智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

融合领域特征的科技学术会议语义相似性计算方法

于润羽, 李雅文, 李昂

引用本文:
于润羽,李雅文,李昂. 融合领域特征的科技学术会议语义相似性计算方法[J]. 智能系统学报, 2022, 17(4): 737–743.
YU Runyu,LI Yawen,LI Ang. Semantic similarity computing for scientific and technological conferences[J]. *CAAI Transactions on Intelligent Systems*, 2022, 17(4): 737–743.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202203050

DOI: 10.11992/tis.202203050
网络出版地址: https://kns.cnki.net/kcms/detail/23.1538.TP.20220621.1156.008.html

融合领域特征的科技学术会议语义相似性计算方法

于润羽[1], 李雅文[2], 李昂[1]
(1. 北京邮电大学 智能通信软件与多媒体北京市重点实验室, 北京 100876; 2. 北京邮电大学 经济管理学院, 北京 100876)

摘 要: 针对目前的语义文本相似度计算方法难以准确估计科技学术会议语义相似性的问题, 提出了一种融合领域特征的科技学术会议语义相似度计算方法(siamese-BERT semantic similarity calculation algorithm fused with domain feature, SBFD)。通过实体识别和关键词抽取等方式获取会议的领域特征信息, 将其作为特征与会议信息共同输入到基于变换器的双向编码器表示网络(bidirectional encoder representations from transformers, BERT)中, 采用孪生网络(Siamese Network)的结构解决 BERT 的各向异性的问题, 并对网络的输出进行池化和标准化, 利用余弦相似度计算两个会议之间的相似程度。实验结果表明 SBFD 方法在不同数据集上都取得了较好的效果, 斯皮尔曼相关系数有一定程度的提高。

关键词: 科技学术会议; 深度学习; 自然语言处理; 语义学习; 知识抽取; 语义相似度; 预训练模型; 孪生网络
中图分类号: TP391 **文献标志码:** A **文章编号:** 1673−4785(2022)04−0737−07

中文引用格式: 于润羽, 李雅文, 李昂. 融合领域特征的科技学术会议语义相似性计算方法 [J]. 智能系统学报, 2022, 17(4): 737–743.
英文引用格式: YU Runyu, LI Yawen, LI Ang. Semantic similarity computing for scientific and technological conferences[J]. CAAI transactions on intelligent systems, 2022, 17(4): 737–743.

Semantic similarity computing for scientific and technological conferences

YU Runyu[1], LI Yawen[2], LI Ang[1]
(1. Beijing Key Laboratory of Intelligent Telecommunication Software and Multimedia, Beijing University of Posts and Telecommunications, Beijing 100876, China; 2. School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract: Aiming at the problem that the current semantic text similarity calculation methods have difficulty in calculating semantic similarity for scientific and technological conference data accurately, a siamese-BERT semantic similarity calculation algorithm for scientific and technological conferences fused with domain features (SBFD) is proposed in this paper. At first, the domain feature information of conference is obtained through entity recognition and keyword extraction, and it is input into the bidirectional encoder representations from transformers (BERT) network as a feature, together with conference information. The structure of the Siamese network is then used to solve the anisotropy problem of BERT. The output of the network is pooled and normalized, and finally the cosine similarity is used to calculate the similarity between the two conferences. Experimental results show that the SBFD algorithm achieves good results on different data sets, with the Spearman's rank correlation coefficient improved in a certain extent.

Keywords: science and technological conference; deep learning; natural language processing; semantic learning; knowledge extraction; semantic similarity; pre-training model; siamese network

科技大数据可以定义为与科研相关的活动产生的海量数据[1-2], 这些数据数量规模大, 特征属性多, 内容专业化[3]。科技学术会议数据包含某个领域内的论文集合, 利用自然语言处理技术挖掘科技学术会议的潜在信息, 判断会议之间的语义相似度, 进而构建知识图谱和画像, 可以帮助

收稿日期: 2022−03−24. **网络出版日期:** 2022−06−22.
基金项目: 国家重点研发计划项目(2018YFB1402600); 国家自然科学基金项目(61772083,61802028); 广西科技重大专项(桂科 AA18118054).
通信作者: 李雅文. E-mail: warmly0716@126.com.

