doi: 10.13413/j.cnki.jdxblxb.2021322

融入新闻标题信息的新闻文本与评论的语义相似度计算方法

李伊仝[1], 王红斌[1], 程 良[2]
(1. 昆明理工大学 信息工程与自动化学院, 昆明 650504; 2. 昆明理工大学 城市学院, 昆明 650051)

摘要: 针对预训练模型在处理新闻这种长文本时会截断一部分文本, 导致文本信息缺失的问题, 提出一种在融入新闻标题信息基础上将 TextRank 算法、隐含 Dirichlet 分布主题模型与预训练模型相结合的方法构建模型, 并将该模型与其他语义相似度计算方法进行对比. 结果表明, 该模型准确率为 82.46%, 召回率为 87.43%, 精确率为 82.68%, F_1 值为 84.99%, 取得了最优结果, 从而有效提高了新闻文本与评论的语义相似度计算性能.

关键词: 语义相似度; 预训练模型; 隐含 Dirichlet 分布; 新闻评论
中图分类号: TP391.1 **文献标志码**: A **文章编号**: 1671-5489(2022)06-1399-08

Semantic Similarity Calculation Method of News Text and Comment Integrated with News Title Information

LI Yitong[1], WANG Hongbin[1], CHENG Liang[2]
(1. *Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming* 650504, *China*;
2. *College of City, Kunming University of Science and Technology, Kunming* 650051, *China*)

Abstract: Aiming at the problem that the pre-training model would cut off part of text when dealing with long text such as news, which led to the loss of text infomation, we proposed a method to build a model by combining TextRank algorithm, implicit Dirichlet distribution topic model and pre-training model on the basis of integrating news title information, and compared the model with other semantic similarity calculation methods. The results show that the accuracy rate of the model is 82.46%, the recall rate is 87.43%, the accuracy rate is 82.68%, and the F_1 value is 84.99%, the optimal results are obtained, which effectively improves the performance of semantic similarity calculation between news texts and comments.

Keywords: semantic similarity; pre-training model; implicit Dirichlet distribution; news comment

在网络时代, 人们需要从海量信息中获取有效内容. 网络新闻文本通常伴有大量的用户评论, 但这些评论中只有一部分评论与新闻文本内容相关, 另一部分评论与新闻文本内容不相关. 语义相似度计算的目的是计算两段文本是否具有相关性, 语义相似度越大, 说明两段文本表述的内容越相关.

收稿日期: 2021-08-27.

第一作者简介: 李伊仝(1996—), 男, 汉族, 硕士研究生, 从事自然语言处理的研究, E-mail: 1007613426@qq.com. **通信作者简介**: 王红斌(1983—), 男, 汉族, 博士, 副教授, 从事自然语言处理与数据分析的研究, E-mail: whbin2007@126.com.

基金项目: 国家自然科学基金(批准号: 61966020)、云南省基础研究计划面上项目(批准号: CB22052C143A)和云南省教育厅科学研究基金(批准号: 2018JS035).

由于文本通常具有不同复杂程度的句法、语法结构，因此如何更好地计算文本之间的语义相似度已成为该领域研究的热点问题之一，目前已取得了许多成果，其中具有代表性的工作主要有基于字符串的方法、基于统计的方法、基于深度学习的方法等.

