doi:10.6041/j.issn.1000-1298.2025.05.037

基于特征增强的农业短文本语义智能匹配方法研究

金　宁[1]　郭宇峰[1,2]　渠丽娜[1]　缪祎晟[2,3]　吴华瑞[2,3]

(1.沈阳建筑大学计算机科学与工程学院,沈阳 110168;2.国家农业信息化工程研究中心,北京 100097;
3.农业农村部农业信息化技术重点实验室,北京 100097)

摘要:针对农业短文本数据特征词语少、语义特征稀疏、冗余度高、价值密度低等问题,构建了一种利用多尺度通道注意力算法融合多语义特征的语义匹配模型 Font_MBAFF,以提升农业短文本的语义匹配性能。首先利用汉字偏旁部首和四角号码丰富短文本特征;然后利用多尺度卷积核通道注意力加权网络 MSCN 和基于多头自注意力的双向长短期记忆网络 Multi_SAB 分别从空间和时间提取语义特征;最后利用文本注意力融合机制 TEXTAFF 对多种特征进行智能融合。试验结果表明,Font_MBAFF 模型可有效弥补短文本特征词少的不足,优化文本特征提取及特征融合,语义匹配正确率达到 96.42%,与 MaLSTM、BiLSTM、BiLSTM_Self-attention、TEXTCNN_Attention、Sentence-BERT 等 5 种语义匹配模型相比优势明显,正确率至少高 2.07 个百分点。

关键词:农业短文本;语义匹配;字形特征表示;多特征融合

中图分类号:TP183　　**文献标识码**:A　　**文章编号**:1000-1298(2025)05-0395-10　　**OSID**:

Exploration of Intelligent Semantic Matching Technique for Agricultural Short Texts Utilizing Feature Enhancement

JIN Ning[1]　GUO Yufeng[1,2]　QU Li'na[1]　MIAO Yisheng[2,3]　WU Huarui[2,3]

(1. *School of Computer Science and Engineering, Shenyang Jianzhu University, Shenyang* 110168, *China*
2. *National Engineering Research Center for Information Technology in Agriculture, Beijing* 100097, *China*
3. *Key Laboratory of Agricultural Information Technology, Ministry of Agriculture and Rural Affairs, Beijing* 100097, *China*)

Abstract: A deep learning model Font_MBAFF was proposed for the task of text similarity calculation, which was mainly applied to the matching of question pairs in Chinese agricultural short texts. In order to solve the problems of sparse semantic features and inadequate understanding of specialized vocabulary in agricultural short texts, it was firstly optimized in the feature representation stage. By introducing the unique font features of Chinese characters to expand the features, including side radicals and four corner numbers, thus enriching the semantic representation of features. In the feature extraction layer, the multi-scale convolution attention channel weighted network MSCN and the bidirectional long short-term memory network Multi_SAB based on multi-head self-attention mechanism were combined respectively, so that the model can further optimize the feature extraction from the spatial and temporal relationship sequences of semantic features. Finally, TEXTAFF, an improved attention fusion mechanism for text, was used in the intelligent fusion stage of features. The experimental results indicated that the Font_MBAFF model can effectively compensate for the lack of feature words in short texts, optimizing text feature extraction and feature fusion. The accuracy of semantic matching reached 96.42%. Compared with five other semantic matching models, including MaLSTM, BiLSTM, BiLSTM_Self-attention, TEXTCNN_Attention, and Sentence-BERT, the Font_MBAFF model demonstrated significant advantages, achieving a correctness rate that was at least 2.07 percentage points higher. Furthermore, the model proved resilient in experiments with datasets of different sizes, showing rapid response times during testing. Font_MBAFF deep learning model exceled at determining the similarity of Chinese agricultural short texts.

Key words: agricultural short text; semantic matching; glyph feature representation; multi-feature fusion

收稿日期:2024-03-05　修回日期:2024-05-28

基金项目:辽宁省教育厅基础研究项目(LJKQZ20222458)和辽宁省科技计划联合计划项目(2024-MSLH-399)

作者简介:金宁(1989—),男,副教授,博士,主要从事农业智能系统研究,E-mail:jinning21@126.com

通信作者:吴华瑞(1975—),男,研究员,主要从事农业智能系统与物联网研究,E-mail:wuhr@nercita.org.cn

0 引言

数字农业是我国由农业大国迈向农业强国的必经之路,也是学术界当前关注的热点问题[1-2]。“中国农技推广”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[3]。这些数据信息量巨大、更新迭代快、冗余度高,多呈无秩化、碎片化状态,难以有效利用。因此,目前存在重复提问多、问题回复时效性差、自动问答准确率低等问题[4-6]。智能的文本语义匹配方法可有效解决上述问题,进一步提高平台信息服务效率及智能化水平。

文本语义匹配任务一般利用 VSM[7]、TF-IDF[8]、词袋模型[9]等方法构建特征工程,然后使用余弦相似度[10]、编辑距离[11]、曼哈顿距离[12]等度量方法计算语义相似度。谭静[13]通过非零权值并集向量空间模型和非零权值基准向量空间模型构建临时向量,来改进传统向量空间模型在相似度计算时的高维度和低效率问题。近年来,深度学习技术在文本语义匹配任务中得到广泛应用[14-17]。为了更精准提取关键语义特征,研究人员将注意力机制引入文本匹配任务,为文本各部分语义特征赋予了不同的注意力权重,使得模型能够更好地捕捉关键信息[18]。基于深度学习的语义匹配方法解决了自动提取语义特征问题,为开展多角度语义特征提取,精细化特征融合研究[19-20]奠定了研究基础。

预训练模型可将外部知识引入到语义匹配任务中,弥补语义特征不足问题[21]。基于 Transformer[22]的 Sentence-BERT[23]将句子向量化后通过平均池化层后利用余弦相似度计算结果。ALBERT[24]实现了对 BERT 的轻量化处理,优化了训练及预测的时间。自我集成 ALBERT 模型[25],利用数据增强增加数据集的大小,通过半监督学习方式提高了模型的学习效率,利用自我集成方法提高模型性能。基于预训练模型方法对通用文本的语义理解上有显著优势,但对农业领域短文本数据的适用性未得到有效验证,此外该方法还存在硬件要求高、部署难度大、训练时间较长等问题。

当前,研究人员在农业文本挖掘领域利用深度学习技术开展了文本分类[26-27]、实体关系抽取[28-29]、实体识别[30-32]等研究,为语义匹配提供了可行性和参考。但在语义特征提取全面性、多语义特征融合的智能性等方面有待进一步优化。为了实现农业短文本智能匹配,本文将从短文本特征增强、多角度语义特征提取以及多特征智能融合 3 个层面进行优化研究,提出一种基于多尺度通道注意力,融合多语义特征的文本相似度计算模型 Font_MBAFF。

