计算机科学
COMPUTER SCIENCE

基于关键语义驱动和对比学习的文本聚类方法

张士举, 郭朝阳, 吴承亮, 吴凌俊, 杨丰玉

引用本文

张士举, 郭朝阳, 吴承亮, 吴凌俊, 杨丰玉. 基于关键语义驱动和对比学习的文本聚类方法[J]. 计算机科学, 2025, 52(8): 171-179.

ZHANG Shiju, GUO Chaoyang, WU Chengliang, WU Lingjun, YANG Fengyu. Text Clustering Approach Based on Key Semantic Driven and Contrastive Learning [J]. Computer Science, 2025, 52(8): 171-179.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240700008

基于关键语义驱动和对比学习的文本聚类方法

张士举[1] 郭朝阳[2] 吴承亮[2] 吴凌俊[2] 杨丰玉[1,2]
1 南昌航空大学软件学院 南昌 330000
2 江西省航空制造数字化仿真工程研究中心 南昌 330000
(2804286469@qq.com)

摘 要 文本聚类是指将大量文本数据按照它们的相似性进行分组的过程,其可以帮助理解文本数据的结构和内容,发现其中的模式和趋势,通常用于信息检索、文档管理等。现有文本聚类模型在信息抽取过程中存在过度依赖原始数据质量和容易造成关键信息提取不充分的问题,而且不同类别的数据在表示空间中会相互重叠。针对以上问题,提出了一种基于关键语义驱动和对比学习的文本聚类方法(KSD-CLTC)。该方法在数据处理环节通过数据增强模块丰富原始数据来提高泛化性,并设计了一个关键语义驱动模块提取文本中的关键词,补足关键语义信息的丢失;在特征提取环节借助预训练模型和自动编码器对数据进行高质量表征;然后,在聚类学习环节借助聚类模块将聚类损失与关键语义驱动模块的重构损失相结合,进一步学习更适用于聚类的特征表示,并利用对比学习模块来实现更好的类别划分效果。实验结果表明,KSD-CLTC 在公共数据集和工业数据集上的聚类效果优于对比的聚类算法,相比先进的 SCCL 方法,其在所有数据集上的 ACC 平均提高了 2.92%,NMI 平均提高了 1.99%。聚类结果也证明了关键语义驱动模块对文本聚类的重要性。

关键词:信息抽取;表示空间;文本聚类;关键语义驱动;对比学习
中图分类号 TP391.9

Text Clustering Approach Based on Key Semantic Driven and Contrastive Learning

ZHANG Shiju[1],GUO Chaoyang[2],WU Chengliang[2],WU Lingjun[2] and YANG Fengyu[1,2]
1 College of Software,Nanchang Hangkong University,Nanchang 330000,China
2 Jiangxi Province Aviation Manufacturing Digital Simulation Engineering Research Center,Nanchang 330000,China

Abstract Text clustering is the process of grouping a large amount of text data according to their similarities,which can help to understand the structure and content of text data,and discover patterns and trends in it,and is usually used in the fields of information retrieval and document management. Existing text clustering models have the problems of over-reliance on the quality of original data and insufficient extraction of key information,and data of different categories overlap each other in the representation space. To solve the above problems,a text clustering method based on key semantic-driven and comparative learning(KSD-CLTC) is proposed. In the process of data processing,a data enhancement module is used to enrich the original data to improve the generalization,and a key semantic-driven module is designed to extract keywords from the text to make up for the loss of key semantic information. In the feature extraction process,the pre-trained model and automatic encoder are used to characterize the data with high quality. Then,in the cluster learning process,the cluster loss is combined with the reconstruction loss of key semantic-driven modules to further learn the feature representation more suitable for clusters,and the contrast learning module is used to achieve better classification results. KSD-CLTC outperforms the comparative clustering algorithms on both public and industrial datasets,improving ACC by an average of 2.92% and NMI by an average of 1.99% across all datasets as compared to the state-of-the-art SCCL method. The clustering results also demonstrate the importance of key semantic drivers for text clustering.

Keywords Information extraction,Denote space,Text clustering,Key semantic-driven,Contrastive learning

聚类的主要目的是对样本进行分组,使相似的样本属于同一类,不相似的样本属于不同类。样本的聚类提供了数据的全局特征,这对进一步分析整个数据集有重要意义。其已在诸多领域得到了广泛引用,如异常检测[1]、领域适应[2]、社区检测[3]和表示学习[4]等。

聚类是无监督学习中的基础难题,已经被广泛研究了几

到稿日期:2024-07-01 返修日期:2024-09-25
基金项目:江西省重点研发计划(20202BBEL53002)
This work was supported by the Key Research and Development Program of Jiangxi Province(20202BBEL53002).
通信作者:杨丰玉(99770277@qq.com)