基于字符串的方法是将字符串放到原文本中进行匹配计算，计算不同文本字符串的共现程度和重复程度，以此作为衡量文本相似度的依据. 主要包括编辑距离、最长公共子序列(longest common substring，LCS)、N-gram 和 Jaccard 相似度等方法. 计算不同文本字符串共现程度和重复程度最常用的方法是计算编辑距离. 编辑距离是将两个字符串文本的差异性转换成一种数学形式上的度量，即通过统计一段文本经过删除、插入、替换等操作后变成另一段文本的操作次数得到编辑距离. 张雷等[1]提出了一种基于编辑距离的词序敏感相似度度量方法，改进了利用余弦相似度计算文本相似度时，因忽略词序而不能理解文本语义的缺点. 编辑距离的方法虽然在计算上比较准确，但计算时间较长. LCS 算法[2]通过计算两个文本重复部分的长度，计算出文本的相似度. 周丽杰等[3]基于 LCS 提出了一种基于关键词数目的语义关联性函数，用于短文本相似度计算. LCS 算法主要针对短文本进行相似度计算，但应用到长文本中效果不佳. N-gram 方法[4]主要思想是设置大小为 N 的滑动窗口，在文本上进行窗口滑动，从而得到长度为 N 的多个文本片段，这些长度为 N 的文本片段称为 N 元组. 通过计算两段文本中公共 N 元组的数量与总的 N 元组数量的比值衡量两段文本的相似度. 黄贤英等[5]提出了一种改进后的问句相似度算法，将 N-gram 及公共词块相结合计算问句向量的相似度，有效提高了问句相似度的准确率. N-gram 方法最大的特点是滑动窗口的大小可根据具体情形进行调节，但也不适用于长文本相似度计算，原因在于文本越长通过 N-gram 得到的文本片段越多，从而增加算法的计算开销. Jaccard 相似度[6]是一种通过计算两段文本中元素交集和并集数量之比，以此表征文本相似度的方法，特点是仅关注两段文本公共元素个数，而不关注元素之间的差异性. 周艳平等[7]将词向量与位置编码相结合后，利用 JS(Jenson's Shannon)散度和 Pearson 积矩相关系数计算词向量之间的相似度，最后用 Jaccard 算法计算句子之间最终的相似度.

基于统计的方法先将文本通过分布假设表征为向量，再将文本映射到同一个向量空间，然后计算文本向量在向量空间的距离，以此作为文本相似度的度量. 目前基于统计的方法主要以向量空间模型(vector space model，VSM)和主题模型(topic model)为主. 向量空间模型主要思想是假设文本语义只与出现在文本中的单词有关，通过计算单词在单篇文档出现的次数和该词在多个文档中出现的次数计算出词频-逆文档频率(term frequency-inverse document frequency，TF-IDF)，根据 TF-IDF 将文本映射为向量表征，最后通过计算向量与向量之间的距离得到文本之间的相似度. 如黄承慧等[8]提出了一种利用 TF-IDF 选取文本中的重要词项，与提出的词项相似度加权树相结合计算文本相似度，有效提升了文本相似度计算的效果. 基于主题模型的方法主要思想是假设每个文档具有多个主题，每个主题都有多个相关的词. 这些主题隐含了文档的语义信息，通过对主题模型的构建和训练，可使模型学习到主题与文本之间的关系，进而可计算文本之间的相似度. 基于主题模型的方法主要以隐含 Dirichlet 分布(latent Dirichlet distribution，LDA)模型[9]为主. 王振振等[10]通过 LDA 主题模型计算两个文本的主题分布后，将主题分布转换成 JS 距离度量两个文本的相似度；张超等[11]通过对名词、动词和其他词集合进行 LDA 主题建模，将三者按一定比例权重进行结合，同样转换为 JS 距离度量相似度；付雨蛟[12]将 LDA 主题模型与变分自编码器相结合，得到了更好的文本表示，最后用向量间夹角余弦值计算文本的语义相似度.

基于深度学习的方法是指通过神经网络生成文本的嵌入表示，从而学习到文本的深层语义，并根据嵌入表示计算文本相似度的方法，也是近年来在自然语言处理(NLP)领域研究较多的方法. 早期主要采用 Word2Vec(word to vector)[13]和 GloVe(global vectors)[14]等词向量模型. Word2Vec 方法可得到词与词之间的向量表示，Word2Vec 分为 CBOW(continuous bag-of-words)模型和 Skip-gram 模型. CBOW 模型以上下文信息预测中心词汇，Skip-gram 模型根据中心词汇预测其上下文信息. GloVe 方法先通过语料库构建所有单词的共现矩阵，然后通过概率计算利用共现矩阵，以此得到文本词向量，最后进行文本相似度计算. 由于该方法进行了全局语料的词频统计，因此在一定程度上考虑了全局信

