Journal of Computer Applications
计算机应用,2023,43(2):365 - 373
ISSN 1001-9081
CODEN JYIIDU
2023-02-10
http://www.joca.cn

文章编号:1001-9081(2023)02-0365-09
DOI:10.11772/j.issn.1001-9081.2021122167

基于池化和特征组合增强BERT的答案选择模型

胡 婕[1,2*],陈晓茜[1],张 龑[1,2]
(1.湖北大学 计算机与信息工程学院,武汉 430062; 2.湖北省教育信息化工程技术研究中心(湖北大学),武汉 430062)
(*通信作者电子邮箱JieHu@hubu.edu.cn)

摘 要:当前主流模型无法充分地表示问答对的语义,未充分考虑问答对主题信息间的联系并且激活函数存在软饱和的问题,而这些会影响模型的整体性能。针对这些问题,提出了一种基于池化和特征组合增强BERT的答案选择模型。首先,在预训练模型BERT的基础上增加对抗样本并引入池化操作来表示问答对的语义;其次,引入主题信息特征组合来加强问答对主题信息间的联系;最后,改进隐藏层的激活函数,并用拼接向量通过隐藏层和分类器完成答案选择任务。在SemEval-2016CQA和SemEval-2017CQA数据集上进行的验证结果表明,所提模型与tBERT模型相比,准确率分别提高了3.1个百分点和2.2个百分点;F1值分别提高了2.0个百分点和3.1个百分点。可见,所提模型在答案选择任务上的综合效果得到了有效提升,准确率和F1值均优于对比模型。

关键词:答案选择;预训练模型;池化;特征组合;激活函数
中图分类号:TP391.1 **文献标志码**:A

Answer selection model based on pooling and feature combination enhanced BERT

HU Jie[1,2*], CHEN Xiaoxi[1], ZHANG Yan[1,2]
(1. *School of Computer Science and Information Engineering, Hubei University, Wuhan Hubei* 430062, *China*;
2. *Hubei Engineering Technology Research Center for Educational Informatization (Hubei University), Wuhan Hubei* 430062, *China*)

Abstract: Current main stream models cannot fully express the semantics of question and answer pairs, do not fully consider the relationships between the topic information of question and answer pairs, and the activation function has the problem of soft saturation, which affect the overall performance of the model. To solve these problems, an answer selection model based on pooling and feature combination enhanced BERT (Bi-directional Encoder Representations from Transformers) was proposed. Firstly, adversarial samples and pooling operation were introduced to represent the semantics of question and answer pairs based on the pre-training model BERT. Secondly, the relationships between topic information of question and answer pairs were strengthened by the feature combination of topic information. Finally, the activation function in the hidden layer was improved, and the splicing vector was used to complete the answer selection task through the hidden layer and classifier. Model validation was performed on datasets SemEval-2016CQA and SemEval-2017CQA. The results show that compared with tBERT model, the proposed model has the accuracy increased by 3.1 percentage points and 2.2 percentage points respectively, F1 score increased by 2.0 percentage points and 3.1 percentage points respectively. It can be seen that the comprehensive effect of the proposed model on the answer selection task is effectively improved, and both of the accuracy and F1 score of the model are better than those of the model for comparison.

Key words: answer selection; pre-training model; pooling; feature combination; activation function

0 引言

答案选择[1-2]是从候选答案池中找到与问题最相符的答案。它是一种句子匹配任务[3],即判断两个句子之间的相似度,是问答(Question Answering, QA)系统中非常重要的子任务。

随着深度学习的发展,许多深度学习技术被应用到答案选择任务上。预训练模型出现之前,用得比较多的是双向长短期记忆(Bi-directional Long Short-Term Memory, BiLSTM)网络。如Neculoiu等[4]提出的Siamese-BiLSTM模型结合BiLSTM与Siamese体系结构来提取问答对的语义特征,然后采用余弦函数计算问答对的相似度;但是单一的BiLSTM网络难以捕捉问题和答案的语义信息以及字符序列上下文关系方面的特征。由于隐含狄利克雷分布(Latent Dirichlet Allocation, LDA)主题模型[5]可以为文本匹配提供主题信息,文献[6-7]中将LDA主题模型应用到文本匹配任务中,帮助

收稿日期:2021-12-29;**修回日期**:2022-06-04;**录用日期**:2022-06-10。 **基金项目**:国家自然科学基金资助项目(61977021)。
作者简介:胡婕(1977—),女,湖北汉川人,教授,博士,主要研究方向:复杂语义大数据管理、自然语言处理; 陈晓茜(1997—),女,河南平顶山人,硕士研究生,主要研究方向:自然语言处理; 张龑(1974—),男,湖北宜昌人,教授,博士,CCF会员,主要研究方向:软件工程、信息安全。

理解句子对的语义信息。在结合主题模型的应用上有不同的策略，Wu 等[7]提出了 ECNU（traditional method of Extracting features and Convolutional Neural Network）模型，它使用包括基于主题模型的特征在内的多种类型的特征与卷积神经网络（Convolutional Neural Network，CNN）结合表示问答对的语义。相较于 Siamese-BiLSTM 模型和 LDA 主题模型，它进一步增强了句子对的语义表示，但没有充分考虑问题和答案之间的交互，容易丢失重要的信息。为了更好地实现问题和答案之间的信息交互，注意力机制被引入到答案选择任务中，Wen 等[8]提出了 UIA-LSTM-CNN（User Interaction Attention-Long Short Term Memory-Convolutional Neural Network）模型。该模型利用 CNN 和 LSTM 混合模式的注意力机制学习问题和答案的语义信息，并将用户信息作为答案选择任务的额外特征。上述模型挖掘了问题和答案中的隐含信息，但是由于引入的用户信息存在噪声问题，Xie 等[9]提出了 AUANN（Attentive User-engaged Adversarial Neural Network）模型，进一步改进引入用户信息的模型，利用对抗训练模块过滤与问题不相关的用户信息。上述模型大多使用 word2vec[10]或者 GloVe（Global Vectors）[11]词嵌入方法进行词向量表示，表示方法都是静态固定的，无法表示上下文语义信息。