十年。基于距离的聚类方法，如 K-means[5] 和高斯混合模型(GMM)[6]依赖于数据空间中测量的距离。还有一些基于密度、词图等方式的聚类方法，如 DBSCAN[7]、OPTICS[8]、谱聚类[9]等。传统聚类方法一般使用统计或者权重计算的方式学习文本中的信息，随着数据的复杂化，传统的方法已经不能有效实现数据的表征[10-11]。

随着深度学习特别是深度无监督学习的巨大成功，许多基于深度架构的表示学习技术被提出。该方法首先学习深度表示，再将其输入聚类方法。然而，这样的表示方法并非直接学习服务于聚类任务的表示，一定程度上限制了聚类性能。一般情况下，学习到的表示要适配于下游任务，因此目前很多研究都是基于表示和深度聚类相结合的方式进行不断的优化[12-14]。

目前许多研究都通过优化文本表示和聚类的目标来提升聚类的效果，但由于文本数据的特性，现有方法仍然存在以下问题[15]。1)语义信息不完整。现有深度聚类方法在将原始高维稀疏的文本数据压缩到低维空间时，可能会导致关键信息提取不充分[16]。2)表示与聚类之间缺乏相关性。特征表示并非直接学习聚类，限制了聚类的性能。3)不同类别数据重叠。不同类别的数据在表示空间中通常会相互重叠，导致很难得到高纯度的聚类结果。

关键信息即关键语义，通常指的是在文本中对理解内容至关重要的语义元素。识别聚类任务中的关键语义对类别的划分十分重要。传统的聚类模型大多利用句法或词法分析达到聚类的目的，这种方法较难捕获到文本与对应类别之间的语义关联信息。此外，以往的研究通过注意力机制的方式优化文本表示，这种方式虽然可以强化文本与类别之间的关联程度，但是在一定程度上忽略了两者之间关键信息的联系。

针对上述问题，本文提出一种基于关键语义驱动和对比学习的文本聚类方法(Key Semantics-Driven and Contrastive Learning based Method of Text Clustering，KSD-CLTC)。首先，为了解决关键信息提取不充分的问题，使用关键语义驱动模块对全局的文本表示进行补充；其次，为了使学习到的表示更适用于聚类任务，设计了聚类模块，该模块将关键词重构损失和 KL 散度损失结合，进而优化指导聚类过程；最后，针对类别重叠问题，设计了对比学习模块，同时将不同实例的样本分开，从而分散了重叠的类别，进一步提高了聚类效果。

本文的主要贡献如下：

1)设计了一种关键语义驱动模块。针对深度学习的文本表征中存在的关键信息提取不充分的问题，使用基于词频的关键词抽取方法获取关键词，对特征学习时的关键语义进行补足。实验结果展示了该模块在聚类任务上的重要性，并通过可视化的方法证明了该算法能有效地捕获关键词或短语。

2)提出了一种在文本聚类任务上兼具通用性和广泛适应性的特征提取方法。针对原始文本质量不高的情况，在增强数据的基础上，在特征提取环节借助预训练模型和自动编码器对文本进行高质量表征，从而更好地兼容各类数据。

3)提出了一种基于关键的语义驱动和对比学习的文本聚类方法。针对传统聚类中存在的关键语义丢失和类重叠问题，强化了文本的关键信息在聚类表征上的作用，并通过对比学习优化了聚类效果。在多个数据集上进行的大量实验表明，本文的方法与之前的方法相比，具有最先进的性能。

1 相关工作

随着技术的不断发展，机器学习和深度学习不断应用于各领域，并被证明具有强大的特征提取和数据处理能力。将深度学习应用到聚类任务中是当前研究的热门方向。相比传统的聚类方法，深度聚类的方法适用于处理高度稀疏的文本数据，且能有效地将聚类的目标融入到神经网络的学习中。

Zhang 等[13]提出 K-means 和自动编码器结合的深度聚类方法。变分深度嵌入式聚类方法[17]使用变分自动编码器和聚类任务结合训练的方式。Xie 等[12]提出 DEC 的聚类方法，DEC 使用浅层的 TF-IDF 作为输入。在 DEC 的基础上，STC[18]用 Word2Vec 训练词向量作为输入，使用聚类策略获取聚类标签。DEC[12]和 STC[18]的研究内容启发了学者对 ARL-Adv[19]的探究，ARL-Adv 将对抗学习引入到聚类中，利用对抗学习增强文本表示。近年来，基于对比学习的方法得到广泛应用，在聚类中应用的是 SCCL[20]。SCCL 在训练中将样本间的对比学习损失引入到聚类中。Bai 等[21]为了弥补表示过程中结构语义的损失，提出结构增强的聚类模型 SEDCN。