科研人员快速获得有价值的科研信息。

语义文本相似度计算方面主要有基于字符串，基于统计机器学习，基于深度学习的方法。目前基于深度学习的方法应用最为广泛，也取得了最好的效果，但在科技学术会议数据上，常规的相似度计算方法并不能挖掘到潜在的语义信息，无法取得最优的效果。同时基于变换器的双向编码器表示网络（bidirectional encoder representations from transformers, BERT）作为自然语言处理领域目前最杰出的预训练模型，却在语义文本相似度上的表现并不是很好，本文提出了融合领域特征的科技学术会议语义相似度计算方法（siamese-BERT semantic similarity calculation algorithm fused with domain feature, SBFD），利用实体识别和关键词抽取等方法获取会议的领域特征信息，将其作为特征与会议信息共同输入到 BERT 网络中，采用孪生网络（Siamese Network）的结构解决 BERT 的各向异性的问题，并对网络的输出进行池化和标准化，利用余弦相似度计算两个会议之间的相似程度，可以有效提升科技学术会议语义相似度的计算性能。

本文的主要贡献如下：

1) 提出了一种融合领域特征的科技学术会议语义相似度计算方法，在预训练模型的基础上微调，提高语义文本相似度计算的准确性。

2) 利用关键词提取，命名实体识别等技术，获取会议中的领域信息，在序列输入层融合了会议的领域特征，提高语义文本相似度计算的准确性。

3) 采用孪生网络结构，解决 BERT 在相似度计算上表现不佳的问题，同时提高模型计算速度。

1　语义相似度计算的研究现状

语义文本相似度计算在文本分类、文本聚类[4-5]、问答系统[6-7]、机器翻译[8-9] 等各个自然语言处理研究分支上被广泛使用。语义文本相似度的计算方式主要有基于字符串，基于机器学习和基于深度学习几种方式。其中基于字符串的方法相对简单，直接对两个字符串原始文本进行比较，主要的计算方法有编辑距离[10-11]、Jaccard 相似度[12] 等，其原理简单，实现方便，但只能识别字符级别的信息，一般用于文本的快速匹配。基于统计的方法主要有 VSM 模型及 LDA[13-14] 模型等。基于深度学习的方法需要在分布式词向量的基础上进行，词向量技术就是将单词映射成可被神经网络识别的向量。Mikolov 等[15] 提出的 word2vec 是最早生成分布式词向量的方法，同时提供了对应的工具。Pennington 等[16] 提出 Glove 模型，Glove 基于语料库构建了单词的共现矩阵，利用概率论的计算方式，结合构建出的矩阵，计算得出最终的词向量。由于矩阵的构建综合了全局语料，因此 Glove 在一定程度上考虑了全局信息。

Peters 等[17] 提出（来自语言模型的嵌入）（embeddings from language models, ELMO）模型，其先用语言模型在一个大的语料库上学习好词的词向量，此时无法区分多义词。Vaswani 等[18] 提出在注意力机制上构建的变换器（Transformer）编码器模型。Radford 等[19] 提出 GPT 模型，引入了 Transformer 架构。Devlin 等[20] 提出的 BERT 模型在 Transformer 的基础上，引入了 mask 遮盖编码的思路和下句子预测方法，在生成动态词向量上取得了更好的表现。Huang 等[21] 提出了基于深度网络的语义模型方法，基于孪生网络架构，模型分为输入层、表示层、匹配层。Palangi等[22] 将长短期记忆网络（long short-term memory, LSTM）引入其中，LSTM 作为特殊的循环神经网络，能够考虑到距离更远的上下文信息和一些序列信息，提升了计算的效果。Pontes 等[23] 将卷积神经网络（convolutional neural networks, CNN）模型和 LSTM 模型同时用于孪生网络架构，利用该网络计算语义文本相似度。Reimers 等[24] 提出了 SBERT 网络结构，SBERT 模型完成仅需 5 s，带来了巨大的效率提升。Li 等[25] 从 BERT 训练向量结果的层面进行了分析，发现了 BERT 预训练出的词向量存在各向异性和低频词汇稀疏的问题，在 STS12-16 和 SICK-R 数据集上有更好的表现。