Google 公司发布的预训练模型 BERT（Bidirectional Encoder Representations from Transformers）[12]改变了自然语言处理（Natural Language Processing，NLP）任务中的模型架构范式，在自然语言处理领域很多任务都取得较好的效果。Laskar 等[13]将预训练模型 BERT 应用到答案选择任务上，使用预训练模型 BERT 中的[CLS]表示问题和答案的语义，然后通过 softmax 层计算问题和答案的相似度。与 CNN 和 BiLSTM 相比，预训练模型 BERT 可以通过深层模型获取问题和答案丰富的语义信息；但是，使用 BERT 学习到的语义特征不够丰富，容易造成文本匹配的语义缺失。近年来，由于图神经网络可以保存全局信息，Chen 等[14]提出了 GMN-BERT（Graph Matching Networks-Bidirectional Encoder Representations from Transformers）模型。该模型使用 BERT 获得每个词的上下文表示，将对应的词嵌入转换为节点嵌入，并利用图匹配模块来计算两个文本的图级表示的相似性。尽管基于图神经网络的方法能够捕捉句子的全局信息，但该方法各节点信息没有充分交互。为了有效提取句子对的语义信息，Peinelt 等[15]提出了 tBERT（topic models with Bidirectional Encoder Representations from Transformers）模型，该模型使用预训练模型 BERT 提取句子对的语义信息，同时利用 LDA 主题模型辅助表示句子对的语义，取得不错的效果。然而，该模型仍然存在如下问题：首先，tBERT 模型使用 BERT 输出的[CLS]来表示句子对语义，不能完整地提取问答对的语义信息。对 BERT 的输出向量取平均作为问答对的语义信息效果更好[16]。其次，tBERT 模型将两个句子的主题信息和它们的语义表示简单地拼接在一起，无法挖掘问答对主题信息的联系，影响模型整体的表达能力。通过对提取的两个句子特征进行特征组合，能更有效地识别句子间的相关性[17]。最后，隐藏层中的 tanh 激活函数存在软饱和与梯度消失问题，影响模型的性能。

针对上述模型存在的问题，本文在 tBERT 模型的基础上进行改进。首先利用 BERT 模型对语料进行微调，为了提升模型的鲁棒性，在嵌入层增加了对抗训练；然后对 BERT 的输出向量和对抗样本取平均作为问答对的语义信息；同时使用 LDA 主题模型提取问题和答案的主题信息辅助模型理解问答对的语义；其次，使用点积操作、按位减操作对问题和答案的主题特征进行组合，提升问题和答案主题信息间的联系；接着，拼接问答对的语义信息与组合后的主题信息，由隐藏层和 softmax 层做最终处理；最后，结合 tanh 和 ReLU（Rectified Linear Unit）激活函数的优点，设计了一种新的激活函数，提升了模型的性能。实验结果表明，本文模型在 SemEval-2016CQA[18]和 SemEval-2017CQA[19]数据集上的 F1 值达到了 76.1% 和 79.9%，相较于基线模型都有所提升，从而验证了本文采用对 BERT 的输出加入对抗训练并取平均作为问答对的语义表示、对问答对的主题特征进行组合以及改进激活函数方法的有效性。

1 本文模型

本文模型结构如图 1 所示。编码层对预处理完成后的问题和答案进行编码并加入对抗训练；池化层对编码层的问题和答案编码以及对抗样本进行语义提取；主题信息提取层对输入问答对的主题信息进行提取；特征组合层通过点积操作、按位减操作进一步关联问答对的主题信息，然后将组合后的主题信息与问答对的语义信息进行拼接；最后将拼接后的结果经过隐藏层和 softmax 层计算得到问题和答案的相似度。

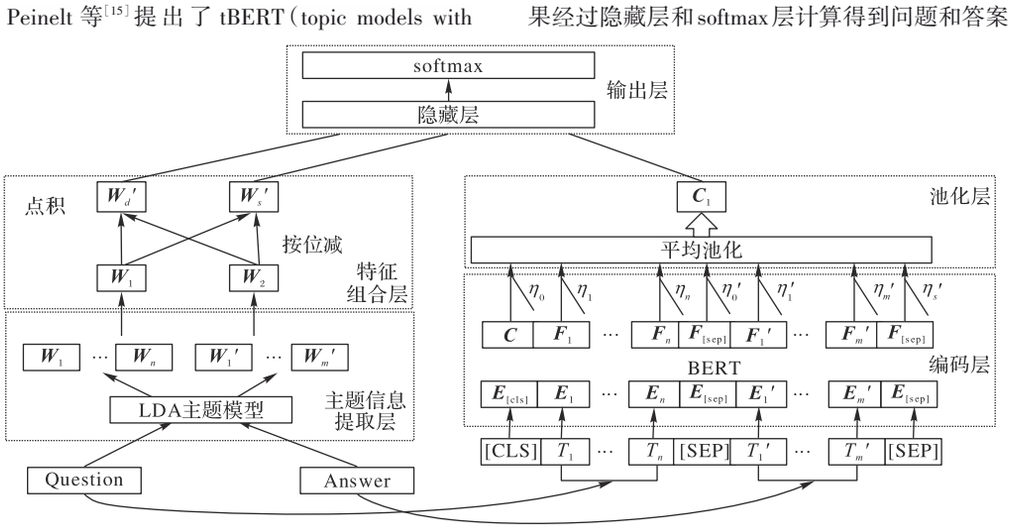

图 1 本文模型结构

Fig. 1 Structure of proposed model

本文模型通过在tBERT模型基础上新增对抗训练、池化层、特征组合层以及改进激活函数,使问答对的语义具有更加丰富的表示,有效完成答案选择任务。

1.1　编码层

编码层使用预训练模型BERT对问题和答案进行编码,预训练模型BERT采用多层双向Transformer的Encoder结构,解决了传统语言模型单向的局限性以及长期依赖的问题。本文使用的BERT-base模型,包含12个编码块。BERT语言模型有两个任务:第一个任务是在句子中随机遮挡一部分单词,模型利用上下文信息来预测这些被遮挡的单词;第二个任务是预测下一个句子。BERT预训练阶段结合这两个任务同时进行,BERT模型结构如图2所示。