文本数据中的关键词语义信息对聚类结果具有关键的指导作用。在实际应用中，存在多种常用的关键词提取方法，例如基于统计特征的算法词频-逆文档频率(TF-IDF)、基于词图模型的算法 TextRank[22]，以及基于主题模型的关键词提取算法 LDA[23]。Wang 等[24]为了捕获更复杂的语义，将文本建模为异构图，利用图神经网络表示句子，从而抽取文本中的关键词和摘要。随着 Transformer[25]和 BERT[26]的出现，Tan 等[27]提出基于 BERT 的摘要生成模型。Transformer 的出现也加速了训练模型的发展，更多生成式预训练模型被提出，如 BART[28]和 T5[29]等，预训练模型在生成摘要和关键词方面也有着较好的表现。

2 本文方法

为了解决在信息抽取过程中存在的关键信息提取不充分、过度依赖原始数据，以及类别在表示空间中相互重叠的问题，本文提出了一种基于关键语义驱动和对比学习的文本聚类方法。如图 1 所示，该方法由 3 个环节组成：在数据处理环节，设计了数据增强模块，以丰富原始文本数据，提高模型泛化性，并使用关键语义驱动模块提取文本中的关键词用于后续关键语义补足；在特征提取环节，针对增强后的文本数据及关键词，结合预训练模型和全连接网络构建的自动编码器进行更精准的语义表示学习，即使用经过原始文本预训练后的权重初始化该模型，这种预训练的初始方法可以保留原始文本中更多的全局信息和细节信息并获取高质量的表示；在聚类学习环节，利用聚类模块和对比学习模块联合优化聚类损失和对比学习损失，实现更好的聚类效果。在此基础上，本文方法可划分为 4 个主要模块：数据增强模块、关键语义驱动模块、对比学习模块和聚类模块。

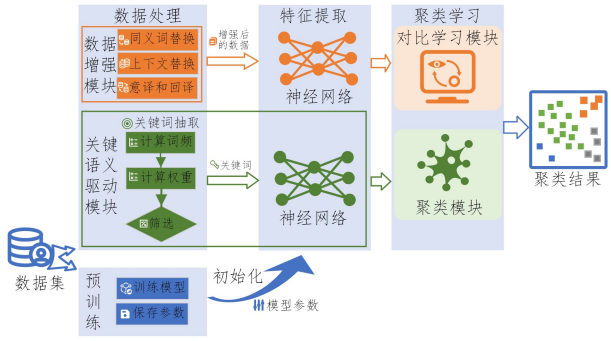


图 1 关键语义驱动和对比学习的文本聚类方法

Fig. 1 Text clustering method based on key semantic-driven and contrastive learning

2.1 预训练模型

预训练模型通过在大规模数据集上进行训练，可以有效地学习数据的通用特征表示，这些特征表示能捕捉到数据底层的统计和语义信息，进而可以进行迁移和重用。

为了获取初步聚类结果和文本的全局语义信息，本文在全连接网络构建的自动编码器上进行了表示学习。该模型由前后对称的编码器和解码器组成。

编码器的作用是将输入序列编码成一个低维稠密的表示(向量)，以便综合和提取整个序列的信息。假设编码器有 L 层，编码器第 l 层的特征可表示为：

$$H(l)=f(\boldsymbol{w}_e \cdot H(l-1)+b_e) \tag{1}$$

其中，f 是激活函数，$\boldsymbol{w}_e$ 和 b_e 是编码器第 l 层的权重矩阵和偏置。

解码器的作用是将编码器生成的隐藏表示逐步重构为原始数据。假设解码器同样有 L 层，解码器第 l 层特征表示为：

$$H(l)=f(\boldsymbol{w}_d \cdot H(l-1)+b_d) \tag{2}$$

其中，f 是激活函数，$\boldsymbol{w}_d$ 和 b_d 是解码器第 l 层的权重矩阵和偏置。

为了令解码器最后一层的输出 $H(L)$ 逼近原始数据，将损失函数表示为：

$$L_x=\sum_{i=1}^{N} \| x_i-x_i' \|_2 \tag{3}$$

其中，x_i 为原始数据，x_i' 为重构数据，N 为样本数量。

2.2 关键语义驱动模块

关键语义驱动是指通过对文本进行深入的语义分析，提取其中关键信息的特征，以此来驱动算法的学习和决策过程。关键语义驱动的方法，可以有效提升算法的准确性和效率，实现更加智能化的数据处理和应用。

1)关键词抽取

关键词通常指在一段文本中具有特定含义或重要性的词语或短语，它们可以帮助总结和概括文本的主题、内容或核心思想。为了解决上述问题，本文受文献[30]和文献[31]的启发，引入了关键语义驱动模块。该模块旨在扩充原始数据的信息量，使最终的文本表示具有更丰富全面的语义信息。