1 多特征融合的语义智能匹配模型

语义匹配模型包括特征增强层、特征提取层、特征融合层、相似性度量层 4 部分,具体结构如图 1 所示。在特征增强层增加了中文字形特征模块 Font,在正常分词基础上,增加字形和四角号码,弥补短文本语义特征稀疏问题。特征提取层利用多尺度卷积通道注意力加权网络(Multi scale convolution-channel-attention-weighted net, MSCN)以及基于多头自注意力机制的双向长短期记忆网络(Multi-head self-attention bilstm, Multi_SAB)分别从空间序列和时间序列提取语义特征。在特征融合层将用于图像特征融合的特征融合算法(Attentional feature fusion, AFF)改进为适用于文本语义特征融合的特征融合算法 TEXTAFF。

1.1 短文本特征增强

(1)文本预处理

首先利用 Jieba 分词模块对文本进行分词,利用加载停用词表方式,去除噪声信息;利用加载专业词库表方式,提高对专业领域词语分词的准确率,然后利用 Word2Vec 将文本向量化,计算出每个词语的向量表示。Word2Vec 计算的词向量可以捕捉词语之间的相似性和语义关系,解决了语义间相互孤立的问题,从而帮助模型更好地理解文本语义。

(2)短文本特征增强

现代汉字由象形文字演变而来[33],字形特征是文本特征的重要组成部分。模型引入中文字形特征进一步丰富文本特征表示,利用汉字偏旁部首和汉字四角号码来增强文本特征表示,弥补短文本特征不足的问题。利用 BS4 爬虫包对汉字查询平台汉典[34]中的康熙字典进行数据爬取,获取 27 973 个常用简体汉字对应的偏旁部首以及四角号码,再通过爬虫遍历中华字典平台[35]中的汉字字形结构,形成完整的汉字字形字典,其主要流程如图 2 所示。

例如按照字形对"病疫"进行拆分,“病”拆分为“疒、丙”,“疫”拆分为“疒、殳”。汉字的四角号码是通过将单个汉字拆分为左上、右上、左下、右下 4 个部分,再按照顺序依次编码[36],“病”的四角号码为“00127”、“疫”的四角号码为“00147”,2 个字形相近的汉字会有着相近的四角号码,通过读取汉字字形字典,将文本转换为分别由拆字字形以及四角

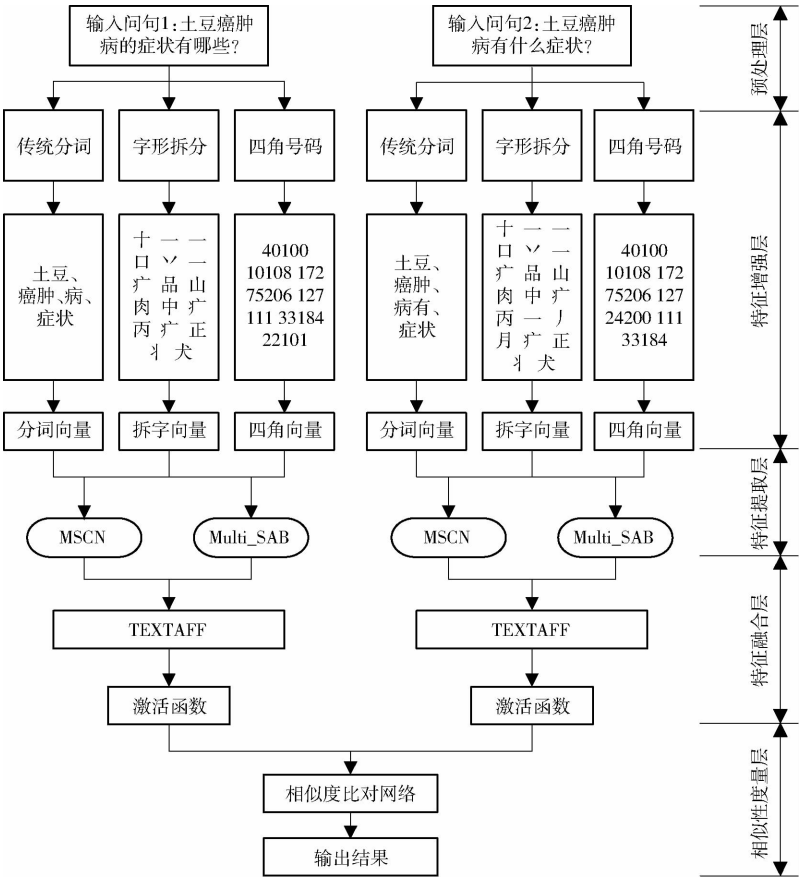

图 1　语义匹配模型结构图

Fig. 1　Schematic of model

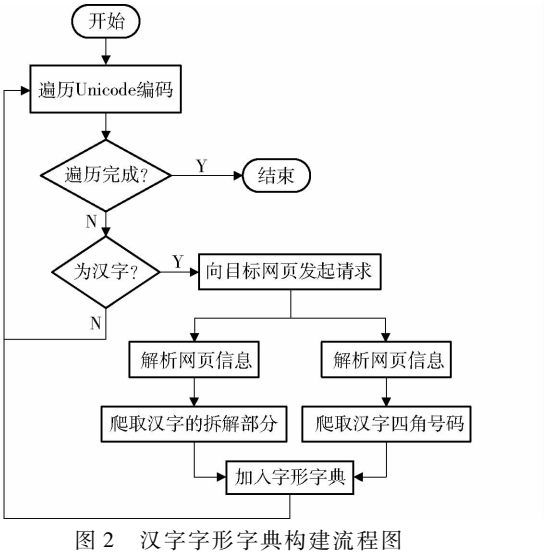

图 2　汉字字形字典构建流程图

Fig. 2　Chinese character dictionary construction process

号码组成的句子。

由于拆字后的字形结构可能为单个字，与原有文本输入重复。例如“畜”拆分为“玄、田”，拆分后的部分在由分词组成的句子及由单字组成句子中会同时存在，在进行文本向量化的过程中会出现同一个字有 2 个不同的编码，影响语义特征表达。因此将扩展后的词向量、偏旁部首向量以及四角号码向量进行两两正交化处理，其公式为

$$\boldsymbol{V}_{\text{words}} = f_{\text{vec}}(s_{\text{words}} - s_{\text{words}} \cap s_{\text{character}}) \cup f_{fc}(s_{\text{words}} \cap s_{\text{character}}) \tag{1}$$

$$\boldsymbol{V}_{ff} = f_{\text{vec}}(s_{ff} - s_{ff} \cap s_{\text{character}}) \cup f_{fc}(s_{ff} \cap s_{\text{character}}) \tag{2}$$

