126-132. (in Chinese)

[8] 张佩云,陈恩红,谢荣见,等. 基于元数据与领域概念树的文本相似度计算[J]. 系统工程与电子技术,2014,36(3):591-597.
ZHANG Peiyun, CHEN Enhong, XIE Rongjian, et al. Computation of document similarity based on metadata and domain concept tree[J]. Systems Engineering and Electronics, 2014, 36(3): 591-597. (in Chinese)

[9] KAUR H, MAINI R. Granularity-based assessment of similarity between short text strings[C]// Proceedings of the Third International Conference on Microelectronics, Computing and Communication Systems. Springer, Singapore, 2019: 91-107.

[10] 李小涛,游树娟,陈维. 一种基于词义向量模型的词语语义相似度算法[J]. 自动化学报,2020,46(8):1654-1669.
LI Xiaotao, YOU Shujuan, CHEN Wei. An algorithm of semantic similarity between words based on word single-meaning embedding model[J]. Acta Automatica Sinica, 2020, 46(8): 1654-1669. (in Chinee)

[11] MIKOLOV T, CORRADO G, CHEN K, et al. Efficient estimation of word representations in vector space[C]// Proceedings of the International Conference on Learning Representations (ICLR 2013), 2013.

[12] AYYADEVARA V K. Word2vec[M]// Pro Machine Learning Algorithms, 2018.

[13] 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 (EMNLP), 2014: 1532-1543.

[14] LIU B, NIU D, WEI H, et al. Matching article pairs with graphical decomposition and convolutions[C]// Proceedings of the

57th Annual Meeting of the Association for Computational Linguistics, 2019: 6284 – 6294.
[15] RAO J, LIU L, YI T, et al. Bridging the gap between relevance matching and semantic matching for short text similarity modeling[C] // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), 2019: 5370 – 5381.
[16] LAN W, XU W. Neural network models for paraphrase identification, semantic textual similarity, natural language inference, and question answering[J]. arXiv Preprint:1806.04330v2, 2018.
[17] ZHANG X, ZHAO J, LECUN Y. Character-level convolutional networks for text classification[C] // Advances in Neural Information Processing Systems, 2015: 649 – 657.
[18] KIM Y. Convolutional neural networks for sentence classification[J]. arXiv Preprint:1408.5882, 2014.
[19] DOS S C, GATTI M. Deep convolutional neural networks for sentiment analysis of short texts[C] // Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers, 2014: 69 – 78.
[20] LIU P, QIU X, HUANG X. Recurrent neural network for text classification with multi-task learning[J]. arXiv Preprint:1605.05101, 2016.
[21] PALANGI H, DENG L, SHEN Y, et al. Deep sentence embedding using long short-term memory networks: analysis and application to information retrieval[J]. IEEE/ACM Transactions on Audio Speech & Language Processing, 2015, 24(4):694 – 707.
[22] WAN S, LAN Y, GUO J, et al. A deep architecture for semantic matching with multiple positional sentence representations[J]. arXiv Preprint:1511.08277v1, 2015.
[23] MNIH V, HEESS N, GRAVES A. Recurrent models of visual attention[C] // Advances in Neural Information Processing Systems, 2014: 2204 – 2212.
[24] 崔晓晖,师栋瑜,陈志泊,等. 基于 Spark 框架 XGBoost 的林业文本并行分类方法研究[J]. 农业机械学报,2019,50(6):280 – 287.
CUI Xiaohui, SHI Dongyu, CHEN Zhibo, et al. Parallel forestry text classification technology based on XGBoost in Spark framework[J]. Transactions of the Chinese Society for Agricultural Machinery, 2019, 50(6):280 – 287. (in Chinese)
[25] 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 Ming, DONG Cuicui, DONG Qiaoxue, 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. (in Chinese)
[26] 冯帅,许童羽,周云成,等. 基于深度卷积神经网络的水稻知识文本分类方法[J]. 农业机械学报,2021,52(3):257 – 264.
FENG Shuai, XU Tongyu, ZHOU Yuncheng, et al. Rice knowledge text classification based on deep convolution neural network[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(3):257 – 264. (in Chinese)
[27] 陈志泊,李钰曼,许福,等. 基于 TextRank 和簇过滤的林业文本关键信息抽取研究[J]. 农业机械学报,2020,51(5):207 – 214,172.
CHEN Zhibo, LI Yuman, XU Fu, et al. Key information extraction of forestry text based on TextRank and clusters filtering[J]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(5):207 – 214, 172. (in Chinese)
[28] 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 Mingyue, WU Huarui, ZHU Huaji. 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. (in Chinese)
[29] 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. Recognition of Chinese agricultural diseases and pests named entity with joint radical embedding and self-attention mechanism[J]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(Supp.2):335 – 343. (in Chinese)
[30] 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 Chinese agricultural text rased on attention mechanism[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(1):185 – 192. (in Chinese)
[31] 郑丽敏,齐珊珊,田立军,等. 面向食品安全事件新闻文本的实体关系抽取研究[J]. 农业机械学报,2020,51(7):244 – 253.
ZHENG Limin, QI Shanshan, TIAN Lijun, et al. Entity relation extraction of news texts for food safety events[J]. Transactions of the Chinese Society for Agricultural Machinery, 2020, 51(7):244 – 253. (in Chinese)
[32] 俞士汶,段慧明,朱学锋,等. 北京大学现代汉语语料库基本加工规范[J]. 中文信息学报,2002(5):49 – 64.
YU Shiwen, DUAN Huiming, ZHU Xuefeng, et al. The basic processing of contemporary Chinese corpus at Peking University specification[J]. Journal of Chinese Information Processing, 2002(5):49 – 64. (in Chinese)
[33] CHENG Z X, DAN L. Chinese text summarization algorithm based on Word2vec[J]. Journal of Physics Conference Series, 2018, 976.
[34] THYAGARAJAN A. Siamese recurrent architectures for learning sentence similarity[C] // Thirtieth AAAI Conference on Artificial Intelligence. AAAI Press, 2016.
[35] CHEN Q, ZHU X, LING Z, et al. Enhanced LSTM for natural language inference[J]. arXiv preprint arXiv: 1609.06038, 2016.
[36] YIN Wenpeng, SCHÜTZE H, XIANG B, et al. ABCNN: attention-based convolutional neural network for modeling sentence pairs[J]. Transactions of the Association for Computatio nal Linguistics, 2016, 4:259 – 272.
[37] ALIAKSEI S, ALESSANDRO M. Learning to rank short text pairs with convolutional deep neural networks[C] // Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR'15), 2015:373 – 382.