2　SBFD 计算方法

2.1　算法整体结构

将 BERT 模型输出的结果进行平滑修正，同时考虑到科技学术会议的特征，提出了 SBFD。该方法的整体结构如图 1 所示。SBFD 算法由序列输入层、神经网络层、池化及标准化层和相似度计算层 4 个部分组成。具体而言，序列输入层作为整个系统结构的输入，融入了领域信息的文本转化为向量，领域信息包括会议中论文研究的技术实体和会议的主要研究方向；在神经网络层采用了 BERT 网络模型，基于 BERT 中文预训练模型进行，在语义文本相似度场景下微调训练出最优的模型。接下来通过全局平均池化和标准化，缓解 BERT 模型训练结果的各向异性的问题；最后在相似度计算层，采用余弦相似度衡量文本相似度。下面将详细描述每层的具体功能与实现原理。

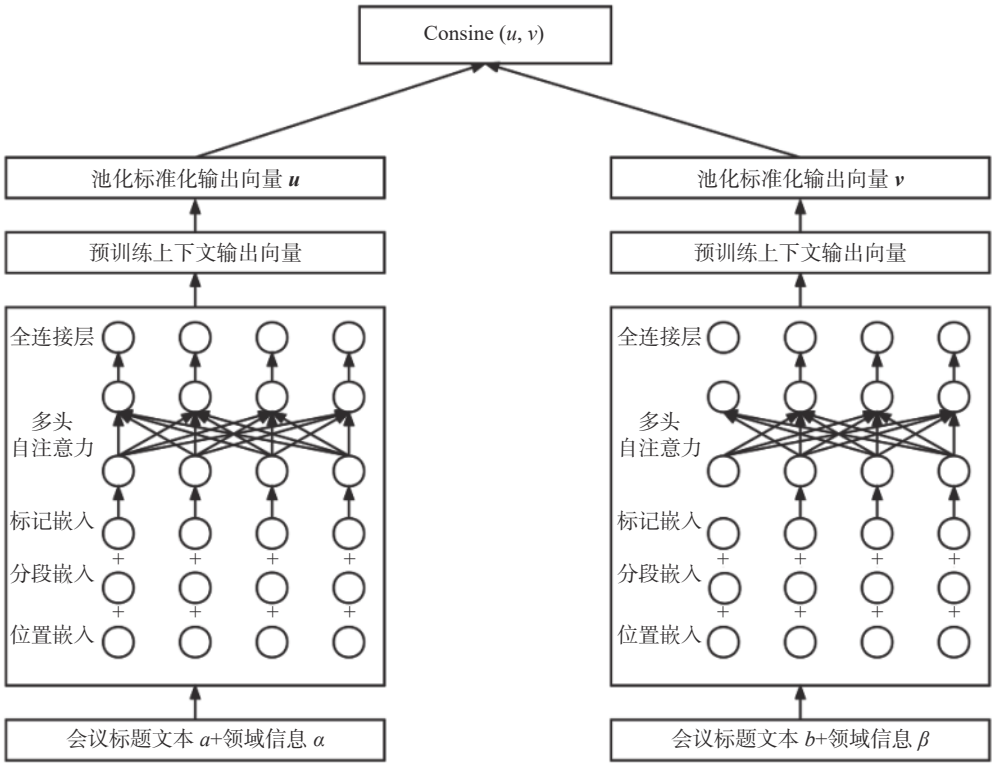

图 1　SBFD 算法整体框架

Fig. 1　Framework of SBFD algorithm

2.2　序列输入层

科技学术会议相似度计算序列输入层将需要对比的相似度的两个文本作为输入，传递到 Siamese 网络结构中，标准的 Siamese 网络结构如图 2 所示，输入为两个要比对的文本序列。孪生网络的优势在于结构简单，训练稳定。在本算法的实现中，根据数据的特点对序列做了优化，因为数据中存在两个会议研究方向相似度很高，但是会议名称单纯从文本层面并没有特别高的相似度，为了降低这个差异的影响，本文中结合了会议领域特征，共同作为序列的输入。如会议标题 1：计算机设计国际会议，2：代码生成和优化国际会议。均为计算机体系结构方面的会议，但会议名称并不能看出这两个会议之间的明显关联，计算机设计国际会议和研究信息安全方向的亚洲计算机与通信会议研究主题更加相近，因此将会议的领域信息，包括研究技术实体，研究主题共同作为序列的输入，传递到神经网络层中获取向量。