$$\boldsymbol{V}_{fc} = f_{fc}(s) \tag{3}$$

式中　f_{vec}——句子向量化转换

s——输入的句子

f_{fc}——四角号码转换

s_{ff}——拆字字形句子

s_{words}——经过了分词以及去停用词后由词语组成的句子

$s_{\text{character}}$——由分离的单字组成的句子

式(1)通过从分词集合 s_{words} 中剔除与单字集合 $s_{\text{character}}$ 的交集，以此消除分词句子中的重复单字，并随后补充利用四角号码编码规则编码的单字编码。类似地，式(2)通过从拆字字形集合 s_{ff} 中剔除了与单字集合 $s_{\text{character}}$ 的交集来消除其中重复的单字，同样再最后补充利用四角号码编码规则编码的单字编码。通过这种处理方式成功地构建了 2 个相互正交的特征向量：分词句子向量 $\boldsymbol{V}_{\text{words}}$ 和拆字字形特征向量 $\boldsymbol{V}_{ff}$。同时利用式(3)生成了四角号码句子向量 $\boldsymbol{V}_{fc}$。这 3 个特征向量随后被输入到神经网络模型的特征提取层中，以供进一步分析和处理。

1.2 多角度语义特征提取

Font_MBAFF 模型提出一种以孪生神经网络结构为基础,对文本时序特征及空间序列特征联合提取方法,特征提取流程如图 3 所示。利用多尺度卷积通道注意力加权网络 MSCN,提取不同视野下,文本多粒度语义特征;利用基于多头自注意力机制的双向长短期记忆网络 Multi_SAB 提取不同时序维度的语义特征,进一步提升模型对文本语义特征提取的全面性和深度。

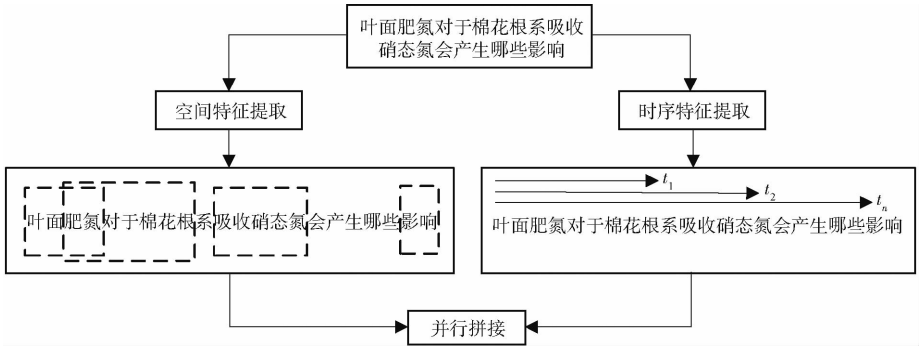

图 3 空间时序特征提取图

Fig. 3 Spatial-temporal feature extraction diagram

1.2.1 孪生神经网络

孪生神经网络包括左右 2 个结构相同、参数共享的神经网络,具有结构复杂度低,训练效率高的特点,是文本语义匹配任务中常用的网络架构。Font_MBAFF 模型分别利用 MSCN 及 Multi_SAB 进行特征提取,将输入的文本映射至新的特征空间中,形成对应的语义特征向量,并将余弦相似算法作为比对策略,衡量向量相似度,具体结构如图 4 所示。

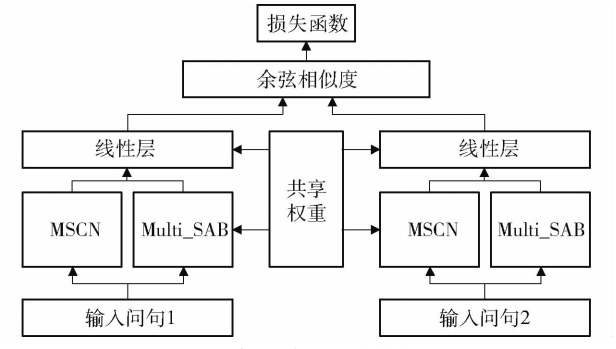

图 4 孪生神经网络结构图

Fig. 4 Siamese neural network architecture diagram

模型将对比损失函数作为优化目标。当真实标签 Z 为 1,2 个输入文本相似时,若计算得到的余弦相似度 D_W 接近预设的相似度阈值 D_{margin},则减小损失值;反之,若 D_W 远离 D_{margin},则增大损失值作为惩罚。当真实标签 Z 为 0,2 个输入文本不相似时,损失函数的计算方式同理。这种损失函数设计有助于模型更好地捕捉文本间的相似度差异,并提高匹配的准确性。损失函数计算公式为

$$L_{\text{loss}} = \frac{1}{2}(1-Z)D_W^2 + \frac{1}{2}Z(\max(0, D_{\text{margin}} - D_W))^2 \tag{4}$$

式中 L_{loss}——孪生神经网络对比损失函数

$\max(\,)$——取最大值函数

1.2.2 多尺度卷积通道注意力加权网络 MSCN

MSCN 由多尺度卷积层和通道注意力 TEXT_MSCAM 层 2 部分组成,网络结构模型如图 5 所示。多尺度卷积层能够提取不同上下文范围的局部文本特征,以便获得多视野范围内的特征表达,并通过堆叠多尺度特征向量方式,形成更加丰富的文本特征表示,其公式为

$$\boldsymbol{x}_{1:n} = \boldsymbol{x}_1 \oplus \boldsymbol{x}_2 \oplus \cdots \oplus \boldsymbol{x}_n \tag{5}$$

$$c_{i,j} = f(\boldsymbol{w}\boldsymbol{x}_{i:i+h_j-1} + \boldsymbol{b}) \tag{6}$$

$$\boldsymbol{c}_j = (c_{1,j}, c_{2,j}, \cdots, c_{n-h_j+1,j}) \tag{7}$$

式中 $\boldsymbol{x}_i$——词向量

$\boldsymbol{x}_{1:n}$——词向量矩阵

$\boldsymbol{w}$——卷积滤波器

$\boldsymbol{b}$——偏置

$\boldsymbol{c}_j$——多尺度卷积核输出

h_j——卷积核尺寸

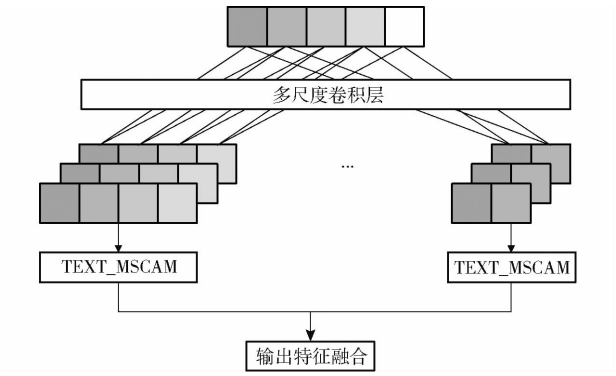

图 5 MSCN 网络模型结构图

Fig. 5 Schematic of MSCN

其中式(5)中的⊕表示将 n 个 $1\times k$ 的词向量拼接为一个 $n\times k$ 矩阵,k 为词向量维度。将此矩阵

作为句子特征表示形式。通过式(6)对输入特征矩阵进行卷积运算，通过式(7)中的多尺度卷积核函数计算得到不同尺度下的文本特征

$$C_0 = f_{tm}(\mathrm{MAX}(\boldsymbol{c}_1), \mathrm{MAX}(\boldsymbol{c}_2), \cdots, \mathrm{MAX}(\boldsymbol{c}_j)) \tag{8}$$