息. 杨德志等[15]首先利用双向递归神经网络对文本进行上下文建模，将上下文信息向量与词嵌入向量进行融合，输入卷积神经网络中对文本进行语义表示，最后采用余弦相似度计算两个文本语义向量之间的语义相似度；Liu 等[16]将句法特征和相对位置特征与词嵌入向量进行深度融合，采用并行 LSTM (long short-term memory)结构获得两种文本的向量表示，最后利用全连接层将文本向量融合表示转换成概率；周圣凯等[17]将 Word2Vec 训练好的词嵌入向量，将句子对转换成句向量后，输入卷积层中提取情感特征和名词特征，再通过双向门控循环单元获取上下文信息，然后通过池化层进行降维，最后通过全连接层获得句子对的整体语义向量，并计算 Manhatton 距离得到句子对的语义相似度. 近年来，随着计算机硬件计算能力的提升，集成大规模外部语料知识的预训练语言模型成为语义相似度计算的主流方法，该方法通常将语义相似度计算转换成分类问题，通过给两段文本打上标签判断两段文本是否相似. 常见的有 ELMo(embeddings from language models)预训练模型[18]、GPT(generative pre-training)预训练模型[19]和 BERT(bidirectional encoder representations from transformers)预训练模型[20]等. 预训练语言模型生成文本向量的主要思想是将模型用于大规模语料上训练，通过强大的学习能力学习到较充分的文本语义表示，再在下游具体任务中微调参数，使模型在面对不同的文本输入时，能根据上下文信息得到不同的文本向量表示. 以 BERT 为例，BERT 预训练模型可同时接收两个文本输入，经过编码得到两种文本各自的向量表示后再进行融合，得到文本融合语义表示后输入全连接层，计算两个文本是否相似的概率. Xu 等[21]先通过 BERT 预训练模型对两个文本进行编码，再通过对两种文本的向量进行交互得到注意力权重，根据注意力权重得到新的文本表征，最后将两个文本向量进行拼接输入到全连接层中.

目前，基于字符串的方法主要停留在文本的字符串和单词层面，虽然这类方法原理直观简单，易实现，但仅停留在文本的表面，并未考虑到文本的深层语义信息，在语法结构复杂场景下的计算效果较差. 基于统计的方法主要通过向量空间模型和主题模型对文本进行建模，再将建模后的向量转换为相似度，对领域的依赖性虽然不强，但该类方法只能对文本进行浅层语义分析. 深度学习方法虽然可对文本进行高质量的向量表征，但应用于新闻文本与评论的相似度计算时，效果较差，这是因为新闻文本与评论之间存在较大的长度差异，使模型很难准确地学习到两种文本之间的语义关系，并且 BERT 预训练模型在处理超长文本序列时会截断一部分文本，无法对新闻文本进行完整的语义建模. 受 Peinelt 等[22]工作的启发，本文认为新闻文本与评论之间同样存在主题相关性，新闻标题也含有丰富的语义信息. 因此本文提出一种在融入新闻标题信息基础上将 TextRank[23]算法、LDA 主题模型与 BERT 预训练模型相结合的方法，将新闻内容与评论之间的语义相似度计算转换为计算新闻文本和新闻评论的主题相似度. 本文将该任务视为分类任务，首先将新闻文本通过 TextRank 算法提取出 k 个关键词，再将 k 个关键词与新闻标题进行拼接得到新的新闻文本，然后将新的新闻文本与评论分别输入到 LDA 主题模型和 BERT 模型中，得到两种文本各自的主题分布及其文本融合表示，最后将两种文本的主题分布和文本融合表示一起输入到全连接层，使用 Softmax 函数得到评论与新闻文本是否相关的概率.

1　模型介绍

新闻文本与新闻评论的相关性分析旨在研究通过计算新闻文本和新闻评论的语义相似度从大量评论中筛选出与新闻文本语义相关的评论. 传统文本相似度方法首先将文本中的每个词转换为词向量，然后通过计算向量之间的余弦相似度得到文本之间的语义相似度. 但当两个文本之间的长度相差较大时效果通常较差，而新闻文本与新闻评论文本长度常存在较大差别，因此本文将新闻内容与评论之间的语义相似度计算方法转换为计算新闻文本和新闻评论的主题相似度. 同时为更好地利用新闻标题中包含的语义信息，在新闻文本表示中融入新闻标题的语义信息，以得到更准确的新闻文本表征. 整体模型结构如图 1 所示.