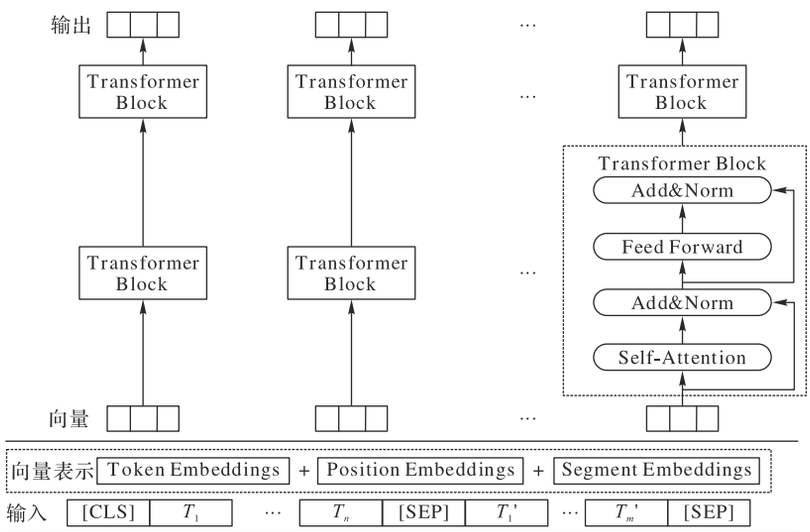

图2　BERT模型结构

Fig. 2　Structure of BERT model

为了应对下游任务,BERT给出了句子级别的表示。如果BERT的输入是两个句子,需要在句子的头部添加标识符[CLS],两个句子之间以及句子末尾使用分隔符号[SEP]。句子中的每个单词由字嵌入向量、分段嵌入向量和位置编码向量三部分组成,其中字嵌入向量的起始单词嵌入为$\boldsymbol{E}_{[CLS]}$,分隔符为$\boldsymbol{E}_{[SEP]}$,最后一个单词嵌入为$\boldsymbol{E}_{[SEP]}$。分段嵌入向量是用来区分两个输入句子,对这两个句子中的单词分配不同的编码(例如:A句子用0编码,B句子用1编码)。位置编码向量表示单词在句中的位置不同可能会导致完全不同的语义。将问题和答案用整合后的向量表示,作为BERT的输入,通过堆叠12个Transformer编码块,得到输出向量,表示为:

$$BERT(\text{Question, Answer}) = [\boldsymbol{C}, \boldsymbol{F}_1 \cdots \boldsymbol{F}_n \cdots \boldsymbol{F}_m'] \tag{1}$$

其中:$\boldsymbol{C}$、$\boldsymbol{F}_1$、$\boldsymbol{F}_n$、$\boldsymbol{F}_m'$分别为问题和答案中的每个单词经过BERT后的向量表示。

对抗训练通过在原始输入上增加对抗扰动来提高模型的鲁棒性。受文献[9]的启发,本文在经过预训练模型BERT处理后获得的初始向量表示后,添加扰动进行对抗训练。将通过预训练模型BERT后的输出向量$\{\boldsymbol{C}, \boldsymbol{F}_1 \cdots \boldsymbol{F}_n \cdots \boldsymbol{F}_m'\}$定义为$\boldsymbol{F}$,对抗样本$\boldsymbol{F}_{adv}$的计算公式为:

$$\boldsymbol{\eta}_{adv} = \varepsilon \frac{\boldsymbol{g}}{\| \boldsymbol{g} \|_2} \tag{2}$$

$$\boldsymbol{g} = \nabla_{\boldsymbol{F}} L(\boldsymbol{F};\ \theta) \tag{3}$$

$$\boldsymbol{F}_{adv} = \boldsymbol{F} + \boldsymbol{\eta}_{adv} \tag{4}$$

其中:ε是扰动$\boldsymbol{\eta}_{adv}$的无限范数最大值,$\boldsymbol{g}$表示梯度,θ是模型参数,$\boldsymbol{\eta}_{adv}$是添加的扰动,$\nabla_{\boldsymbol{F}}$表示输入$\boldsymbol{F}$求导得到的梯度。对抗训练在一个步骤中做了两次梯度更新:第一次做梯度上升是为了找到最佳扰动$\boldsymbol{\eta}_{adv}$,使得损失最大;第二次做梯度下降是为了找到最佳模型参数,使得损失最小。

生成对抗样本之后,原始向量表示和对抗样本一同输入到池化层。

1.2　池化层

经过预训练模型BERT后,得到问题和答案的句子表示。先前的工作一般使用BERT中的[CLS]作为问答对的语义表示,[CLS]虽然可以代表整个句子的语义,但是没有考虑到每个token的信息,可能造成语义信息的丢失。考虑到这些问题以及池化层可以去除冗余信息,压缩信息的特征。为了进一步挖掘问题和答案的语义信息,本文对BERT的输出以及对抗样本取平均作为问答对的语义表示。池化通常分为平均池化和最大池化两种:最大池化不能将整个句子的语义信息保留下来,因此本文使用平均池化提取问答对的语义信息;平均池化方法沿着文本长度和嵌入维度求均值,实现隐藏序列到向量的转换。问答对的语义表示$\boldsymbol{C}_1$表示为:

$$\boldsymbol{C}_1 = mean\_pooled[\boldsymbol{F}, \boldsymbol{F}_{adv}] \in \mathbf{R}^{batch\_size*hidden\_size} \tag{5}$$

1.3　主题信息提取层

为了增强问题和答案的语义特征,该层使用LDA主题模型[5]提取问答对的主题信息。LDA主题模型是一个三层贝叶斯概率模型,该模型认为一篇文档有多个主题,每个主题对应不同的词。在LDA模型中,假设n篇文档中含有K个主题,w表示每篇文档对应的单词,z表示每篇文档中单词对应的主题,θ表示每篇文档的主题概率,服从先验参数α的狄利克雷分布,Φ表示每个主题对应词的概率,服从先验参数β的狄利克雷分布,模型结构如图3所示。

LDA提取主题词的过程为:首先,对于给定的文档生成一个隐狄利克雷分布模型,得到θ作为主题-文档分布;然后,对于每个主题,根据先验参数β的狄利克雷分布得到主题词的分布Φ;最后,根据主题词分布和主题文档分布得到z

和w。本文采用当前LDA模型中主流的采样方法——吉布斯采样算法求解得到全局主题z的分布和词语的分布，需要确定3个超参数α、β和最优主题数K，其中α和β使用默认值，K通过困惑度计算确定。计算公式为：

$$Perlexity(D) = \exp\left\{-\frac{\sum_{i=1}^{M}\ln p(d_i)}{\sum_{i=1}^{M}N_i}\right\} \tag{6}$$