通常而言，关键词在文档中出现的频率远高于其他词，并且从频率上看，关键词与其他词之间的方差较小。此外，相比深度学习模型，基于规则的方法或基于词袋模型的方法能更直接地捕捉到某些关键的词汇或短语[32]。基于以上分析，本文提出了一种基于词频的关键词抽取的方法，从词语的出现频率与词语的方差分布出发，计算各个词语的关键度，如式(4)、式(5)所示。

$$K(c,m)=\frac{TF(w,c)-\frac{1}{M}\sum_{1\leqslant k\leqslant m} TF(w,k)}{var(TF(w))} \tag{4}$$

$$Keywords=\{c \mid K(c,m)>\tau\} \tag{5}$$

其中，w 为整个词库的集合，c 为 w 中的某个词，m 为 w 中不包含 c 的词库集合，$var()$ 为方差，TF 为词频，τ 为关键度阈值。

可以得知，如果词语 c 的频率远高于文档中各词出现的平均频率，且在频率上该词与其他词之间的方差分布更小，则该词语为重要的关键词。当文档中某个词的关键度大于 τ 时，将该词加入关键词集合 $K=\{k_1,k_2,\cdots,k_n\}$。为了确保选取的关键词与原始数据维度一致，使用一个与预训练模块完全相同的自动编码器对筛选出的关键词数据进行特征学习。

2)损失函数

在训练过程中，使用预训练模型保存的参数来初始化关键词自动编码器。这种方式不仅可以利用预训练模型中已经学习到的知识来加速关键词自动编码器的收敛，还能够更好地补足关键语义信息的缺失，从而提高特征表示的质量和语义信息的完整性。为了实现数据重构，令该解码器的最后一层的 $H(L)$ 逼近此数据。关键词重构损失函数表示为：

$$L_{\mathrm{kw}}=\sum_{i=1}^{N} \| k_i-k_i' \|_2 \tag{6}$$

其中，N 为样本数量，k_i 为原始关键词，k_i' 为重构关键词。

本文采用预训练初始化策略主要有以下优势：1)关键词丢失了部分原始数据包含的全局的、细节的语义信息，使用原始数据训练的模型可以更好地捕捉输入数据的共性特征，从而提高模型的泛化能力；2)参数共享策略可以使模型对输入数据中的噪声和干扰具有更好的鲁棒性。

2.3 聚类模块

通过关键语义驱动模块，模型获得了更丰富的特征表示，然而该特征仅仅是基于关键词重构损失优化到的表示，缺少与聚类任务之间的联系，可能会导致该表示不能很好地适用于聚类。表示学习和聚类学习是相互依赖的，需要相互促进。基于此，本文设计了聚类模块，使用自监督聚类机制对编码器进行调整，使得获取到的表示更适合聚类任务。经过编码器得到中间层表示 Z，根据文献[33]，使用学生分布作为计算节点 i 的表示 z_i 和中心 j 的表示 μ_j 的相似度。

$$q_{ij}=\frac{\left(1+\frac{\| z_i-\mu_j \|^2}{\alpha}\right)^{-\frac{\alpha+1}{2}}}{\sum_{j'}\left(1+\frac{\| z_i-\mu_{j'} \|^2}{\alpha}\right)^{-\frac{\alpha+1}{2}}} \tag{7}$$

其中，α 为学生分布的自由度。

根据 q_{ij} 定义优化的目标分布：

$$p_{ij}=\frac{q_{ij}^2/f_j}{\sum_{j'} q_{ij'}^2/f_{j'}} \tag{8}$$

其中，$f_j=\sum_i q_{ij}$，即软聚类频度。式(8)的目标分布首先通过软分配概率 q_{ij} 强调高置信分配的作用，然后通过相关的聚类频率对其进行归一化。通过这种方式，鼓励聚类方法在高置信度的集群分配中学习，同时对抗不平衡集群造成的偏见。

聚类损失函数定义为当前分布和目标分布的 KL 散度，将聚类分配概率向目标分布推进，因此 KL 散度损失函数 L_{kl} 表示为：

$$L_{kl}=KL(P\parallel Q)=\sum_i\sum_j p_{ij}\log\frac{p_{ij}}{q_{ij}} \tag{9}$$

为了在学习过程中既得到适合聚类任务的特征表示，又不破坏原有特征表示的局部结构，本文采用同时优化关键词重构损失和 KL 散度损失的方式。最终的聚类损失函数由关键词重构损失 L_{kw} 与 KL 散度损失 L_{kl} 共同组成。

$$L_c=L_{kw}+L_{kl} \tag{10}$$

2.4 对比学习模块

在聚类过程中，不同类别的数据在表示空间中通常会相互重叠，导致很难得到高纯度的聚类结果，即产生类别重叠问题。为了解决该问题，本文设计了对比学习模块，引入了对比学习中的 InfoNCE 损失，用于拉近同类样本的距离，推远不同类样本的距离，从而促进类别的划分。对比学习能够提高特征表示的区分性，通过拉近正样本间的距离并推远负样本间的距离，使模型学会更具辨识度的特征；同时，它无需依赖大量标注数据，可以利用未标注的数据进行训练，这在标注资源有限的情况下尤为重要；此外，对比学习通过学习数据的不同视角，增强了模型的泛化能力，使其在执行多种下游任务时展现出更加稳健的性能。