模型主要分为三部分，即新闻文本内容表示、新闻文本与评论主题分布计算和 BERT 语义编码. 首先，用 TextRank 对新闻文本内容做关键词抽取，通过计算新闻文本中每个词的得分进行关键词抽

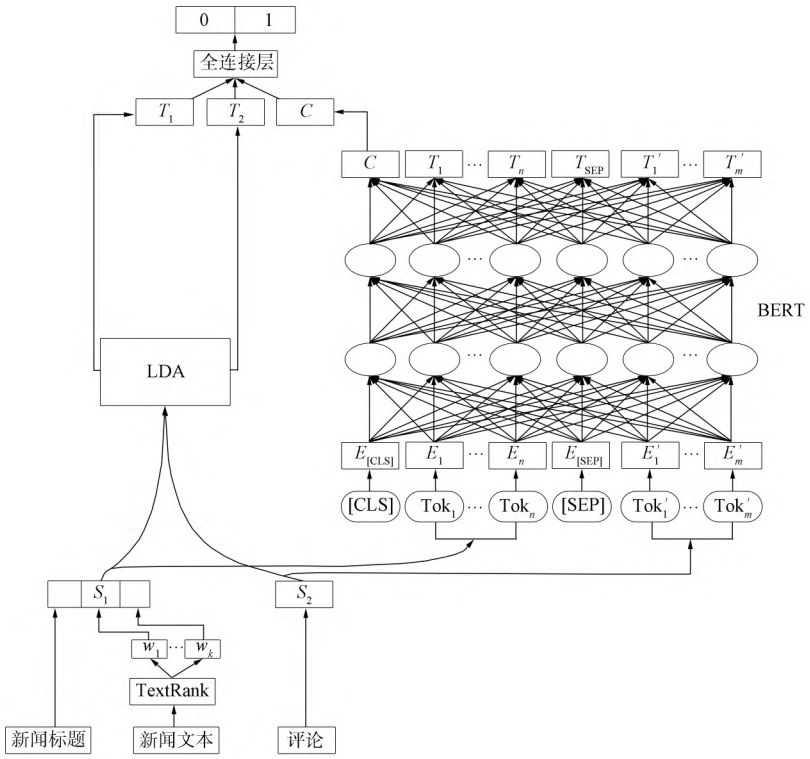

图1 模型示意图

Fig. 1 Schematic diagram of model

取. 抽取关键词过程可用公式表示为

$$WS(V_i)=(1-d)+d\times\sum_{j\in \mathrm{In}(V_j)}\left(\frac{1}{|\mathrm{Out}(V_j)|}\times WS(V_j)\right), \tag{1}$$

其中：V_i 和 V_j 分别表示新闻的第 i 个词和第 j 个词；$WS(V_i)$和 $WS(V_j)$分别表示 V_i 和 V_j 的分数；d 为阻尼系数，防止一些词的得分为 0；$\mathrm{In}(V_i)$表示所有指向节点 V_i 的节点集合，即入链集合；$\mathrm{Out}(V_j)$表示 V_j 所指的所有节点集合，即出链集合，$|\mathrm{Out}(V_j)|$表示出链数量.

首先，将新闻文本通过 jieba 分词工具进行分词处理，然后通过设置窗口向右滑动建立词与词之间的链接关系. 每个词的得分都要平均贡献给每个与其链接的词，最后从所有词中选出得分最高的前 k 个词作为新闻的关键词，用公式表示为

$$w_1,w_2,\cdots,w_k=\mathrm{Top}_k(WS(V_1,V_2,\cdots,V_n)), \tag{2}$$

其中 $w_1,w_2,\cdots,w_k$ 表示从文章中提取出的 k 个关键词，$WS(V_1,V_2,\cdots,V_n)$表示所有词的分数. 为充分利用标题包含的丰富语义信息，将标题与提取的 k 个关键词拼接得到新的新闻文本表示，从而在不丢失语义信息的情况下，将较长的新闻文本转换成较短的文本，即图 1 中的 S_1，用公式表示为

$$S_1=(\text{新闻标题},w_1,w_2,\cdots,w_k). \tag{3}$$

为计算新的新闻文本与评论之间的主题分布，对 S_1 和评论 S_2 分别做分词、去除停用词处理，并输入到 LDA 主题模型中，从而计算出 S_1 和 S_2 的主题分布向量 $\boldsymbol{T}_1$ 和 $\boldsymbol{T}_2$：

$$\boldsymbol{T}_1=\mathrm{LDA}(S_1),\qquad \boldsymbol{T}_2=\mathrm{LDA}(S_2). \tag{4}$$