其中：D为语料库中的训练集，共M个问答对，d_i为句子i中的词，$p(d_i)$为句子中词d_i产生的概率，N_i为每个句子中的单词数。

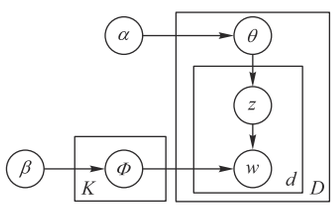

图3 LDA模型结构

Fig. 3 Structure of LDA model

通过LDA主题模型，可以判断问答对的主题信息分布是否一致。首先将问题和答案输入到主题模型，计算每个词的主题分布，计算公式为：

$$w_i = TopicModel(T_i) \tag{7}$$

其中：i表示问答对中单词的序号，T_i表示每个问题和答案中对应的单词。然后，分别对问题和答案所有词的主题分布取平均获得问题和答案的主题信息表示$\boldsymbol{W}_1$和$\boldsymbol{W}_2$，如下式：

$$\boldsymbol{W}_1 = \sum_{i=1}^{n} w^i \Big/ n \tag{8}$$

$$\boldsymbol{W}_2 = \sum_{i=1}^{m} w^{i'} \Big/ m \tag{9}$$

其中：n表示每个问题中的单词个数，m表示每个答案中的单词个数，w_i表示每个问题中对应单词的主题信息，w^i表示每个答案中对应单词的主题信息。

1.4 特征组合层

仅使用LDA主题模型提取问题和答案主题特征，容易忽略上下文词汇间的语义关联，需要进一步组合这些特征加强问题和答案的主题信息之间的逻辑关系。特征组合层对问题和答案的主题信息实现组合并捕捉它们的主题信息，从而使得问题的主题特征向量包括答案的主题特征，这样可以过滤掉与问题主题不相关的答案。例如，问题的主题是“去哪里吃饭”，相对应的答案的主题是“在饭店吃炒鸡”，通过特征组合层可以加强两个主题的联系，把注意力重点放在主题上，判断这两个主题是否有关联。

文献[15]使用拼接、按位乘和按位减操作提取两个句子的特征。该方法能够有效识别句子间的相关性，但是对于问答对主题特征而言，拼接和按位乘操作不能有效建立特征之间的联系，而点积操作可以使它们进行更充分的组合。为了更好地捕捉问答对主题特征间的联系，本文使用点积和按位减操作对问答对的主题特征$\boldsymbol{W}_1$、$\boldsymbol{W}_2$进行组合。

通过使用点积和按位减两种操作，得到两种不同的特征$\boldsymbol{W}_1 \cdot \boldsymbol{W}_2$和$\boldsymbol{W}_1 - \boldsymbol{W}_2$。将这两种特征与问答对的语义特征进行拼接得到最终的组合特征$\boldsymbol{F}'$，如式(10)所示：

$$\boldsymbol{F}' = [\boldsymbol{C}_1;\ \boldsymbol{W}_1 \cdot \boldsymbol{W}_2;\ \boldsymbol{W}_1 - \boldsymbol{W}_2] \tag{10}$$

其中：“;”表示拼接，“·”表示点积操作，“-”表示按位减操作，$\boldsymbol{C}_1$为问答对的语义表示，$\boldsymbol{W}_1$和$\boldsymbol{W}_2$分别为问题和答案的主题信息表示。

1.5 输出层

得到组合特征向量$\boldsymbol{F}$后，应用隐藏层和softmax层得到模型的输出，最终得到问答对相似度概率y'，如式(11)所示：

$$y' = \text{softmax}(hidden\_layer(\boldsymbol{F})) \tag{11}$$

在隐藏层中，tBERT模型使用tanh激活函数。激活函数的作用是把激活神经元的特征通过非线性函数把特征保留并映射出来，也决定了信号网络中能否传递。因此，激活函数影响整个网络的性能。由于tanh函数的取值范围在[-1, 1]区间内，存在软饱和性以及梯度消失的问题，近年来，深度学习网络中经常使用ReLU函数，相较于tanh函数具有较快的收敛性，但是没有负值激活会导致权重无法更新，存在神经元死亡的现象。为了克服tanh函数和ReLU函数的缺点，本文提出一种新的激活函数，定义如下：

$$f(x) = \begin{cases} \dfrac{e^{2x} - e^{-x/2}}{e^{2x} + e^{-x/2}}, & x \leqslant 0 \\ x, & x > 0 \end{cases} \tag{12}$$

该激活函数的正半轴使用ReLU的正半轴，负半轴使用的函数在tanh基础上增大了中心区域的梯度，降低了饱和的速度。

相较于使用tanh作为隐藏层的激活函数，本文提出的激活函数有以下优点：

1)正半轴使用ReLU激活函数，缓解了梯度消失的问题。由于在正半轴，函数的导数值恒为1，这样保证了正半轴梯度不衰减。

2)激活负值，在tanh函数的基础上增大中心区域的梯度，缓解了快速饱和的问题。

3)降低噪声，该函数在负半轴具有软饱和性，意味着可以减小输出到下一层信息的变化。因此，它表现出来的特征可以降低噪声。

1.6 模型训练

在本文中，通过捕捉问答对之间的语义信息来计算它们的相似度。本文应用交叉熵损失函数来衡量答案标签的真实概率分布和预测概率分布之间的差值：

$$Loss = -\sum_{1}^{M}[y \cdot \log(\hat{y}) + (1 - y) \cdot \log(1 - \hat{y})] \tag{13}$$

其中：M为问答对的训练数量，y为样本真实值，$\hat{y}$为样本预测值。

为了防止过拟合，本文采用早期停止法训练模型[13]：为了降低梯度，当测试误差停止降低并开始增大时，就停止训练。模型训练是为了在训练过程中不断地更新参数，尽可能地减小模型的总损失。

2 实验与分析

2.1 数据集与评价指标

2.1.1 数据集

为了验证本文模型的有效性，在两个被广泛使用的社区问答公开数据集SemEval-2016CQA和SemEval-2017CQA上做对比实验。它们的验证集相同，训练集和测试集不同。这

两个数据集是由卡塔尔生活论坛创建的,每个评论上都被贴上“好”“坏”或者“可能有用”的标签。本文将“好”作为正样本,其他标签作为负样本。为了验证激活函数改进的可行性,在一个公开数据集MSRP上做对比实验。该数据集是用于释义识别的数据集,两个句子是否互为释义,是微软研究释义构建的语料库。详细信息如表1所示。

表1　数据集描述

Tab. 1　Description of datasets

数据集	样本数			文本的平均长度
	训练集	验证集	测试集	
SemEval-2016CQA	20 340	2 440	3 270	42
SemEval-2017CQA	14 110	2 440	2 930	46
MSRP	3 576	500	1 725	18