本文使用的数据除了原始数据外，还有增强数据对。对原始数据中的每条数据生成成对的增强数据对。假设原始数据表示为 X，大小为 M，则增强数据对 X^a 的大小则为 $2M$。假设 $x_{i1},x_{i2}\in X^a$（x_{i1},x_{i2} 为同一条数据产生的数据增强对）为正样本，则剩余的 $2M-2$ 条均为负样本。z_{i1},z_{i2} 为 x_{i1},x_{i2} 经过特征提取环节后的表示。对于 x_{i1}，通过最小化式(11)将 x_{i2} 与其他所有负样本区分开。

$$L_{cl}=-\log\frac{\exp\left(\frac{sim(z_{i1},z_{i2})}{t}\right)}{\sum_{j=1}^{2M}\Gamma_{j\neq i}\cdot\exp\left(\frac{sim(z_{i1},z_j)}{t}\right)} \tag{11}$$

其中，$\Gamma_{j\neq i}$ 为指示函数；t 为温度参数；$sim(\cdot)$ 为一对归一化输出之间的点积，即 $sim(z_i,z_j)=z_i^{\mathrm{T}}z_j/z_{i2}z_{j2}$。

本文探索了 3 种不同的无监督文本增强方法。

1) WordNet Augmenter[34]：通过用 WordNet 同义词替换输入文本的单词来转换输入文本。

2) Contextual Augmenter[35]：利用预训练的 Transformers在输入文本中找到 top-n 个合适的单词进行插入或替换。通过单词替换对数据进行扩充，并选择 BERT 生成扩充对。

3) Paraphrase via back translation[36]：首先将输入文本翻译成另一种语言，然后翻译成原始语言，生成对输入文本的意译。

本文最终的整体损失如式(12)所示，即聚类损失与对比学习损失相结合。

$$L=L_c+\lambda L_{cl} \tag{12}$$

其中，L_c 和 L_{cl} 在式(10)和式(11)中定义，λ 为对比学习损失和聚类损失之间的平衡因子。

2.5 改进的文本聚类算法

综上所述，基于关键语义驱动和对比学习的文本聚类方法的算法主要包括 3 个阶段：数据处理、特征提取以及聚类学习。具体的伪代码如算法 1 所示。

算法 1 基于关键语义驱动和对比学习的文本聚类方法

输入：原始数据集 text；预训练迭代次数 N；对比学习的关键语义驱动文本聚类方法迭代次数 N1；聚类方法 Model；关键词抽取算法 KSD；数据增强算法 Aug；评价函数 Metric；真实标签 True_label

输出：关键词集合 K；聚类标签 Pre_labe

```
//数据处理阶段
K=KSD(text) //提取关键词
text1,text2=Aug(text) //生成增强数据
//特征提取阶段
for i=0 to N do:
    使用 text 和式(1)－式(3)训练模型
    保存模型参数 W
end for
Model=init(Model,W) //初始化 Model
//聚类学习阶段
for i=0 to N1 do:
    用 K-means 算法计算初始化聚类中心
    feat1,feat2=Model(text1,text2) //获取增强数据表示
    cl_loss=Model.contrast_loss(feat1,feat2) /*计算对比学习损失*/
    feat=Model(K) //获取关键词表示
    cluster_loss=Model.cluster_loss(feat) //计算聚类损失
    loss=cluster_loss+λcl_loss //联合优化
    loss.backward() //反向传播
end for
Pre_label=Model.get_cluster_prob(feat)
return K,Pre_label
Metric(True_label,Pre_label)
```

3 实验

3.1 实验数据与参数设置

采用深度学习框架 PyTorch 对相关模型进行编码，并在 Ubuntu18.04 上采用 GPU(NVIDIA GeForce RTX 3060 * 2)对模型进行调试和训练。采用 Hyperopt 进行参数的选取，优化器为 Adam，batch size 设置为 16，学习率 $lr=1\times10^{-5}$。