2.3　神经网络层

基于 Siamese 网络结构的网络层可以选取不同的神经网络来实现，例如最基本的 RNN 递归神经网络，或者在 RNN 基础上进行改进的传统 LSTM、双向 LSTM、LSTM+Attention、GRU 等，在本文的实现中采用了 BERT 作为网络层。BERT 是一种双向编码的预训练模型，训练时同时考虑到了上下文信息。它的网络架构基于 Transformer 编码器，采用了 Masked Language Model 和下一句预测两个训练任务训练网络参数，MLM 模型主要是用来获取上下文信息，它不像 CBOW 一样把所有的词都预测一遍，而是随机遮挡部分字符。传统的语言模型（language model, LM）的获取概率的公式为

$$\log p(x_{1:T}) = \sum_{t=1}^{T} \log p(x_t \mid c_t)$$

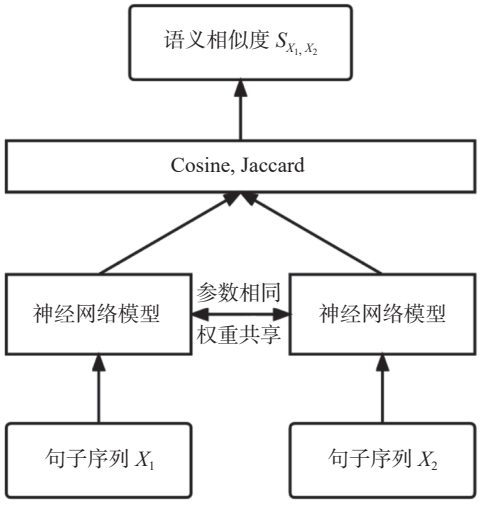

图 2　孪生网络结构示意图

Fig. 2　Structure of Siamese Network

式中：$x_{1:T}=(x_1,x_2,\cdots,x_T)$表示 token 序列，$c_t=x_{1t-1}$。传统语言模型计算 token 出现概率$p(x_{1T})$时，采用自回归方式进行因式分解，例如句子$x_1x_2x_3$出现的概率为$p(x_1)^*p(x_2\mid x_1)^*p(x_3\mid x_1x_2)$。

与传统语言模型不同，BERT 采用 MLM 模型屏蔽了输入文本中的一部分文本，训练出模型后再对被屏蔽的文本进行预测，其计算公式为

$$p(\bar{x}\hat{x})=\sum_{t=1}^{T}m_t p(x_t\mid c_t)$$

式中：$\hat{x}$为带有被掩盖的 token 的序列；$\bar{x}$为被掩盖的 token；m_t表示该 token 是否被掩盖，只有两种取值，其中 1 为是，0 为否。

MLM 模型是针对于单词量级的训练，有许多任务是在句子量级上的。这就需要语言模型理解句子之间的关系，BERT 的下一句预测任务是对于句子级别的任务的训练，利用二值预测方法预测句子 X 是否是句子 Y 的下一句。思路简单，但在句子级别问题，如智能问答上有很显著的效果。在完成以上两个部分的参数训练后，Bert 采用 Transformer 结构，其中编码单元的核心模块利用了自注意力机制。在 BERT 模型中，为了扩展模型专注于不同位置的表达能力，采用了 MultiHead，即“多头”模式，即

$$\begin{gathered}\text{MultiHead}\,(Q,K,V)=\\ \text{Concat}^2(\text{head}_1,\text{head}_2,\cdots,\text{head}_n)\\ \text{head}_i=\ \text{Attention}\left(Q\cdot W_i^Q,K\cdot W_i^K,V\cdot W_i^V\right)\end{gathered}$$

注意力层的输出即为 BERT 输出的高维向量。

2.4 池化及标准化层

BERT 网络输出的向量可能存在训练后的向量长度不相同的问题，难以计算这两个不同维度的结果之间的距离，因此采用全局平均池化提取句子级别的语义表示 U 和 V，得到 U 和 V 可以进行相似度计算。BERT 训练后的结果由于存在各向异性的问题，影响最终的预测效果，因此在这里增加一层向量标准化（normalized），定义了一个从潜在空间观测空间的可逆变换，标准化流的生成过程描述为

$$z\sim p_Z(z),\ u=f_\phi(z)$$

其中$p_Z(z)$为先验分布，$z->u$是可逆变换。

通过变量代换定理，可观测变量x的概率密度函数可以表示为

$$p_U(\boldsymbol{u})=p_z\left(f_\phi^{-1}(\boldsymbol{u})\right)\left|\det\frac{\partial f_\phi^{-1}(\boldsymbol{u})}{\partial\boldsymbol{u}}\right|$$