在计算 S_1 和 S_2 主题分布的同时，将 S_1 和 S_2 输入 BERT 模型中，获得两种文本的融合语义表示 $\boldsymbol{C}$：

$$\boldsymbol{C}=\mathrm{BERT}(S_1,S_2). \tag{5}$$

其次，将 BERT 编码的向量与 LDA 模型计算的主题分布向量拼接得到最终的向量表示，并输入全连接层中，用 Softmax 函数转换为分类概率：

$$y=\mathrm{Softmax}(\boldsymbol{W}([\boldsymbol{T}_1,\boldsymbol{T}_2,\boldsymbol{C}])+\boldsymbol{b}), \tag{6}$$

其中 y 表示分类概率，$\boldsymbol{W}$ 为权重，$\boldsymbol{b}$ 为偏置.

最后，通过交叉熵损失函数更新模型参数.

2　实验结果与分析

2.1　实验数据集

由于目前缺少新闻文本与评论的相关性分析公共数据集，因此本文首先利用爬虫技术从网易新闻爬取了 198 篇新闻文章以及对应的评论，涵盖了数码、金融、娱乐明星、刑事案件、民事纠纷、新冠肺炎和国际新闻等内容；然后对数据集进行人工标注，用标签 0 表示评论与新闻文本不相关，标签 1 表示评论与新闻文本相关，只要评论出现新闻中的人物、地名、组织机构、事件以及用户的主观意见，均可视为与新闻相关；最后将数据集随机划分为训练集、验证集和测试集，数据集分布列于表 1.

表 1　数据集分布

Table 1　Distribution of datasets

类别及总数	训练集	验证集	测试集
1	6 488	2 478	2 076
0	3 512	2 483	2 885
总数	10 000	4 961	4 961

2.2　评价指标

本文采用准确率、精确率、召回率和 F_1 值作为评价指标，定义为

$$\text{准确率}=\frac{\mathrm{TP}+\mathrm{TN}}{N}, \tag{7}$$

$$\text{精确率}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}}, \tag{8}$$

$$\text{召回率}=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}, \tag{9}$$

$$F_1=\frac{2\times\text{精确率}\times\text{召回率}}{\text{精确率}\times\text{召回率}}, \tag{10}$$

其中 TP 表示模型正确预测评论与新闻相关的样本数，FP 表示模型错误预测评论与新闻相关的样本数，TN 表示模型正确预测评论与新闻不相关的样本数，FN 表示模型错误预测评论与新闻不相关的样本数，N 表示样本总数.

2.3　参数设置

设置文本最大序列长度为 200，Batch 大小为 24，学习率为 2×10^{-5}，关键词数量设为 20. 将文本输入主题模型前先对文本进行分词、去除停用词等预处理，训练时使用交叉熵损失函数更新模型参数.

2.4　消融实验

本文模型在 TextRank，BERT 和 LDA 主题模型的基础上融入了新闻标题信息，因此需设置实验验证融入新闻标题信息的有效性，实验结果列于表 2. 由表 2 可见，BERT 模型在融入新闻标题信息时 4 个指标均有提高，表明融入新闻标题信息可提高新闻文本与评论的语义相似度. 这是因为标题本身相当于外部信息，融入该外部信息理论上可提高语义相似度模型的性能，因此实验结果比未融入标题信息的 BERT 模型效果更好.

表 2　不同模型的消融实验结果 1

Table 2　Results 1 of ablation experiment of different models

模型	准确率	精确率	召回率	F_1 值
BERT(未融入标题信息)	0.816 7	0.845 1	0.829 2	0.837 1
BERT(已融入标题信息)	0.824 9	0.845 9	0.851 4	0.848 6

下面通过另一组消融实验分别验证 TextRank 算法与 LDA 主题模型的有效性. 本文在融入标题信息的基础上设置 3 组实验与本文模型进行对比. 3 组实验信息如下：

1) 融入标题信息的 BERT，该实验的目的是将 TextRank 算法与 LDA 主题模型分离出来. 将新闻标题信息与新闻文本进行拼接得到新的新闻文本，并与评论一起输入到 BERT 模型中计算新的新闻文本表示与评论之间的语义相似度.

2) 融入标题信息的BERT与TextRank算法组合，该实验目的是检验TextRank算法的有效性. 将新闻文本利用TextRank算法提取出前k个关键词，再将k个关键词与新闻标题一起组成新的新闻文本，最后与评论一起输入BERT模型中计算语义相似度.