式中　MAX——最大池化函数

$f_{tm}(\,)$——多尺度通道注意力模块

式(8)通过最大池化以降低过拟合以及增加模型训练的效率。将经过了多尺度卷积的特征代入其中，得到了最后输出的特征 C_0。

为了更加精准赋予不同尺度特征的权重，模型提出了一种基于 SENet 与注意力机制的多尺度通道注意力(Text multi-scale channel attention, TEXT_MSCAM)。其核心思想为通过改变空间池大小的方式在多个不同特征尺度上实现通道关注，同时为了让模型保持轻量级，将局部上下文特征添加到注意力模块内部的全局上下文中，使 TEXT_MSCAM 能够同时感受到全局和局部通道上下文特征，从而更有效地捕捉和处理不同尺度的特征信息。这种设计有助于提升模型对不同尺度语义特征的识别能力，具体公式为

$$G(\boldsymbol{X}) = \frac{1}{1 \times W} \sum_{i=1}^{1} \sum_{j=1}^{W} \boldsymbol{X}[:, i, j] \tag{9}$$

$$L(\boldsymbol{X}) = B(\mathrm{PWConv2}(\delta(\mathrm{PWConv1}(\boldsymbol{X})))) \tag{10}$$

$$M(\boldsymbol{X}) = \sigma(L(\boldsymbol{X}) \cup g(\boldsymbol{X})) \tag{11}$$

$$\boldsymbol{X}' = \boldsymbol{X} \otimes M(\boldsymbol{X}) \tag{12}$$

式中　W——特征向量尺寸

$\boldsymbol{X}$——输入向量

$G(\,)$——平均池化

PWConv1(　)、PWConv2(　)——点卷积

$\delta(\,)$——激活函数 ReLU

$B(\,)$——BatchNorm 归一化处理

$L(\,)$——局部通道注意力

$g(\,)$——全局通道注意力

$\sigma(\,)$——Sigmoid 函数

$M(\,)$——全局局部通道注意力加权

$\boldsymbol{X}'$——加权后输出

式(9)表示全局平均池化，其中 $1 \times W$ 为一维特征向量尺寸，式(10)表示局部通道注意力，PWConv1 以及 PWConv2 表示尺寸为 1×1 的点卷积，通过点积将通道数目恢复为输入特征向量的通道数。全局通道注意力则是在局部通道注意力的基础上在特征向量输入前先经过全局平均池化 $G(\boldsymbol{X})$。通过式(11)对全局通道注意力的输出以及局部通道注意力的输出进行计算得出向量加权，∪表示输出向量并行拼接。最后将输入向量 $\boldsymbol{X}$ 通过式(12)得到输出向量 $\boldsymbol{X}'$，其中⊗表示 2 个特征向量对应元素相乘。

TEXT_MSCAM 内部结构如图 6 所示，图中的 $C \times 1$ 表示经过全局池化后的特征矩阵的形状，$(C/r) \times 1$、$(C/r) \times W$、$C \times 1$、$C \times W$ 表示经过一维点积后的特征矩阵的形状，其中 r 表示一维点积输出的缩放倍数。

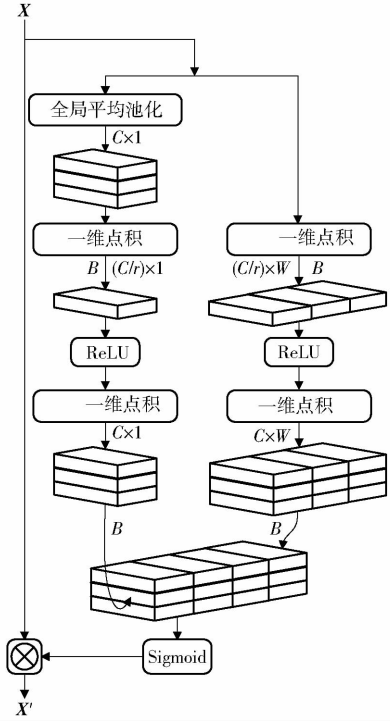

图 6　TEXT_MSCAM 模型结构图

Fig. 6　Schematic of TEXT_MSCAM

1.2.3　基于多头自注意力机制的双向长短期记忆网络 Multi_SAB

Multi_SAB 由双向长短期记忆神经网络和多头自注意力机制模块 2 部分组成，结构如图 7 所示。

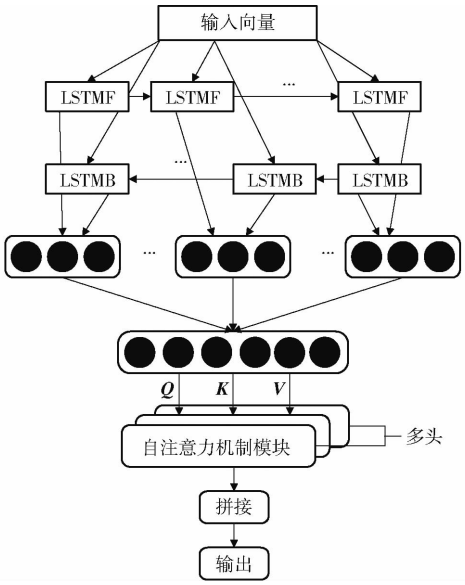

图 7　Multi_SAB 网络模型结构图

Fig. 7　Schematic of Multi_SAB

模型使用 LSTM 来提取时序特征。在循环神经网络的基础上,LSTM 添加遗忘门、输入门、输出门和细胞状态,这使得神经网络能够有效保存长序列的历史信息,缓解了 RNN 所引起的梯度爆炸以及梯度消失等问题,LSTM 网络模型结构图如图 8 所示。