训练目标为最大化预定好的 BERT 句向量的似然函数，即

$$\log_z\left(f_\phi^{-1}(\boldsymbol{u})\right)+\log\left|\det\frac{\partial f_\phi^{-1}(\boldsymbol{u})}{\partial\boldsymbol{u}}\right|$$

式中：p_Z是标准高斯分布；$\boldsymbol{u}$是 BERT 句向量分布；det 为矩阵的行列式。

2.5 科技学术会议相似度计算层

余弦相似度将文本置于向量空间，更适合本文数据集，解释性较强，因此被选为算法的相似度计算层方法，衡量最后输出的两个文本向量的距离。

2.6 SBFD 计算方法步骤

输入：文档 D，包含 $2n$ 个句子文本序列α及其语义特征β，每行有两对文本序列及特征，用空格分隔，分别为α_1、β_1、α_1、β_2。

输出：余弦相似度序列

for $(\alpha,\beta)\in D$ do

特征拼接：$(\alpha_1,\beta_1)->e_1$，$(\alpha_2,\beta_2)->e_2$

孪生网络训练：

$\text{Bert\_left}(e_1)->h_1,\text{Bert\_right}(e_2)->h_2$

池化及标准化：

mean & normalized$(h_1)->o_1$，

mean & normalized$(h_2)->o_2$

相似度计算：

cosine similarity$(o_1,o_2)->v$

将结果添加到列表中：v add to list

return list

3 实验结果

3.1 数据集

本实验分为两部分，为了验证方法的泛化能力，在公开数据集上，对没有融合领域特征的方法进行了测试，采用的数据集包括广泛应用的 STS12-STS16 数据集，以及 SICK-R 数据集。科技学术会议数据集为从知网和万方爬取的文本内容，由于数据需要人工标注，因此在选取了信息科技 3 种学科下的数据 1 000 条，其中 800 条用于训练，200 条用于测试，由于受到标注的限制，整体数据量有限，将数据采用交叉验证的方式，取平均结果作为最终模型性能。

3.2 评价指标

语义文本相似度计算算法的指标一般采用相关系数，相关系数常用的有皮尔逊相关系数以及斯皮尔曼相关系数。皮尔逊相关系数，定义为两个变量的协方差除以它们标准差的乘积，计算公式为

$$\rho_{X,Y}=\frac{\text{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E\left[(X-\mu_X)(Y-\mu_Y)\right]}{\sigma_X\sigma_Y}$$

斯皮尔曼相关系数为式 (1)，对于样本容量为 n 的样本，n 个原始数据X_i，Y_i被转换成等级数据x_i、y_i，其中等级数据x_i、y_i是每个原始数据的降序位置的平均。

$$\rho_{X,Y}=\frac{\sum_i(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_i(x_i-\bar{x})^2\sum_i(y_i-\bar{y})^2}} \tag{1}$$

已有研究表明，皮尔逊相关系数由于对线性关系更加敏感，离心数据对整体评价指标影响较大，其相关性的内在评估可能会产生误导，皮尔逊相关系数并不是适合检测语义文本相似度任务的最佳选择。而斯皮尔曼相关系数更多的是衡量正相关关系，因此更加适合作为本节的评价指标，故本节采用斯皮尔曼相关系数对结果进行评估。

3.3　实验结果

在本文为实验中，BERT 的 batch size 设置为 32，学习率设置为 1×10^{-5}。向量输出的池化方式全局平均池化，优化器选择 Adam，Dropout 设置为 0.1。

本节使用斯皮尔曼相关系数对文本相似度分析的结果进行评价，对不同的方法效果进行评估，首先在 STS12-STS16 和 SICK-R 数据集数据集上，在孪生网络的结构中，对比了 GloVe，BERT，SBERT 和 SBERT 方法的效果，这里的 SBERT 并没有输入领域信息，主要为了验证整体网络框架的能力，实验结果如表 1 所示。

表 1　不同方法在 STS12-16 及 SICK-R 上的表现

Table 1　Performance of different methods on STS12-16 and SICK-R

算法	STS12	STS13	STS14	STS15	STS16	SICK-R
GloVe	0.5525	0.6728	0.6215	0.6746	0.6423	0.5608
BERT	0.4238	0.5766	0.5825	0.6322	0.6207	0.5889
SBERT	0.6881	0.7276	0.7322	0.7423	0.7133	0.7206
SBFD	0.6920	0.7321	0.7426	0.7618	0.7380	0.7354