为了验证本文方法的有效性，参照相关研究选取了 8 个不同的文本数据集进行实验，每个数据集的详细信息如表 1 所列。

20-news[37]：收集了大约 20000 条的新闻档，将其均匀分为 20 个不同主题的新闻组集合。

StackOverflow[38]：Kaggle 发布的一个数据集，该数据集包含 20 个不同类别相关的 20000 个问题标题。

AgNews[39]：新闻标题的数据集，包含 4 个主题。

Tweet[40]：包含 2472 条推文，89 个类别。

GoogleNews[40]：包含与 152 个事件相关的 11109 篇新闻文章的标题和摘要。完整的数据集被命名为 GoogleNews-TS，分别通过提取标题和摘要得到 GoogleNews-T 和 GoogleNews-S[41]。

工业数据集：数据来源于某航空制造企业，收集了 2018－2022 年期间某分厂的故障文本数据大约 30000 条。按照故障发生的阶段、部件及发生的位置进行分类，直接定位到故障发生的具体位置及故障内容，共 76 类。该数据集中故障文本

由人工录入，数据中存在大量的噪声（错别字、专业符号、专业代码），且数据多为短文本。

表 1 数据集详细情况

Table 1 Details of datasets

数据集	词汇量	类别数	最大类别与最小类别的大小之比
工业数据集	7000	76	104
20-news	12000	20	1
StackOverflow	15000	20	1
AgNews	21000	4	1
Tweet	5000	89	249
GoogleNews-TS	20000	152	143
GoogleNews-T	8000	152	143
GoogleNews-S	18000	152	143

经过分析与相关的实验探索，将方法中涉及的参数最终设置如下。

1）τ 的设置

在关键语义驱动模块中，关键词选取的阈值是非常重要的参数。为了寻找在关键语义驱动模块中的最佳阈值，同时证明所提模型具有良好的鲁棒性，针对数据集使用不同阈值探索关键词阈值选取对聚类结果的影响，如表 2 所列。可以看出，在一定区间内，模型结果随着阈值的增大而提高，关键词质量高有助于提升模型的性能；当阈值上升到一定程度时，聚类结果不再有明显提升，反而趋于下降，这是由于关键词阈值过高导致关键词数量减少，模型学习到的知识有限，进而使聚类结果有较大幅度的下降。因此，本文在所有实验中设置 $\tau=0.7$。

表 2 不同 τ 对模型结果的影响

Table 2 Effect of different τ on model results

（%）

数据集	评价指标	$\tau=0.5$	$\tau=0.6$	$\tau=0.7$	$\tau=0.8$	$\tau=0.9$
工业数据集	*ACC*	73.46	82.08	83.47	80.32	73.46
	NMI	61.21	62.78	63.45	61.36	60.83
20-news	*ACC*	21.11	25.17	26.71	24.37	21.11
	NMI	28.95	34.81	35.78	32.26	28.95

2）α 的设置

由于无法在无监督的环境中对验证集上的 α 进行交叉验证，因此依据文献[28]，在所有实验中设置 $\alpha=1$。

3）λ 的设置

λ 为对比损失和聚类损失之间的平衡因子。为了探究 λ 的最佳取值，在 StackOverflow 上进行了实验，如图 2 所示。实验表明，当 λ 小于 1 时，模型的 ACC 效果较差；当 λ 为 5 时，模型的效果最好。因此，本文在所有实验中设置 $\lambda=5$。

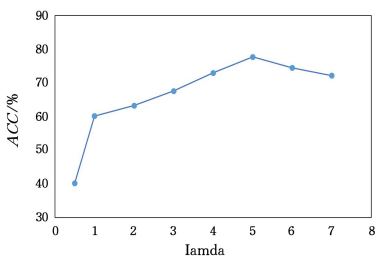


图 2 λ 对方法性能的影响

Fig. 2 Impact of λ on the performance of the method

3.2 对比方法

为了证明本文提出的方法可以在文本聚类上实现最先进的或极具竞争力的性能，将其与 6 种聚类方法进行了对比，包括经典的聚类方法、当前较为先进的基于语义的聚类方法以及深度聚类方法。