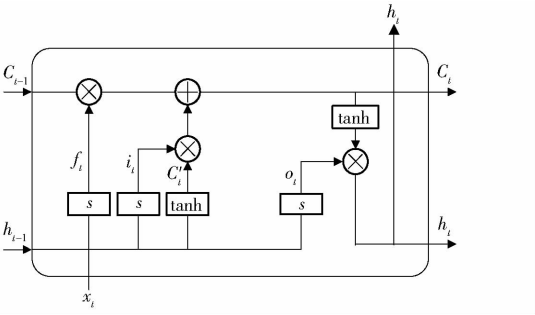

图 8 长短记忆神经网络模型结构图

Fig. 8 Schematic of LSTM

其相关公式为

$$f_t = \sigma(\boldsymbol{W}_f[h_{t-1}, x_t] + \boldsymbol{b}_f) \tag{13}$$

$$i_t = \sigma(\boldsymbol{W}_i[h_{t-1}, x_t] + \boldsymbol{b}_i) \tag{14}$$

$$C'_t = \tanh(\boldsymbol{W}_c[h_{t-1}, x_t] + \boldsymbol{b}_c) \tag{15}$$

$$C_t = f_t C_{t-1} + i_t C'_t \tag{16}$$

$$o_t = \sigma(\boldsymbol{W}_o[h_{t-1}, x_t] + \boldsymbol{b}_o) \tag{17}$$

$$h_t = o_t \tanh(C_t) \tag{18}$$

式中 x_t——输入数据

$\tanh(\,)$——双曲正切函数

f_t、i_t、c_t、o_t——遗忘门、输入门、细胞状态、输出门

$\boldsymbol{W}_f$、$\boldsymbol{W}_i$、$\boldsymbol{W}_c$、$\boldsymbol{W}_o$——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

C'_t——单元状态更新值

h_t——隐藏节点

通过式(13)将输入的数据 x_t以及上一个隐藏节点 h_{t-1}代入,计算输出得到遗忘门 f_t,LSTM 通过遗忘门来丢弃细胞状态中的一些数据信息,再利用式(14)和式(15)来计算得出输入门 i_t及单元状态的更新值 C'_t,并通过式(16)来计算当前的单元状态 C_t,最后使用式(17)和式(18)来计算得到输出门 o_t以及当前隐藏节点 h_t。

在长短期记忆网络 LSTM 的基础上引入 2 个不同方向的独立 LSTM 组成双向长短期记忆网络 BiLSTM,一个从前往后(前向 LSTMF),一个从后往前(后向 LSTMB),分别对其进行编码,有效地捕捉到文本前后双向文本特征。BiLSTM 基本结构:LSTMF 输入为(t_1, t_2, …, t_n),LSTMB 输入为(t_n, t_{n-1}, …, t_1),经过隐藏层后输出得到包含了前后向特征信息$\{h_1, h_2, \cdots, h_n\}$,其网络模型结构图如图 9 所示。

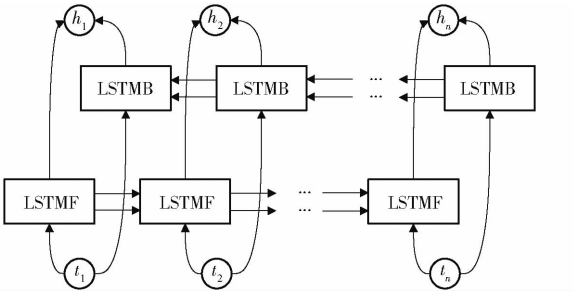

图 9 双向长短记忆神经网络模型结构图

Fig. 9 Schematic of BiLSTM

在 BiLSTM 基础上添加多头自注意力模块(Multi-head self-attention, Multi_SA)对模型进行进一步优化。将通过 BiLSTM 编码层的输出接入多头自注意力模块,得到多组注意力结果,然后将这些结果进行拼接和线性投影得到最终输出。利用多注意力头在嵌入维度上对文本语义进行分割,并通过自注意力提取不同注意力头中的语义向量特征,可以有效地学习句子内部依赖关系,捕捉句子中更重要的特征信息,其具体公式为

$$\text{Attention}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_K}}\right)\boldsymbol{V} \tag{19}$$

$$\text{Multihead}(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) = \text{Concat}(\text{head}_1, \text{head}_2, \cdots, \text{head}_n)\boldsymbol{A}^0 \tag{20}$$

$$\text{head}_i = \text{Attention}(\boldsymbol{Q}\boldsymbol{A}_i^Q, \boldsymbol{K}\boldsymbol{A}_i^K, \boldsymbol{V}\boldsymbol{A}_i^V) \tag{21}$$

式中 $\text{Attention}(\,)$——注意力机制

$\text{softmax}(\,)$——概率分布激活函数

d_k——$\boldsymbol{K}$ 向量维度 $\boldsymbol{A}^0$——权重矩阵

$\text{Multihead}(\,)$——多头注意力

$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$——查询、键、值矩阵

$\text{Concat}(\,)$——注意力头拼接

head_i——注意力头输出

式(19)计算得到输入矩阵在 Self - attention 输出,其中 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 均来自同一输入,式(20)引入多头注意力机制来连接输出,与矩阵 $\boldsymbol{A}^0$相乘得到最后的结果。式(21)为每个注意力头输出函数。

1.3 多特征智能融合处理

特征融合是指来自相同或是不同的网络层输出特征的组合,一般为简单的线性操作,如逐个元素的相加、相乘或是特征的串并联拼接等。模型提出了一种注意力特征融合模型 TEXTAFF,其网络结构如图 10 所示,首先将 2 个特征向量先进行前向拼接融合,再利用 TEXT_MSCAM 计算得到全局局部通道注意力加权向量,最后将加权向量分别与 2 个初始的特征向量并行拼接,确保特征矩阵在特征融合时能够尽量保留原始特征信息而不失真,公式为

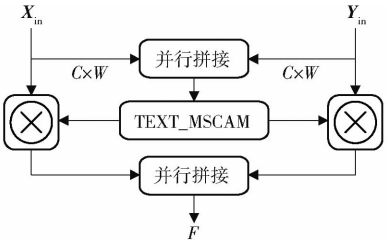

图 10　TEXTAFF 模型结构图

Fig. 10　Schematic of TEXTAFF

$$F(\boldsymbol{X}_{in},\boldsymbol{Y}_{in}) = [M(\boldsymbol{X}_{in}\cup\boldsymbol{Y}_{in})\otimes\boldsymbol{X}_{in}]\cup [(1-M(\boldsymbol{X}_{in}\cup\boldsymbol{Y}_{in}))\otimes\boldsymbol{Y}_{in}] \tag{22}$$