3) 融入标题信息的BERT与LDA组合，该实验目的是检验LDA主题模型的有效性. 将新闻标题信息与新闻文本进行拼接得到新的新闻文本后，与评论分别输入到LDA主题模型和BERT模型中，获得两种文本的主题分布向量及其文本融合表示，最后输入全连接层计算是否相关的概率.

3组实验结果列于表3. 由表3可见，在只结合LDA主题模型的情况下，本文模型与BERT模型相比4个指标均有提高，表明融入LDA主题模型有效. 其原因是利用LDA主题模型提取出两种文本的主题分布向量，加强了语义表示，再融入到BERT模型中，效果相对更好，从而表明可从新闻文本中提取出更多信息进一步提高新闻文本与评论语义相似度计算模型的性能.

表3 不同模型的消融实验结果2

Table 3 Results 2 of ablation experiment of different models

模型(已融入标题信息)	准确率	精确率	召回率	F_1值
BERT	0.820 2	0.815 5	0.863 1	0.838 6
TextRank+BERT	0.822 1	0.806 9	0.867 2	0.836 0
LDA+BERT	0.823 3	0.815 9	0.872 2	0.843 1
本文	0.824 6	0.826 8	0.874 3	0.849 9

在融入标题信息基础上将TextRank算法与LDA主题模型结合后，即为本文模型. 从实验结果可见，与其他组相比本文模型在4个指标上均有提高，表明了本文模型的有效性. 虽然TextRank算法忽略了文本顺序，但LDA主题模型的引入使模型的性能仍然有提高，因此本文模型表现相对较好.

2.5 对比实验

下面将本文模型与ABCNN(attention-based convolutional neural network)[24]，DecomposableAttention[25]和SiaGRU(siamese gated recurrent unit)[26]模型进行性能对比. ABCNN模型的核心思想是先利用宽卷积的方式捕获句子对的完整信息，再利用注意力机制捕获句子对之间的相互依赖关系；DecomposableAttention模型的核心思想是利用注意力机制捕获句子对中词与词之间的对应关系判断句子对之间的关系；SiaGRU模型的核心思想是利用两个权重共享的LSTM网络[27]将长度不一致的句子对编码成向量，从而计算句子之间的相似度. 不同模型的对比实验结果列于表4.

表4 不同模型的对比实验结果

Table 4 Comparative experimental results of different models

模型	准确率	精确率	召回率	F_1值
ABCNN	0.780 4	0.806 4	0.807 0	0.806 7
DecomposableAttention	0.780 6	0.784 4	0.846 1	0.814 1
SiaGRU	0.756 1	0.783 3	0.788 6	0.785 9
本文	0.824 6	0.826 8	0.874 3	0.849 9

由表4可见，本文模型与其他模型相比在4个指标上均有提高，证明了本文方法的有效性. 这是因为本文模型首先将较长的新闻文本通过TextRank算法转换为较短的文本，使数据更适用于BERT模型，此外，本文还通过LDA模型计算文本与评论的主题分布进一步加强语义表示，最后融入新闻标题信息，从而性能更好. 而ABCNN和DecomposableAttention模型的性能相对较低，可能是因为新闻文本与评论之间的长度相差较多，不能准确捕捉到文本与文本、词与词之间的关系. 而SiaGRU模型则忽略了文本的上下文信息.

2.6 公共数据集上的实验

下面在BQ(http://icrc.hitsz.edu.cn/info/1037/1162.htm)和LCQMC(http://icrc.hitsz.edu.cn/info/1037/1146.htm)两个中文公共数据集上进行实验，以检验本文模型的泛化能力. 两个数据集均由哈尔滨工业大学智能计算研究中心构造并公开，数据集BQ含有12万条从金融领域采集的问题对，其中10万条为训练集，1万条为验证集，1万条为测试集. 数据集LCQMC覆盖了更多领域的问

题匹配，含有 260 068 条手工标注的问题对，其中 238 766 条为训练集，8 802 条为验证集，12 500 条为测试集. 之所以采用问题匹配的数据集，是因为在问答领域中，同样需要将输入的问题与设定好的问题进行语义相似度计算，如果相似就寻找问题库中的答案. 评价指标仍采用准确率、精确率、召回率和 F_1 值. 不同模型在两个数据集上的实验结果分别列于表 5 和表 6. 由表 5 和表 6 可见，本文模型在两个数据集上的实验结果与其他模型相比性能均较好，表明本文模型有一定泛化能力. 这是因为本文模型使用的 LDA 主题模型不需要依赖领域数据即可推断出主题分布，因此泛化能力相对更好.