表 1 中，SBERT 为 Siamese-BERT 算法，NS-BERT 为标准化后的 Siamese -BERT 算法。由表 1 可以看出 BERT 表现欠佳，甚至在某些数据集表现还不如模型更简单的静态词嵌入模型 GloVe，分析其中原因，常规的 BERT 训练出的结果不具有语义信息，导致两个相似的句子得到的句向量可能有很大的差别，为了解决这个问题，孪生网络的两侧分别输入需要计算语义相似度的两个句子，获得含有句子潜在语义的 Embedding。再将其输入到 BERT 网络中进行计算，可以看到效果对比与 GloVe 和 BERT 有非常明显地提升，证明孪生网络结构的有效性。最后，NSBERT 由于对输出的向量进行标准化，解决了 BERT 生成向量空间分布不均匀的问题，进一步提升了算法的表现。

根据表 1 的分析结果可以证明孪生网络结构在计算语义文本相似度的有效性，因此，在科技学术会议数据集上，均采用孪生网络作为网络框架，在神经网络层选择不同的网络结构进行比较，比较结果如表 2 所示。

表 2　不同网络结构相似度分析表现

Table 2　Similarity analysis performance of different network structures

计算领域	LSTM	LSTM+ATT	SBERT	SBFD_1	SBFD_2
信息科技	0.6036	0.6221	0.6682	0.7029	0.7525
工程科技	0.5258	0.5882	0.6236	0.6822	0.7057
农业科技	0.5620	0.6436	0.6822	0.6918	0.7231
平均	0.5638	0.6178	0.6580	0.6923	0.7271

由表 2 的分析可知，不同的神经网络模型在文本相似度计算领域，对于孪生网络有很大的影响。SBFD_1 不考虑领域信息的 SBFD 方法，SBFD_2 表示包含领域信息的 SBFD 方法，SBFD 在不同的网络结构中取得了更好的效果，对比方法包括 LSTM、LSTM+ATT，通过 LSTM 识别，可能会忽略掉长距离依赖的信息和下文信息，因此在 LSTM 基础上结合注意力机制，可以考虑到全局语义信息，提高相似度计算的效果。SBERT 利用 BERT 作为孪生网络的神经网络层，预训练模型提取特征，也取得了很好的效果。最后对比 SBFD 方法，对 BERT 输出进行标准化，就取得了更好的效果，解决向量分布异化的问题。引入领域信息带来了效果的提升，验证了本文方法的有效性。

3.4　实验参数对性能的影响

将 BERT 的 batch size 设置不同数值进行实验，确定其对论文数据命名实体识别效果的影响，在信息科技数据集实验结果如图 3 所示。

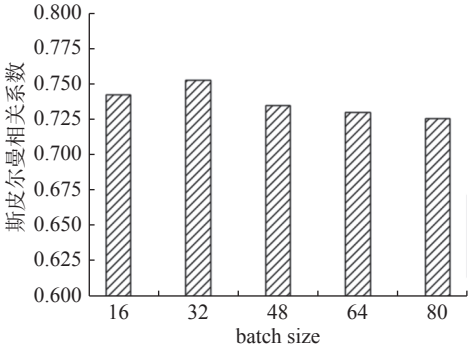

图 3　batch size 对信息科技数据集计算效果影响

Fig. 3　Impact of batch size on the computing effect of information technology datasets

根据图 3 可以看出，在信息科技数据集中，batch size 在 32 时方法取得了最好的效果，随着 batch size 的升高，效果有一定的下滑，但下滑并不明显。因此从信息科技数据集上看，batch size 对于方法的效果有一定影响，但并不是正相关或者负相关的趋势，影响程度有限。接下来在工程科技数据集上进行实验，结果如图 4 所示。

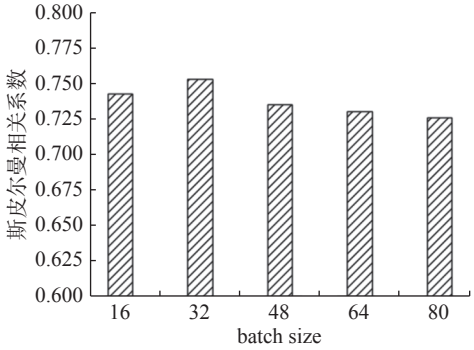

图 4　batch size 对工程科技数据集计算效果影响

Fig. 4　Impact of batch size on the calculation effect of engineering science and technology datasets