2.1.2　评估标准

在实验中,不同的评价指标能够从不同的角度反映模型的性能。本文采用准确率(Accuracy, *Acc*)和F1值(*F*1)作为评价指标,这两种评价指标越高,代表模型的准确率和综合性能越好。准确率是模型正确预测答案的样本数占总样本数的比例。F1值是精确率和召回率的调和平均数,精确率(Precision, *P*)表示的是正确预测标准答案占实际预测为标准答案的比例,而召回率(Recall, *R*)则表示预测标准答案占标准答案的比例。评价指标的计算公式如下:

$$Acc = \frac{TP + TN}{TP + TN + FP + FN} \times 100\%$$

$$P = \frac{TP}{TP + FP} \times 100\%$$

$$R = \frac{TP}{TP + FN} \times 100\%$$

$$F1 = \frac{2 \times P \times R}{P + R} \times 100\%$$

其中:*TP*(True Positive)是预测该答案是正例且判定正确的次数,*TN*(True Negative)是预测该答案为负例且判定正确的次数,*FP*(False Positive)是预测该答案为正例但判断错误的次数,*FN*(False Negative)是预测该答案为负例但判断错误的次数。

2.2　实验设置

本文实验在Python3.6,Tensorflow1.1,GTX5000平台上运行。本文模型使用BERT-base构建,层数$L = 12$,自注意力头数$A = 12$。其他超参数的设置如表2所示。

表2　参数设置

Tab. 2　Parameter setting

参数	值	参数	值
learning-rate	3E − 5	batch_size	16
optimization	Adam	numbers of topics	70
epochs	3	LDA alpha values	50
hidden_size	768		

2.3　对比模型

实验对比的基线模型分为三类:第一类是传统的文本匹配方法LDA主题模型;第二类是基于CNN和LSTM的方法,分别是ECNU、Siamese-BiLSTM、UIA-LSTM-CNN以及AUANN模型;第三类是基于预训练模型BERT的方法,分别是BERT、GMN-BERT、BERT-pooling以及tBERT模型。这些模型的特点如下。

1)LDA主题模型[5]:一般用于获取文档的主题信息,是传统且简单的算法。在每个数据集的训练部分建立一个主题模型,并计算两个句子主题分布之间的JS散度(Jensen-Shannon Divergence, JSD)。如果JSD大于阈值,该模型预测为负标签,否则预测为正标签。

2)ECNU[7]:该模型采用GloVe方法获取静态词向量,结合传统特征的监督模型和CNN表示问答对特征,最后拼接两个特征向量并使用softmax函数计算问题和答案的得分。

3)Siamese-BiLSTM[4]:该模型是孪生神经网络,左右两边的每一层网络权重共享,使用BiLSTM网络对问题和答案进行编码,然后用余弦函数计算两个编码向量之间的相似度。

4)UIA-LSTM-CNN[8]:该模型采用GloVe方法获取静态词向量,使用CNN和LSTM混合模式注意力机制,计算问题和答案句子中单词的注意力权重以及对应句子中每个单词对一个句子中单词的注意力,结合这两种注意力机制能够使问答对中的大部分信息进行句子匹配。此外,利用学习到对问答有用的用户信息完成答案选择任务。

5)AUANN[9]:该模型与UIA-LSTM-CNN方法相似,利用用户信息学习问答对的上下文信息。不同的是,为了解决引入用户信息而产生的噪声问题,设计一个去噪机制,采用粗粒度和细粒度的选择过程,通过计算答案和用户信息的相似性以及对抗训练两个方法过滤掉与问答不相关的用户信息。

6)BERT[13]:该模型仅使用BERT中的[CLS]表示问题和答案的语义,然后通过全连接层和分类器来完成答案选择任务。

7)GMN-BERT[14]:该模型使用预训练模型BERT获得每个单词的上下文表示,然后将上下文节点作为图节点的初始表示,通过消息传播和更新表示两个步骤,使得每个节点既包含可到达节点的信息又包含了与另一个图中所有节点成对比较的信息。最后使用两个图级表示预测句子对的相似性。

8)BERT-Pooling:该模型在预训练模型BERT的输出层仅使用池化操作表示问答对的语义。

9)tBERT[15]:该模型分别用LDA和GSDMM(Gibbs Sampling algorithm for the Dirichlet Multinomial Mixture model)主题模型与BERT结合完成语义相似度检测任务,其中LDA主题模型适合长文本,GSDMM主题模型适合短文本。由于本文使用的数据集是长文本,故与LDA主题模型结合BERT的tBERT模型对比。该方法仅使用BERT中的[CLS]表示问答对的语义,简单地与问答对的主题信息进行拼接。

2.4　实验分析

首先在tBERT模型中加入对抗训练(Adversarial Training, AT)并引入池化操作,然后引入主题信息特征组合并改进激活函数。为了验证模型改进思路的可行性,分别对tBERT模型及其改进模型进行实验对比。

首先验证在tBERT模型中加入对抗训练并引入池化操作改变问答对语义特征提取方式的可行性。问答对语义信息的提取非常重要,本文考虑到每个token的信息以及模型的鲁棒性,先对BERT的输出加入扰动生成对抗样本,然后对原始向量和对抗样本取平均作为问答对的语义表示。在tBERT模型中分别加入对抗训练和引入池化操作并将两者叠加的结果如表3所示。

表3 tBERT、tBERT-AT、tBERT-pooling和tBERT-AT-pooling模型的准确率和F1值的对比 单位：%

Tab. 3 Comparison of accuracy and F1 scores of tBERT, tBERT-AT, tBERT-pooling, and tBERT-AT-pooling models unit: %

模型	SemEval-2016CQA		SemEval-2017CQA	
	准确率	F1	准确率	F1
tBERT	77.6	74.1	78.3	76.8
tBERT-AT	78.6	74.9	79.4	77.9
tBERT-pooling	78.0	74.4	78.6	77.2
tBERT-AT-pooling	**78.8**	**75.3**	**79.6**	**78.1**