表 5　不同模型在数据集 BQ 上的实验结果

Table 5　Experimental results of different modls on BQ dataset

模型	准确率	精确率	召回率	F_1 值
ABCNN	0.779 5	0.808 6	0.732 4	0.768 6
DecomposableAttention	0.797 0	0.791 4	0.806 6	0.798 9
SiaGRU	0.790 8	0.823 8	0.739 8	0.779 6
本文	0.840 4	0.841 9	0.838 2	0.840 0

表 6　不同模型在数据集 LCQMC 上的实验结果

Table 6　Experimental results of different models on LCQMC dataset

模型	准确率	精确率	召回率	F_1 值
ABCNN	0.811 5	0.757 9	0.915 4	0.829 3
DecomposableAttention	0.815 5	0.751 1	0.943 8	0.836 5
SiaGRU	0.829 8	0.772 7	0.934 4	0.845 9
本文	0.863 9	0.815 9	0.940 0	0.873 5

综上所述，针对预训练模型在处理新闻这种长文本时会截断一部分文本，导致文本信息缺失的问题，本文提出了一种结合 TextRank、LDA 主题模型和 BERT 预训练模型的新闻文本与评论语义相似度计算方法，同时融入了新闻标题信息. 实验结果表明了本文方法的有效性.

参　考　文　献

[1] 张雷，崔荣一. 基于编辑距离的词序敏感相似度度量方法 [J]. 延边大学学报(自然科学版)，2020(2)：140-144.（ZHANG L，CUI R Y. A Word Order Sensitive Similarity Measure Based on Edit Distance [J]. Journal of Yanbian University (Natural Science)，2020(2)：140-144.）

[2] XU X H，CHEN L，HE P. Fast Sequence Similarity Computing with LCS on LARPBS [C]//International Symposium on Parallel and Distributed Processing and Applications. Berlin：Springer，2005：168-175.

[3] 周丽杰，于伟海，郭成. 基于词项语义组合的文本相似度计算方法研究 [J]. 计算机工程与应用，2016，52(19)：90-93.（ZHOU L J，YU W H，GUO C. Research on Text Similarity Calculation Strategy Based on Semantic Combination of Keywords [J]. Computer Engineering and Application，2016，52(19)：90-93.）

[4] KONGRAK G. *N*-Gram Similarity and Distance [C]//String Processing and Information Retrieval. Berlin：Springer，2005：115-126.

[5] 黄贤英，谢晋，龙姝言. 基于公共词块及 *N*-gram 模型的问句相似度算法 [J]. 重庆理工大学学报(自然科学)，2017，31(10)：175-179.（HUANG X Y，XIE J，LONG S Y. Question Similarity Algorithm Based on Common Chunks and *N*-Gram Model [J]. Journal of Chongqing University of Technology (Natural Science)，2017，31(10)：175-179.）

[6] NIWATTANAKUL S，SINGTHONGCHAI J，NAENUDORN E，et al. Using of Jaccard Coefficient for Keywords Similarity [J]. Lecture Notes in Engineering and Computer Science，2013，1(3)：13-15.

[7] 周艳平，李金鹏. 一种基于词向量及位置编码的 Jaccard 相似度算法 [J]. 青岛科技大学学报(自然科学版)，2020，41(6)：93-98.（ZHOU Y P，LI J P. Jaccard Similarity Algorithm Based on Word Embedding and Position Encoding [J]. Journal of Qingdao University of Science and Technology (Natural Science Edition)，2020，41(6)：93-98.）

[8] 黄承慧，印鉴，侯昉. 一种结合词项语义信息和 TF-IDF 方法的文本相似度量方法 [J]. 计算机学报，2011，34(5)：856-864.（HUANG C H，YIN J，HOU F. A Text Similarity Measurement Combining Word Semantic Information with TF-IDF Method [J]. Chinese Journal of Computers，2011，34(5)：856-864.）

[9] BLEI D M，NG A Y，JORDAN M I，et al. Latent Dirichlet Allocation [J]. Journal of Machine Learning

Research, 2012(3): 993-1022.