式中　$\boldsymbol{X}_{in}$、$\boldsymbol{Y}_{in}$——输入向量

$F()$——注意力融合特征

式(22)中的 F 表示融合后特征，$\boldsymbol{X}_{in}$、$\boldsymbol{Y}_{in}$表示输入融合初始特征，并行拼接操作使得特征矩阵在融合时能够尽量保留特征向量的信息而不失真，保证了语义全面性。

2　试验与结果分析

2.1　试验数据

试验数据源自“中国农技推广”的农业问答社区。该数据集涵盖了病虫草害、动物疫病、土壤肥料等多个专业领域，共计 19 968 条句子对。为了确保模型的泛化能力，每个问句对都属于同一类别，部分数据集样例如表 1 所示。为确保数据标注的准确性与可靠性，采用人工标注，采用人工交叉核验的方法对句子对的语义相似性进行了多次核对，2 个问句语义相同或相近，则标注为 1；反之，则标注为 0，其中标注为 1 的句子对共有 9 221 条，占总数据集的 46.18%；标注为 0 的句子对共有 10 747 条，占总数据集的 53.82%。

表 1　数据集样例

Tab. 1　Sample of dataset

编号	问句 1	问句 2	真实标签
1	土豆癌肿病的症状有哪些？	土豆癌肿病有什么症状？	1
2	牡丹缺钾症防治方法有哪些？	杜鹃缺钾症防治方法有哪些？	0
3	防治夏季玉米钻心虫危害，危害症状是什么？	如何防治夏季玉米钻心虫危害？	0
4	种植土豆应该选择什么样的土地种植？	应该选择什么样的种植土豆土地种植？	1
5	柑橘幼树能不能追施尿素肥料吗？	柑橘幼树如何施尿素肥料吗？	0
6	请问各位老师这是什么虫，它正在取食樱桃树叶，该如何防治？	各位老师，正在取食樱桃树叶的虫子是什么，该如何防治这种虫子？	1
7	客源市场对休闲观光农业有什么影响？	休闲观光农业有哪些影响因素？	0

整个数据集划分成训练集、验证集、测试集 3 个部分，其中训练集占比 80%，主要用于模型对于数据集的学习；验证集占比 10%，根据验证集的效果调整模型超参数；测试集占比 10%，调整完成参数的模型经过测试集验证其泛化性，并且训练集、验证集、测试集相互独立。

2.2　模型评价指标

模型使用正确率（Accuarcy）、精确率（Precision）、召回率（Recall）、F1 值作为评价指标，衡量模型在语义匹配任务中的效果。其中正确率是指模型相似度预测正确的样本数与预测的总样本数之比，反映模型整体正确程度。精确率是指模型预测为相似的样本中真正为相似的样本数与总预测为相似的样本数之比，反映模型在预测为相似时的精确性。召回率是指真正为相似的样本中被模型预测为相似的样本数与总真正为相似的样本数之比，反映模型在找出所有真正为相似的样本时的能力。F1 值是精确率和召回率的调和平均数，旨在平衡精确率和召回率，使得模型在两者之间取得一个较好的折衷。

2.3　模型参数设置

试验数据集词向量维度为 256，多尺度卷积通道注意力加权网络 MSCN 中的卷积核数量为 64，基于多头自注意力机制的双向长短记忆网络 Multi_SAB 中的输出特征维度为 256，注意力头数量为 4，多尺度通道注意力模块 TEXT_MSCAM 中的通道缩放倍数为 16，dropout 值为 0.4。表 2 为试验环境与配置参数。

表 2　试验环境与配置参数

Tab. 2　Experimental configuration environment

试验环境	环境配置及硬件参数
操作系统	Windows 11 22H2
内存	DDR4 32 GB 3 200 MHz
CPU	AMD RYZEN 5 5600X 3.7 GHz
Python	3.9
Pytorch	2.1.0

2.4　模型性能

模型将与语义匹配领域内的基于孪生神经网络的深度学习模型进行比对，包括 MaLSTM[14]、BiLSTM[15]、BiLSTM_Self - attention[37]、TEXTCNN_

Attention[38]、Sentence - BERT[23],其中 BiLSTM_Self - attention 是在 BiLSTM 的基础上融合自注意力机制模块,TEXTCNN_Attention 是在 TEXTCNN 的基础上融合注意力机制。

2.4.1 语义匹配效果试验

表 3 展示了 Font_MBAFF 与对比试验模型在语义匹配任务的试验结果,其中基于预训练模型的 Sentence - BERT 正确率、精确率、召回率、F1 值均超过 94%,取得了较好的效果;基于注意力机制的 BiLSTM_Self - attention 的 4 项评价指标均超过 91%,TEXTCNN_Attention 的 4 项评价指标均超过 91%,而传统的 BiLSTM 及 MaLSTM 正确率均低于 90%。Font_MBAFF 4 项评价指标均超过 96%,均为最高值,明显优于其他的 5 种比对模型。正确率、精确率、召回率、F1 值比 Sentence - BERT 高出 2.07、2.12、2.22、2.17 个百分点。

表 3 试验模型结果比对

Tab.3 Comparison of model results %

试验模型	正确率	精确率	召回率	F1 值
MaLSTM	85.79	88.31	80.83	84.40
BiLSTM	87.85	91.60	81.99	86.53
BiLSTM_Self - attention	92.49	91.89	92.37	92.13
TEXTCNN_Attention	91.99	91.79	91.42	91.56
Sentence - BERT	94.35	94.07	94.07	94.07
Font_MBAFF	96.42	96.19	96.29	96.24

2.4.2 消融试验

Font_MBAFF 以 TEXTCNN_BiLSTM 为基础,分别增加字形特征模块 Font、文本多尺度通道注意力模块 TEXT_MSCAM、多头自注意力机制模块 Multi_SA 和文本注意力融合模块 TEXTAFF。为了验证模型综合优化效果,将 TEXTCNN_BiLSTM 作为 baseline,对比 Font_MBAFF 与其在文本比配任务的正确率和 F1 值,结果为:baseline 正确率、F1 值分别为 92.39%、92.05%,而 Font_MBAFF 正确率、F1 值为 96.42%、96.24%,分别比 baseline 的高 4.03、4.19 个百分点,模型整体优化效果显著。

为了验证各模块对模型性能提升效果的影响,将各优化模块分别与 baseline 进行对比试验,对比在文本比配任务的正确率和 F1 值,具体结果如图 11 所示。Font 模块对模型性能提升效果影响最大,正确率和 F1 值分别提高 2.57、2.85 个百分点,说明丰富的语义特征有助于提高语义匹配质量,特别面对特征词较少的短文本数据集,性能提升更加明显。TEXT_MSCAM 模块将正确率和 F1 值提高 0.60、0.65 个百分点,Multi_SA 将正确率和 F1 值提高 0.96、0.97 个百分点,说明在模型中引入注意力机制,对各部分语义特征进行精准赋值,可有效提高模型性能。TEXTAFF 将正确率和 F1 值提高 0.65、0.65 个百分点,说明相比于简单的特征拼接,基于注意力特征融合方法可智能分析各部分特征权重,对于模型正确率的提升有显著效果。