由表3可知，仅加入对抗训练，提高了模型的鲁棒性。对tBERT和tBERT-pooling分别引入对抗训练后，改进后的模型的准确率和F1值在SemEval-2016CQA数据集下相较于原模型分别提高了1和0.8个百分点，0.8和0.9个百分点；在SemEval-2017CQA数据集上分别提高了0.9和1.1个百分点，1和0.9个百分点。仅引入池化操作改变问答对语义特征的提取方式，改进后模型的准确率和F1值在SemEval-2016CQA数据集上相较于tBERT、tBERT-AT模型分别提高了0.4和0.3个百分点，0.2和0.4个百分点；在SemEval-2017CQA数据集上分别提高了0.3和0.4个百分点，0.2和0.2个百分点。加入对抗训练并引入池化操作后模型的准确率和F1值在SemEval-2016CQA数据集上相较于tBERT模型分别提高了1.2和1.2个百分点，在SemEval-2017CQA数据集上分别提高了1.3和1.3个百分点。验证了在tBERT模型上加入对抗训练并引入池化操作是可行的，而且两者的叠加具有更好的效果。

然后验证引入主题信息特征组合的可行性。为了让模型更好地识别问答对主题信息的相关性，提升模型的整体预测能力，本文对问答对的主题信息进行按位减、点积操作从而得到主题信息特征组合。分别对比没有进行特征组合的模型，结果如表4所示。

表4 tBERT、tBERT-AT、tBERT-pooling以及tBERT-AT-pooling模型引入主题信息特征组合前后的准确率和F1值的对比 单位：%

Tab. 4 Comparison of accuracy and F1 scores of tBERT, tBERT-AT, tBERT-pooling and tBERT-AT-pooling models before and after introducing combination of topic information features unit: %

模型	SemEval-2016CQA		SemEval-2017CQA	
	准确率	F1	准确率	F1
tBERT	77.6	74.1	78.3	76.8
tBERT-特征组合	77.9	74.3	78.5	77.0
tBERT-AT	78.6	74.9	79.4	77.9
tBERT-AT-特征组合	78.9	75.1	79.5	78.1
tBERT-pooling	78.0	74.4	78.6	77.2
tBERT-pooling-特征组合	78.4	74.7	78.8	77.5
tBERT-AT-pooling	78.8	75.3	79.6	78.1
tBERT-AT-pooling-特征组合	**79.2**	**75.6**	**79.9**	**78.6**

由表4可知，对tBERT、tBERT-AT、tBERT-pooling以及tBERT-AT-pooling模型分别引入主题信息特征组合后，改进后模型的准确率和F1值在SemEval-2016CQA数据集下相较于各自原模型分别提高了0.3和0.2个百分点，0.3和0.2个百分点，0.4和0.3个百分点，0.4和0.3个百分点；在SemEval-2017CQA数据集下相较于各自原模型分别提高了0.2和0.2个百分点，0.1和0.2个百分点，0.2和0.3个百分点，0.3和0.5个百分点，验证了只引入主题信息特征组合对提升模型的性能是有效的。此外，tBERT-AT-pooling-特征组合模型相较于tBERT模型，准确率和F1值在SemEval-2016CQA数据集下分别提高了1.6和1.5个百分点，在SemEval-2017CQA数据集下分别提高了1.6和1.8个百分点，验证了同时加入对抗训练并引入池化操作和主题信息特征组合对提升模型的性能有更好的效果。

最后验证改进激活函数的可行性。由于tanh激活函数存在软饱和性和梯度消失的问题，影响模型的性能。本文在tanh和ReLU激活函数的基础上对隐藏层中的激活函数进行改进，分别对基线模型tBERT和改进后的tBERT-pooling-特征组合、tBERT-AT-特征组合、tBERT-AT-pooling-特征组合模型使用不同的激活函数，其准确率和F1值的对比如表5所示。

由表5可知，对tBERT、tBERT-AT-特征组合、tBERT-pooling-特征组合以及tBERT-AT-pooling-特征组合模型分别改进激活函数后，改进后模型的准确率和F1值在SemEval-2016CQA数据集下相较于各自原模型分别提高了0.9和0.2个百分点，0.2和0.2个百分点，0.9和0.9个百分点，1.5和0.5个百分点；在SemEval-2017CQA数据集下相较于各自原模型分别提高了0.7和0.5个百分点，0.2和0.3个百分点，0.3和0.7个百分点，0.6和1.3个百分点，验证了改进激活函数对提升模型的性能是有效的。可以看出，仅改进tBERT模型中的激活函数在SemEval-2016CQA数据集上的准确率和F1值提升的效果差别较大，主要原因是该数据集的噪声多，只改进激活函数对模型的整体性能的提升效果不明显；并且对tBERT、tBERT-AT-特征组合、tBERT-pooling-特征组合、tBERT-AT-pooling-特征组合模型分别改进激活函数后，在两个数据集的准确率和F1值的提升效果有明显的差别，这是因为SemEval-2016CQA数据集中的答案包含较多的复杂背景信息，引入特征组合以及改进激活函数后，能够过滤更多的噪声词。引入对抗训练之后，本文模型在两个数据集上的准确率和F1值有所提升，进一步证明对抗训练可以提升模型的鲁棒性。此外，本文模型即tBERT-AT-pooling-特征组合-改进的激活函数模型相较于tBERT模型，准确率和F1值在SemEval-2016CQA数据集下分别提高了3.1和2个百分点，在SemEval-2017CQA数据集下分别提高了2.2和3.1个百分点，验证了同时加入对抗训练并引入池化操作和主题信息特征组合以及对激活函数进行改进对提升模型的性能有更好的效果，而且几个方面的改进具有明显的叠加效果。

表5　tBERT、tBERT-AT-特征组合、tBERT-pooling-特征组合以及tBERT-AT-pooling-特征组合模型改进激活函数前后的准确率和F1值的对比　　单位:%

Tab. 5　Comparison of accuracy and F1 scores of tBERT, tBERT-AT-feature combination, tBERT-pooling-feature combination and tBERT-AT-pooling-feature combination models before and after improving activation function　　unit: %

模型	SemEval-2016CQA		SemEval-017CQA	
	准确率	F1	准确率	F1
tBERT-tanh	77.6	74.1	78.3	76.8
tBERT-改进的激活函数	78.5	74.3	79.0	77.3
tBERT-AT-特征组合-tanh	78.9	75.1	79.5	78.1
tBERT-AT-特征组合-改进的激活函数	79.1	75.3	79.7	78.4
tBERT-pooling-特征组合-tanh	78.4	74.7	78.8	77.5
tBERT-pooling-特征组合-改进的激活函数	79.3	75.6	80.1	78.2
tBERT-AT-pooling-特征组合-tanh	79.2	75.6	79.9	78.6
本文模型	**80.7**	**76.1**	**80.5**	**79.9**