[10] 王振振, 何明, 杜永萍. 基于LDA主题模型的文本相似度计算[J]. 计算机科学, 2013, 40(12): 229-232. (WANG Z Z, HE M, DU Y P. Text Similarity Computing Based on Topic Model LDA [J]. Computer Science, 2013, 40(12): 229-232.)

[11] 张超, 陈利, 李琼. 一种PST_LDA中文文本相似度计算方法[J]. 计算机应用研究, 2016, 33(2): 375-377. (ZHANG C, CHEN L, LI Q. Chinese Text Similarity Algorithm Based on PST_LDA [J]. Application Research of Computers, 2016, 33(2): 375-377.)

[12] 付雨蛟. 基于融合主题信息的深度VAE算法的蒙古文短语义相似度计算[D]. 呼和浩特: 内蒙古大学, 2018. (FU Y J. Mongolian Short Text Semantic Similarity Calculation Based on Deep VAE Integrated with Topic Information [D]. Hohhot: Inner Mongolia University, 2018.)

[13] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient Estimation of Word Representations in Vector Space [EB/OL]. (2013-09-07)[2021-02-01]. https://arxiv.org/abs/1301.3781v3.

[14] 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 (EMNLP). [S.l.]: Association for Computational Linguistics, 2014: 1532-1543.

[15] 杨德志, 柯显信, 余其超, 等. 基于RCNN的问题相似度计算方法[J]. 计算机工程与科学, 2021, 43(6): 1076-1080. (YANG D Z, KE X X, YU Q C, et al. A Question Similarity Calculation Method Based on RCNN [J]. Computer Engineering & Science, 2021, 43(6): 1076-1080.)

[16] LIU L Q, WANG Q L, LI Y. Improved Chinese Sentence Semantic Similarity Calculation Method Based on Multi-feature Fusion: Regular Papers [J]. Journal of Advanced Computational Intelligence and Intelligent Informatics, 2021, 25(4): 442-449.

[17] 周圣凯, 富丽贞, 宋文爱. 基于深度学习的短文本语义相似度计算模型[J]. 广西师范大学学报(自然科学版), 2022, 40(3): 49-56. (ZHOU S K, FU L Z, SONG W A. Semantic Similarity Computing Model for Short Text Based on Deep Learning [J]. Journal of Guangxi Normal University (Natural Science Edition), 2022, 40(3): 49-56.)

[18] PETERS M E, NEUMANN M, IYYER M, et al. Deep Contextualized Word Representations [EB/OL]. (2018-02-05)[2021-02-20]. https://arxiv.org/abs/1802.05365.

[19] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving Language Understanding by Generative Pre-training [J]. Computer Science, 2018, 3(2): 324-336.

[20] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding [EB/OL]. (2018-10-11)[2021-02-20]. https://arxiv.org/abs/1810.04805.

[21] XU F F, ZHENG S T, YU T. Bert-Based Siamese Network for Semantic Similarity [J]. Journal of Physics: Conference Series, 2020, 1684: 012074-1-012074-12.

[22] 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. [S.l.]: Association for Computational Linguistics, 2020: 7047-7055.

[23] MIHALCEA R, TARAU P. Textrank: Bringing Order into Text [C]//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. [S.l.]: Association for Computational Linguistics, 2004: 404-411.

[24] YIN W P, 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.

[25] PARIKH A P, TÄCKSTRÖM O, DAS D, et al. A Decomposable Attention Model for Natural Language Inference [EB/OL]. (2016-09-25)[2021-03-01]. https://arxiv.org/abs/1606.01933.

[26] MUELLER J, THYAGARAJAN A. Siamese Recurrent Architectures for Learning Sentence Similarity [C]// Proceedings of the AAAI Conference on Artificial Intelligence. New York: ACM, 2016: 2786-2792.

[27] 马玉环, 张瑞军, 武晨, 等. 深度残差网络和LSTM结合的图像序列表情识别[J]. 重庆邮电大学学报(自然科学版), 2020, 32(5): 874-883. (MA Y H, ZHANG R J, WU C, et al. Expression Recognition of Image Sequence Based on Deep Residual Network and LSTM [J]. Journal of Chongqing University of Posts and Telecommunications (Natural Science Edition), 2020, 32(5): 874-883.)

(责任编辑: 韩 啸)