1）K-means[5]：使用 TF-IDF 在维度为 1 500 的文本特征上应用 K-means 来评估。

2）DEC[12]：深度嵌入式聚类方法，是经典的深度聚类方法。

3）SEDCN[21]：解决结构语义信息在特征学习过程中消失的问题的深度聚类方法。

4）STC[18]：一种增强的预训练词嵌入，采用了分层预训练获得的自编码器的深度聚类方法。

5）HAC-SD[41]：在低于所选阈值的相似性得分归零获得的稀疏成对相似性矩阵上应用分层聚类。

6）SCCL[20]：在 DEC 的基础上引入样本级别的对比损失函数。

3.3 评价指标

使用聚类精度（ACC）和正则互信息量（NMI）这两个常用的聚类指标来评估聚类性能。

$$ACC(y,c)=\max_{m}\frac{\sum_{i=1}^{n}1\{y_i=g(c_i)\}}{n} \tag{13}$$

其中，y_i 和 c_i 分别为样本 x_i 的真实标签和使用聚类算法预测的标签，$g(\cdot)$ 为预测标签到真实标签的一对一映射。

$$NMI(y,c)=\frac{I(y,c)}{\frac{1}{2}[H(y)+H(c)]} \tag{14}$$

其中，$I(,)$ 代表互信息，$H(\cdot)$ 代表熵，y，c 含义同式（13）。

ACC 和 NMI 两个值的取值范围均为[0,1]，结果越接近 1，说明聚类效果越好。

3.4 实验结果分析

为了证明本文模型对噪声输入具有鲁棒性，在实验时对原始数据未进行除关键词提取以及数据增强以外的去噪声预处理。表 3 列出了 $\tau=0.7$ 时，使用 Paraphrase via back translation 数据增强方法在 8 个数据集上的聚类结果。从评估指标上看，本文方法明显优于大多数方法。基于深度学习的聚类方法往往优于基于传统聚类的方法，例如 DEC 的结果优于 K-means，这说明将原始数据降维到低维特征空间的深度聚类算法较传统聚类方法能够学到更好的表示。SCCL 和本文方法在多数指标上优于没有使用对比学习的方法，这说明对比学习的方法可以提高聚类的效果。本文通过将聚类损失和对比损失相结合，不仅提高了簇内的内聚性，使数据更接近聚类中心，而且能够拉开不同类别间的距离。

本文方法在 20-news，StackOverflow，AgNews 以及工业数据集这 4 个数据集上的效果均优于对比方法，这表明关键语义的补足和对比学习的方法能够促进聚类结果的改善。在 20-news 数据集上，本文方法效果较差，原因可能是该数据集相较于其他数据集不同类别的类别重叠问题更为复杂，导致聚类效果不佳，但本文方法相较于对比的方法已经有明显的性能提升。在 GoogleNews 数据集上，本文方法在个别指标

上略低于其他方法。经过分析，可能的原因是：1）GoogleNews 和 Tweet 数据集相比于其他数据集有更多类别，且最大类别与最小类别在样本量上的大小之比相差较大，因此利用关键语义驱动模块提取关键词时，较少类别中关键词的数量通常远少于非关键词，导致关键词被非关键词淹没，从而造成较少样本类别数据的关键词较难提取；2）GoogleNews 和 Tweet 相比其他数据集样本数量更少，对于对比学习来说，通常需要更多的训练数据才有更好的效果。

表 3 ACC 和 NMI 对比结果

Table 3 Comparison results of ACC and NMI

（%）

数据集	评价指标	K-means	DEC	SEDCN	STC	HAC-SD	SCCL	Ours
20-news	*ACC*	10.23	23.34	24.73	25.36	24.98	26.57	**26.71**
	NMI	5.62	25.38	32.29	32.67	35.12	33.54	**35.78**
StackOverflow	*ACC*	58.41	63.58	70.21	59.80	64.80	75.50	**77.69**
	NMI	58.78	63.45	55.24	54.80	59.50	74.50	**76.92**
AgNews	*ACC*	34.56	76.31	79.56	78.63	81.80	88.20	**90.45**
	NMI	11.94	60.56	68.91	59.87	54.60	68.20	**71.97**
Tweet	*ACC*	57.05	73.59	86.47	78.24	**89.60**	78.20	87.53
	NMI	80.77	85.91	84.59	82.14	85.20	89.20	**90.32**
GoogleNews-TS	*ACC*	68.08	78.59	85.69	83.45	85.80	89.80	**91.87**
	NMI	88.96	91.25	89.96	92.25	88.00	**94.92**	93.94
GoogleNews-T	*ACC*	58.94	70.45	73.71	73.46	**81.80**	75.80	79.24
	NMI	79.35	84.12	83.45	84.61	84.20	88.30	**91.65**
GoogleNews-S	*ACC*	61.97	69.84	79.49	80.79	80.60	83.10	**85.93**
	NMI	83.21	85.54	84.59	85.74	83.50	**90.40**	89.22
工业数据集	*ACC*	57.89	75.42	80.71	81.29	81.65	82.34	**83.47**
	NMI	43.12	48.31	58.29	59.21	60.81	58.28	**63.45**

3.5 消融实验

为了更好地验证本文方法各模块的性能，分别进行了消融实验，从多个方面证明本文所提框架具有良好的性能。

1）联合优化聚类损失和对比学习损失的模型性能更优

本文将聚类模块（Clustering Module）命名为 cluster，对比学习模块（Comparative Learning Module）命名为 CL，在 StackOverflow 和 Tweet 数据集上进行了实验，分别减少了 cluster 和 CL 模块，实验结果如图 3 所示（图中“－”代表减少）。从图 3 中可以看出，联合优化聚类损失和对比学习损失的模型性能在 StackOverflow 和 Tweet 数据集上的准确率（ACC）和标准化互信息（NMI）均优于单独优化单一的损失。减少聚类模块后 ACC 和 NMI 都较低的原因可能是，聚类模块使用关键语义补足了原始语义的缺失，且该模块将关键词重构损失与 KL 散度损失结合，使学习到的表示更适用于聚类任务。这一结果验证了本文所提的联合优化框架在利用聚类和对比学习的优势互补方面具有有效性和重要性。