为了进一步验证改进激活函数在其他任务上的有效性,在MSRP数据集上对比基线模型tBERT以及tBERT-改进后的激活函数,其准确率和F1值如表6所示。

表6　tBERT改进激活函数前后在MSRP数据集上的准确率和F1值的对比　　单位:%

Tab. 6　Comparison of accuracy and F1 scores of tBERT, tBERT before and after improving activation function on MSRP dataset　　unit: %

模型	MSRP	
	准确率	F1
tBERT-tanh	89.5	88.4
tBERT-改进后的激活函数	89.8	88.6

由表6可知,对tBERT模型中的激活函数进行改进,改进后模型在MSRP数据集上的准确率和F1值相较于原模型分别提升了0.3和0.2个百分点。相较于Semeval-2016CQA和Semeval-2017CQA数据集,改进tBERT模型中的激活函数在MSRP数据集的准确率和F1值的提升效果有所差别,主要原因是激活函数与数据集的特征有关;但在两个不同任务上性能均有提高,表明改进激活函数是有效的。

综上所述,在tBERT模型基础上加入对抗训练并引入池化操作改变问答对语义特征提取方式、引入主题信息特征组合以及改进激活函数是可行的。

为了进一步验证本文模型的有效性,本文还对2.3节所述的9种模型进行了实验对比,结果如表7所示。

表7　相关模型准确率和F1值的对比　　单位:%

Tab. 7　Comparison of accuracy and F1 scores of related models　　unit: %

模型	SemEval-2016CQA		SemEval-2017CQA	
	准确率	F1	准确率	F1
LDA主题模型	70.3	67.6	71.4	68.4
ECNU	74.3	66.7	78.4	77.6
Siamese-BiLSTM	74.6	68.7	75.3	67.1
UIA-LSTM-CNN	78.2	68.4	77.1	76.4
AUANN	80.5	74.5	78.5	79.8
BERT	75.6	71.9	76.2	70.4
GMN-BERT	76.7	72.8	77.5	71.6
BERT-pooling	76.1	72.5	77.1	71.1
tBERT	77.6	74.1	78.3	76.8
本文模型	**80.7**	**76.1**	**80.5**	**79.9**

从表7可以看出,本文模型在SemEval-2016CQA数据集上的准确率和F1值分别达到80.7%和76.1%,在SemEval-2017CQA上的准确率和F1值分别达到80.5%和79.9%,均取得了最好的效果。实验结果中,第二类模型即ECNU、Siamese-BiLSTM、UIA-LSTM-CNN以及AUANN模型整体性能优于第一类模型即LDA主题模型。这是因为CNN和BiLSTM比LDA主题模型能够更有效地对问题和答案潜在的上下文语义信息进行建模。在第二类模型中,ECNU和Siamese-BiLSTM模型在两个数据集上的结果有明显的差异,在SemEval-2017CQA数据集上的表现能力更好,这是因为SemEval-2017CQA数据集相较于SemEval-2016CQA数据集更加规整,噪声词较少。ECNU模型相较于Siamese-BiLSTM模型,在SemEval-2017数据集上的表现能力较好。这是因为ECNU模型除了用CNN表示问答对的语义,还引入5个额外的特征,有助于建立问答对的相关性。值得注意的是,AUANN模型在第二类模型中性能表现最佳,这是因为AUANN模型在UIA-LSTM-CNN模型的基础上加入了问题–答案、问题–用户信息交互模块并引入了对抗训练过滤掉与问题无关的用户信息。本文模型相较于AUANN模型,在两个数据集上的准确率分别提高了0.2和2个百分点,F1值分别提高1.6和0.1个百分点。这是由于本文模型在嵌入层增加对抗样本并引入池化操作改变问答对语义的提取方式,能够更有效地表示问题和答案深层的语义特征。此外,本文模型还引入特征组合整合问答对的主题信息丰富问答对的语义表示,相较于用户信息作为额外特征,引入问答对主题信息特征组合更能增强问答对的语义表示。

第三类模型都使用预训练模型BERT表示问答对的语义信息,相较于前两类模型,BERT模型使用Transformer的编码器提取问答对信息。BERT-pooling相较于BERT模型,在两个数据集上的准确率分别提高了0.5和0.9个百分点,F1值分别提高了0.6和0.7个百分点。说明对BERT的输出取平均作为问答对的表示效果更好。GMN-BERT相较于BERT-pooling模型,在两个数据集上的准确率分别提高了0.6和0.4个百分点,F1值分别提高了0.3和0.5个百分点。这是由于图神经网络能够捕捉全局结构信息,与预训练模型相结合能够利用问答对潜在的语义关系。在第三类模型中,tBERT在两个数据集上的性能表现最佳。这是因为tBERT模型增加了额外的主题信息特征,更加丰富了问答对的语义

特征表示。

综上所述，本文模型的综合性能优于所有对比模型。这是因为：本文模型加入对抗训练并引入池化操作改变问答对语义特征的提取方式；其次，引入主题信息特征组合加强文本主题信息间的联系，增强问答对的语义表示；最后，本文模型改进了隐藏层的激活函数，增强了特征的鲁棒性，从而提高模型的整体性能。

2.5 案例分析

为了直观地看出本文模型的有效性，首先选用Semeval-2017CQA数据集中的一个例子并可视化对抗训练和池化方法对问题的注意权重。颜色深浅表示单词的重要程度，颜色越深越重要。其结果如表8所示。

从表8可以看出，tBERT模型和本文模型都关注了“salary”“negotiating”“mechanical engineer”“grade 5 in a government company”“Qatar”词语。不同的是，tBERT模型对“mechanical engineer”和“Qatar”赋予最高的注意权重，把注意力重点放在“mechanical engineer”和“Qatar”上。而本文模型由于使用对抗训练和池化方法，使得模型对“How much salary”“salary of mechanical engineer”“grade 5 in a government company”及“benefits”赋予最高的注意权重。两个模型对同一问题预测的答案如表9所示。

由表9可知，tBERT模型错误地预测了该例子的答案，但是本文模型预测正确。由于两个模型对问题中的词语的注意权重不同导致不同的结果。tBERT模型根据问题主题词“mechanical engineer”“Qatar”来预测答案；由于本文使用主题信息特征组合方法，使得模型预测答案的主题词与问题的主题词相关联，如答案中的“12-15”“free government housing、3 000 mobile and internet allowance”分别与问题中的“How much”“benefits”相对应。从注意力可视化到预测答案的结果可以看出，对抗训练和池化方法使得模型关注重点词语并弱化不重要的词语的权重；主题信息特征组合方法，加强问题和答案的主题信息间的联系，最终提升了模型的性能。

表8 tBERT模型与本文模型对同一例子的注意力可视化对比

Tab. 8 Comparison of attention visualization to the same example between tBERT and proposed model

模型	注意力可视化示例
tBERT模型[15]	问题：How much salary? Hi everyone I'm in the process of negotiating my salary but I have no idea how much should be the salary of mechanical engineer with grade 5 in a government company and the benefits. This will be my first time in Qatar. Kindly help me. Thanks in advance.
本文模型	问题：How much salary? Hi everyone I'm in the process of negotiating my salary but I have no idea how much should be the salary of mechanical engineer with grade 5 in a government company and the benefits. This will be my first time in Qatar. Kindly help me. Thanks in advance.