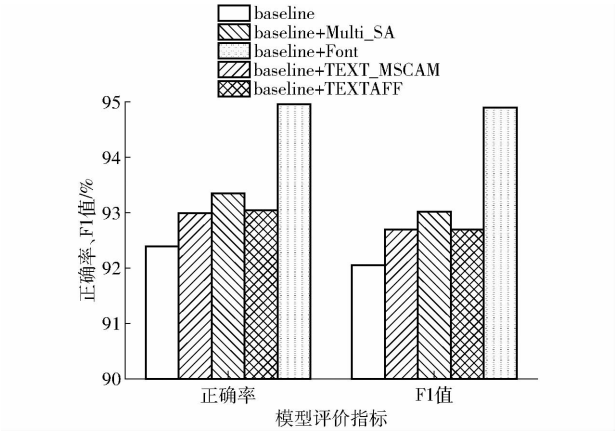

图 11 各模块消融结果比对

Fig. 11 Comparison of ablation results of each module

2.4.3 响应时间试验

为验证模型在实际应用场景中的有效性,选择正确率较高的 BiLSTM_Self - attention、Sentence - BERT 作为对比模型,总共对 1 984 对农业领域短文本问句进行测试。Font_MBAFF 的实时响应时间为 1.42 s,与 BiLSTM_Self - attention 响应时间(1.09 s)基本持平,但远低于 Sentence - BERT 的响应时间(2.91 s),基本满足实际任务对快速反馈语义匹配结果的要求。

2.4.4 不同规模数据集适应性试验

为了验证 Font_MBAFF 在不同规模数据集上依旧保持较强的鲁棒性,选择 BiLSTM_Self - attention 及 Sentence - BERT 作为比对模型,分别在 4 992、9 920、16 000 组问句对的数据集上进行试验验证,图 12 展示了 3 种试验方法在不同规模数据集下的文本相似度正确率。所有的试验模型在小规模数据集正确率低于大规模数据集,说明基于深度学习模

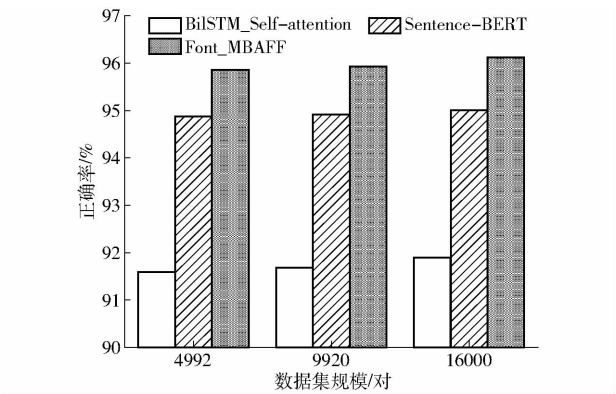

图 12 不同数据规模结果对比

Fig. 12 Comparison of results of different data scales

型需要海量数据集支撑。但 Font_MBAFF 在 3 种规模的数据中均得到了最优效果，特别在小规模的数据集中，正确率达到 95.61%，比 BiLSTM_Self-attention 高 3.9 个百分点，比 Sentence-BERT 高 1.72 个百分点，优势明显，表明了 Font_MBAFF 具有较强的鲁棒性。

3　结论

（1）Font_MBAFF 模型在文本语义匹配任务中取得了较好效果，正确率、精确率、召回率、F1 值分别达到 96.42%、96.19%、96.29%、96.24%，与 5 种比对模型相比优势明显，测试响应时间较短，满足快速获取匹配结果要求，即使在面对小规模数据集时，模型仍保持较好的稳定性，验证了模型在农业短文本语义匹配任务中的适用性。

（2）模型引入汉字字形模块 Font，模型正确率提高 2.57 个百分点，有效地弥补了中文农业短文本语义特征稀疏的问题。在特征提取方面，模型构建了基于空间序列和时间序列联合提取文本语义特征策略，提出了多尺度卷积核通道注意力加权网络 MSCN，模型正确率提高 0.60 个百分点，能够提取不同视野范围的文本特征信息，有效解决特征语义尺度不一致问题；提出一种基于多头自注意力的孪生双向长短期记忆网络 Multi_SAB，模型正确率提高 0.96 个百分点，有效弥补了循环神经网络在不同空间位置上的上下文特征提取能力不足的问题。特征融合方面，利用由多尺度通道注意力组成的文本注意力融合模块实现了在不同尺度语义特征以及不同分支输出特征之间的统一融合方案，模型正确率提高 0.65 个百分点，实现了多语义特征的智能、精准融合。

参　考　文　献

[1] 赵春江，李瑾，冯献. 面向 2035 年智慧农业发展战略研究[J]. 中国工程科学，2021，23(4)：1-9.
ZHAO Chunjiang，LI Jin，FENG Xian. Development strategy of smart agriculture for 2035 in China[J]. Strategic Study of CAE，2021，23(4)：1-9.（in Chinese）

[2] 舒圣宝，房瑞，邓醒，等. 数字农业发展研究文献综述[J]. 农业工程，2022，12(9)：144-150.
SHU Shengbao，FANG Rui，DENG Xing，et al. Literature review on development of digital agriculture[J]. Agricultural Engineering，2022，12(9)：144-150.（in Chinese）

[3] 中国农技推广信息平台[DB/OL]. [2023-10-20]. http://njtg.Nercita.org.cn/user/index.Shtml.

[4] 王郝日钦，吴华瑞，冯帅，等. 基于 Attention_DenseCNN 的水稻问答系统问句分类[J]. 农业机械学报，2021，52(7)：237-243.
WANG Haoriqin，WU Huarui，FENG Shuai，et al. Classification technology of rice questions in question answer system based on Attention_DenseCNN[J]. Transactions of the Chinese Society for Agricultural Machinery，2021，52(7)：237-243.（in Chinese）

[5] 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）

[6] 饶海笛. 基于语义的作物病虫害多模态知识问答方法研究[D]. 合肥：安徽农业大学，2023.
RAO Haidi. Semantic-based multimodal knowledge question and answer method for crop pests and diseases[D]. Hefei：Anhui Agricultural University，2023.（in Chinese）

[7] SALTON G，WONG A，YANG C S. A vector space model for automatic indexing[J]. Communications of the ACM，1975，18(11)：613-620.

[8] AIZAWA A. An information-theoretic perspective of tf-idf measures[J]. Information Processing & Management，2003，39(1)：45-65.