从图 4 可以看出，在工程科技数据集中，batch size 在 16 时方法取得了最好的效果，与在信息科技数据集中有一定的区别，但是从分布趋势上来看是一致的，即 batch size 对于性能有影响，但影响不明显，在某一个值时取得最好效果，随着其继续增大，效果对比 16 时有一定的下降。

接下来在农业科技数据集上进行实验，结果如图 5 所示。

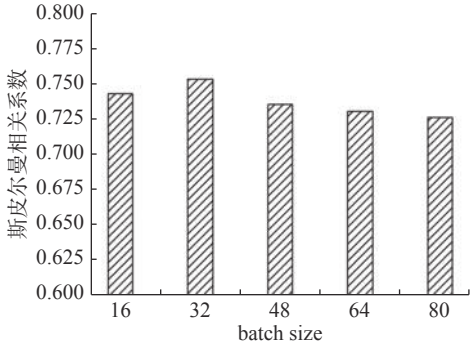

图 5　batch size 对农业科技数据集计算效果影响

Fig. 5　Impact of batch size on the computing effect of agricultural science and technology datasets

从图 5 可以看出，在农业科技数据集中，batch size 在 32 时取得了最好的效果，结合图 3、4、5 分析可以得出结论：batch size 对整个网络的识别效果有一定影响，batch size 越大，训练速度越快，但太大的 batch size 导致模型泛化能力下降。在一定范围内，增加 batch size 有助于收敛的稳定性是随着 batch size 的增加，方法的性能会下降，对精度造成影响。在训练速度允许的范围内，尽可能的选择一个合适的 batch size，可以提高方法的性能。

4　结束语

本文提出了一种融合领域特征的科技学术会议语义相似度计算方法，结合科技学术会议自身的特点，融入研究领域特征，共同作为向量输入。基于孪生网络结构对文本语义相似度进行计算，其中神经网络层选择了 BERT 模型，充分利用了 BERT 预训练模型的优势，挖掘文本中的深层语义信息。同时，考虑到 BERT 在文本相似度计算上计算开销大、准确性的缺点，基于孪生网络结构，构建了 SBERT 网络，并对训练出的向量进行标准化，计算文本的相似度。将 SBFD 方法与 Glove、LSTM 等方法比较，实验结果表明，SBFD 在学术会议数据集上有更好的表现。相似度计算的结果可以用于学术会议知识图谱和画像的构建，帮助科研人员快速从中获取到想要的科研信息。

参考文献：

[1] 周园春, 王卫军, 乔子越, 等. 科技大数据知识图谱构建方法及应用研究综述 [J]. 中国科学: 信息科学, 2020, 50(7): 957−987.
ZHOU Yuanchun, WANG Weijun, QIAO Ziyue, et al. A survey on the construction methods and applications of sci-tech big data knowledge graph[J]. Scientia sinica (informationis), 2020, 50(7): 957−987.

[2] 苏晓娟, 张英杰, 白晨, 等. 科技大数据背景下的中英双语语料库的构建及其特点研究 [J]. 中国科技资源导刊, 2019, 51(6): 87−92.
SU Xiaojuan, ZHANG Yingjie, BAI Chen, et al. Research of bilingual corpus construction and its characteristics in big data[J]. China science & technology resources review, 2019, 51(6): 87−92.

[3] 胡吉颖, 谢靖, 钱力, 等. 基于知识图谱的科技大数据知识发现平台建设 [J]. 数据分析与知识发现, 2019, 3(1): 55−62.
HU Jiying, XIE Jing, QIAN Li, et al. Constructing big data platform for sci-tech knowledge discovery with knowledge graph[J]. Data analysis and knowledge discovery, 2019, 3(1): 55−62.

[4] TONG Yuqiang, GU Lize. A news text clustering method based on similarity of text labels[M]//Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering. Cham: Springer International Publishing, 2019: 496-503.

[5] LI Wenling, JIA Yingmin, DU Junping, et al. Distributed multiple-model estimation for simultaneous localization and tracking with NLOS mitigation[J]. IEEE transactions on vehicular technology, 2013, 62(6): 2824–2830.

[6] DAS A, MANDAL J, DANIAL Z, et al. A novel approach for automatic Bengali question answering system using semantic similarity analysis[J]. International journal of speech technology, 2020, 23(4): 873–884.