表9 tBERT模型与本文模型对同一问题的预测答案的对比

Tab. 9 Comparison of answers to the same question predicted by tBERT and proposed model

问题	答案	
	tBERT模型[15]	本文模型
How much salary? Hi everyone I'm in the process of negotiating my salary but I have no idea how much should be the salary of mechanical engineer with grade 5 in a government company and the benefits. This will be my first time in Qatar. Kindly help me. Thanks in advance.	Hey; I am a Mechanical Engineer as well and working in Qatar. You can email me and we can discus it further.	That should be around 12-15 and you should get free government housing and a 3 000 mobile and internet allowance. That's it.

3 结语

对于答案选择任务存在的问答对语义信息表示不完整的问题，本文在tBERT模型基础上，引入对抗训练和池化操作来表示问答对的语义信息。由于简单地拼接主题特征不能有效地建立特征之间的联系，本文引入点积操作和按位减操作加强问答对主题特征间的联系，进而增强问答对的语义表示。实验结果表明本文模型相较于tBERT模型能更好地提取问答对的语义特征以及提高模型预测能力；但是深入挖掘问答对中潜在的语义特征，仅利用主题模型作为辅助是不够的。在未来的工作中，我们将进一步研究使用图神经网络或者将知识库嵌入到预训练模型中完成答案选择任务。

参考文献（References）

[1] ASKAR M T R, HUANG J X, HOQUE E. Contextualized embeddings based transformer encoder for sentence similarity modeling in answer selection task[C]// Proceedings of the 12th Language Resources and Evaluation Conference. [S. l.]: European Language Resources Association, 2020: 5505-5514.

[2] YANG L, AI Q Y, GUO J F, et al. aNMM: ranking short answer texts with attention-based neural matching model[C]// Proceedings of the 25th ACM International Conference on Information and Knowledge Management. New York: ACM, 2016: 287-296.

[3] YANG R Q, ZHANG J H, GAO X, et al. Simple and effective text matching with richer alignment features[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2019:4699-4709.

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

[5] BLEI D M, NG A Y, JORDAN M I. Latent Dirichlet allocation[J]. The Journal of Machine Learning Research, 2003, 3: 993-1022.

[6] MIHAYLOV T, NAKOV P. SemanticZ at SemEval-2016 Task 3: ranking relevant answers in community question answering using semantic similarity based on fine-tuned word embeddings[C]// Proceedings of the 10th International Workshop on Semantic Evaluation. Stroudsburg, PA: ACL, 2016: 879-886.

[7] WU G S, SHENG Y X, LAN M, et al. ECNU at SemEval-2017 task 3: using traditional and deep learning methods to address

community question answering task[C]// Proceedings of the 11th International Workshop on Semantic Evaluation. Stroudsburg, PA: ACL, 2017: 365-369.

[8] WEN J H, MA J W, FENG Y L, et al. Hybrid attentive answer selection in CQA with deep users modelling[C]// Proceedings of the 32nd AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2018: 2556-2563.

[9] XIE Y X, SHEN Y, LI Y L, et al. Attentive user-engaged adversarial neural network for community question answering[C]// Proceedings of the 34th AAAI Conference on Artificial Intelligence. Palo Alto, CA: AAAI Press, 2020:9322-9329.

[10] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. (2013-09-07)[2021-01-06]. https://arxiv. org/pdf/1301. 3781. pdf.

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

[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, Volume 1 (Long and Short Papers). Stroudsburg, PA: ACL, 2019:4171-4186.

[13] LASKAR M T R, HOQUE E, HUANG J X. Utilizing bidirectional encoder representations from transformers for answer selection[C]// Proceedings of the 2019 International Conference on Applied Mathematics, Modeling and Computational Science, PROMS 343. Cham: Springer, 2021: 693-703.

[14] CHEN L, ZHAO Y B, LV B, et al. Neural graph matching networks for Chinese short text matching[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020:6152-6158.

[15] PEINELT N, NGUYEN D, LIAKATA M. tBERT: topic models and BERT joining forces for semantic similarity detection[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: ACL, 2020: 7047-7055.

[16] REIMERS N, GUREVYCH I. Sentence-BERT: sentence embeddings using Siamese BERT-networks[EB/OL]. (2019-08-27)[2020-03-24]. http://arxiv. org/abs/1908. 10084. pdf.

[17] 栾克鑫,杜新凯,孙承杰,等. 基于注意力机制的句子排序方法[J]. 中文信息学报, 2018, 32(1):123-130. (LUAN K X, DU X K, SUN C J, et al. Sentence ordering based on attention mechanism[J]. Journal of Chinese Information Processing, 2018, 32(1):123-130.)

[18] NAKOV P, MÀRQUEZ L, MOSHITTI A, et al. SemEval-2016 Task 3: community question answering[C]// Proceedings of the 10th International Workshop on Semantic Evaluation. Stroudsburg, PA: ACL, 2016:525-545.

[19] NAKOV P, HOOGEVEEN D, MÀRQUEZ L, et al. SemEval-2017 Task 3: community question answering[C]// Proceedings of the 11th International Workshop on Semantic Evaluation. Stroudsburg, PA: ACL, 2017:27-48.

This work is partially supported by National Natural Science Foundation of China (61977021).

HU Jie, born in 1977, Ph. D., professor. Her research interests include complex semantic big data management, natural language processing.

CHEN Xiaoxi, born in 1997, M. S. candidate. Her research interests include natural language processing.

ZHANG Yan, born in 1974, Ph. D., professor. His research interests include software engineering, information security.