[9] ZHANG Y，JIN R，ZHOU Z H. Understanding bag-of-words model：a statistical framework[J]. International Journal of Machine Learning and Cybernetics，2010，1：43-52.

[10] RAHUTOMO F，KITASUKA T，ARITSUGI M. Semantic cosine similarity[C]// The 7th International Student Conference on Advanced Science and Technology ICAST，2012.

[11] RISTAD E S，YIANILOS P N. Learning string-edit distance[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence，1998，20(5)：522-532.

[12] KRAUSE E F. Taxicab geometry[J]. The Mathematics Teacher，1973，66(8)：695-706.

[13] 谭静. 基于向量空间模型的文本相似度算法研究[D]. 成都：西南石油大学，2015.
TAN Jing. Research on text similarity algorithm based on vector space model[D]. Chengdu：Southwest Petroleum University，2015.（in Chinese）

[14] MUELLER J，THYAGARAJAN A. Siamese recurrent architectures for learning sentence similarity[C]// Proceedings of the AAAI Conference on Artificial Intelligence，2016.

[15] NECULOIU P，VERSTEEGH M，ROTARU M. Learning text similarity with siamese recurrent networks[C]// Proceedings of the 1st Workshop on Representation Learning for NLP，2016：148-157.

[16] ZHAO Weidong，LIU Xiaotong，JING Jun，et al. Re-LSTM：a long short-term memory network text similarity algorithm based on weighted word embedding[J]. Connection Science，2022，34(1)：2652-2670.

[17] SHI H, WANG C, SAKAI T. A Siamese CNN architecture for learning Chinese sentence similarity[C]//Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing: Student Research Workshop, 2020: 24 - 29.
[18] MORAVVEJ S V, JOODAKI M, KAHAKI M J M, et al. A method based on an attention mechanism to measure the similarity of two sentences[C]//2021 7th International Conference on Web Research (ICWR). IEEE, 2021: 238 - 242.
[19] 孟金旭, 单鸿涛, 万俊杰, 等. BSLA: 改进 Siamese - LSTM 的文本相似模型[J]. 计算机工程与应用, 2022, 58(23): 178 - 185.
MENG Jinxu, SHAN Hongtao, WAN Junjie, et al. BSLA: improved text similarity model for Siamese - LSTM[J]. Computer Engineering and Applications, 2022, 58(23): 178 - 185. (in Chinese)
[20] YIN W, SCHÜTZE H, XIANG B, et al. ABCNN: attention-based convolutional neural network for modeling sentence pairs [J]. Transactions of the Association for Computational Linguistics, 2016, 4: 259 - 272.
[21] 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)
[22] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in Neural Information Processing Systems, 2017, 30: 5998 - 6008.
[23] REIMERS N, GUREVYCH I. Sentence-bert: sentence embeddings using siamese bert-networks[J]. arXiv Preprint, arXiv: 1908. 10084, 2019.
[24] LAN Z, CHEN M, GOODMAN S, et al. ALBERT: a lite bert for self-supervised learning of language representations[J]. arXiv Preprint, arXiv: 1909. 11942, 2019.
[25] LI J, ZHANG X, ZHOU X. ALBERT-based self-ensemble model with semisupervised learning and data augmentation for clinical semantic textual similarity calculation: algorithm validation study[J]. JMIR Medical Informa Tics, 2021, 9(1): e23086.
[26] 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)
[27] 韦婷婷, 葛晓月, 熊俊涛. 基于层级多标签的农业病虫害问句分类方法[J]. 农业机械学报, 2024, 55(1): 263 - 269, 435.
WEI Tingting, GE Xiaoyue, XIONG Juntao. Hierarchical multi-label classification of agricultural pest and disease interrogative questions[J]. Transactions of the Chinese Society for Agricultural Machinery, 2024, 55(1): 263 - 269, 435. (in Chinese)
[28] 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)
[29] 杨鹤, 于红, 孙哲涛, 等. 基于双重注意力机制的渔业标准实体关系抽取[J]. 农业工程学报, 2021,37(14): 204 - 212.
YANG He, YU Hong, SUN Zhetao, et al. Fishery standard entity relation extraction using dual attention mechanism[J]. Transactions of the CSAE, 2021, 37(14): 204 - 212. (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 based on attention mechanis[J]. Transactions of the Chinese Society for Agricultural Machinery, 2021, 52(1): 185 - 192. (in Chinese)
[31] 李书琴, 张明美, 刘斌. 融合字词语义信息的猕猴桃种植领域命名实体识别研究[J]. 农业机械学报, 2022, 53(12): 323 - 331.
LI Shuqin, ZHANG Mingmei, LIU Bin. Kiwifruit planting entity recognition based on character and word information fusion [J]. Transactions of the Chinese Society for Agricultural Machinery, 2022, 53(12): 323 - 331. (in Chinese)
[32] 蒲攀, 张越, 刘勇, 等. Transformer 优化及其在苹果病虫命名实体识别中的应用[J]. 农业机械学报, 2023, 54(6): 264 - 271.
PU Pan, ZHANG Yue, LIU Yong, et al. Transformer optimization and application in named entity recogniton of apple diseases and pests[J]. Transactions of the Chinese Society for Agricultural Machinery, 2023, 54(6): 264 - 271. (in Chinese)
[33] 朱一鸣, 赵阳, 唐宁, 等. 笔画节点在手写体汉字识别中的作用[J]. 心理学报, 2023, 55(12): 1903 - 1916.
ZHU Yiming, ZHAO Yang, TANG Ning, et al. The role of stroke nodes in the recognition of handwritten Chinese characters [J]. Acta Psychologica Sinica, 2023, 55(12): 1903 - 1916. (in Chinese)
[34] 汉典[DB/OL]. [2023 - 09 - 20]. https://www.zdic.net.
[35] 中华字典[DB/OL]. [2023 - 09 - 20]. https://www.zhonghuazidian.com.
[36] 周雨昊, 孙哲, 吴晓非, 等. 基于门控特征融合的中文错别字纠正模型[J]. 北京邮电大学学报, 2023, 46(4): 91 - 96, 122.
ZHOU Yuhao, SUN Zhe, WU Xiaofei, et al. Chinese spelling correction model based on gated feature fusion[J]. Journal of Beijing University of Posts and Telecommunications, 2023, 46(4): 91 - 96, 122. (in Chinese)
[37] XIANG H, GU J. Research on question answering system based on Bi - LSTM and self-attention mechanism[C]//2020 IEEE 7th International Conference on Industrial Engineering and Applications (ICIEA). IEEE, 2020: 726 - 730.
[38] ALSHUBAILY I. TextCNN with attention for text classification[J]. arXiv Preprint, arXiv: 2108. 01921, 2021.