[7] FANG Yuke, DENG Weihong, DU Junping, et al. Identity-aware CycleGAN for face photo-sketch synthesis and recognition[J]. Pattern recognition, 2020, 102: 107249.

[8] QIAN Ming, LIU J, LI Chaofeng, et al. A comparative study of English-Chinese translations of court texts by

machine and human translators and the Word2Vec based similarity measure's ability to gauge human evaluation biases[C]//Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks. Dublin: ACL, 2019: 95−100.

[9] XUE Zhe, DU Junping, DU Dawei, et al. Deep low-rank subspace ensemble for multi-view clustering[J]. Information sciences, 2019, 482: 210–227.

[10] RISTAD E S, YIANILOS P N. Learning string-edit distance[J]. IEEE transactions on pattern analysis and machine intelligence, 1998, 20(5): 522–532.

[11] HU Weiming, GAO Jun, LI Bing, et al. Anomaly detection using local kernel density estimation and context-based regression[J]. IEEE transactions on knowledge and data engineering, 2020, 32(2): 218–233.

[12] SUPHAKIT Niwattanakul, JATSADA Singthongchai, EKKACHAI Naenudorn, et al. Using of Jaccard coefficient for keywords similarity[C]//Proceedings of the international multiconference of engineers and computer scientists. Hong Kong: Newswood Limited, 2013, 1(6): 380−384.

[13] KOU Feifei, DU Junping, HE Yijiang, et al. Social network search based on semantic analysis and learning[J]. CAAI transactions on intelligence technology, 2016, 1(4): 293–302.

[14] LI Wenling, JIA Yingmin, DU Junping. Variance-constrained state estimation for nonlinearly coupled complex networks[J]. IEEE transactions on cybernetics, 2018, 48(2): 818–824.

[15] MIKOLOV T, CHEN KAI, CORRADO G, et al. Efficient estimation of word representations in vector space[EB/OL]. New York: arXiv, 2013. (2013−01−16)[2022−03−24]. https: //arxiv. org/abs/1301.3781.

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

[17] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[EB/OL]. New York: arXiv, 2018. (2018−03−22)[2020−07−01]. https://arxiv.org/abs/1802.05365.

[18] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all You need[C]//NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems. New York: ACM, 2017: 6000−6010.

[19] RADFORD A, NARASIMHAN K. Improving language understanding by generative pre-training[EB/OL]. (2018−11−05)[2020−07−01].https://www.semanticscholar.org/paper/Improving-Language-Understanding-by-Generative-Radford-Narasimhan/cd18800a0fe0b668a1cc19f2ec95b5003d0a5035.

[20] DEVLIN J, CHANG MING-WEI, LEE K, et al. BERT: pre-training of deep bidirectional transformers for language understanding[EB/OL]. New York: arXiv, 2018. (2018−10−11)[2022−03−20]. https: //arxiv. org/abs/1810.04805.

[21] HUANG Posen, HE Xiaodong, GAO Jianfeng, et al. Learning deep structured semantic models for web search using clickthrough data[C]//CIKM'13: Proceedings of the 22nd ACM international conference on Information & Knowledge Management. New York: ACM, 2013: 2333−2338.

[22] PALANGI H, DENG L, SHEN Y, et al. Semantic modelling with long-short-term memory for information retrieval[EB/OL]. New York: arXiv, 2014. (2014−12−20)[2022−03−20].https://arxiv.org/abs/1412.6629.

[23] PONTES E L, HUET S, LINHARES A C, et al. Predicting the semantic textual similarity with Siamese CNN and LSTM[EB/OL]. New York: arXiv, 2018. (2018−10−24) [2022−03−20]. https: //arxiv. org/abs/1810.10641.

[24] REIMERS N, GUREVYCH I. Sentence-BERT: sentence embeddings using Siamese BERT-networks[EB/OL]. New York: arXiv, 2019. (2019−08−27) [2022−03−20]. https: //arxiv. org/abs/1908.10084.

[25] LI BOHAN, ZHOU HAO, HE JUNXIAN, et al. On the sentence embeddings from pre-trained language models[EB/OL]. New York: arXiv, 2020. (2020−11−02)[2022−03−20]. https: //arxiv. org/abs/2011.05864.

作者简介：

于润羽，硕士研究生，主要研究方向为深度学习、数据挖掘。

李雅文，副教授，主要研究方向为企业创新、人工智能、大数据。

李昂，博士研究生，主要研究方向为信息检索、数据挖掘、机器学习。