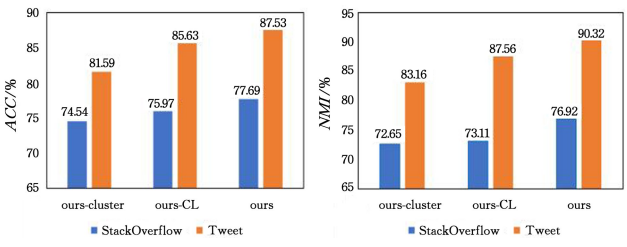
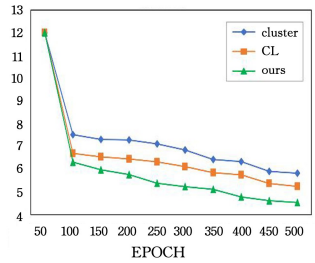


图 3 对比学习模块和聚类模块对模型性能的影响

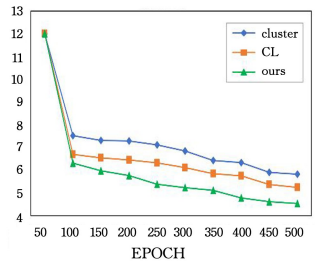
Fig. 3 Impact of comparing learning module and clustering module on model performance

2）联合优化聚类损失和对比学习损失的模型簇内和簇间距离更优

在模型学习过程中评估文本在表示空间中的簇内距离和簇间距离。对于一个给定的簇，簇内距离是簇中心与该簇中所有样本之间的平均距离。簇间距离，即最近的相邻簇之间的距离。本文采用欧氏距离计算簇内距离和簇间距离。在 StackOverflow 数据集上进行 500 次迭代实验，结果如图 4 所示（图中 cluster 代表去除对比学习模块，保留聚类模块，CL 则相反）。可以看出，联合优化聚类损失和对比学习损失在 StackOverflow 数据集上获得了更小的簇内距离和更大的簇间距离，这一目标与聚类分析的目标是一致的。



（a）簇内距离



（b）簇间距离

图 4 迭代过程中簇内距离和簇间距离的变化

Fig. 4 Variation of intra-cluster distance and inter-cluster distance during iterations

为了更直观地对比，本文参考文献[42]中的可视化方法，使用 Matplotlib 将聚类结果可视化，结果如图 5 所示。可以看出，联合优化聚类损失和对比学习损失在 StackOverflow

数据集上获得了更小的簇内距离和更大的簇间距离，聚类效果更优。

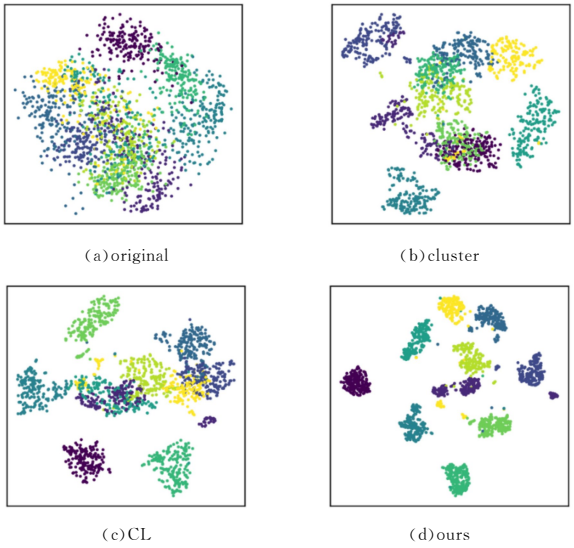


图 5 聚类结果的可视化

Fig. 5 Visualization of clustering results

3)关键语义驱动模块对聚类有较强的指导作用

本文将关键语义驱动模块(Key Semantic Driver Modules)命名为 KSD。为了研究使用 KSD 对模型性能的影响，对本文所提方法进行消融实验，结果如表 4 所列(表中"—"代表减少)。可以看出，本文提出的关键语义驱动模块在工业数据集、20-news 和 StackOverflow ACC 上都有明显的性能提升，表明关键语义驱动模块提取出的关键词对聚类有较强的指导作用。

表 4 有无语义驱动模块的模型性能

Table 4 Model performance with and without semantics-driven modules

(%)

数据集	Ours	Ours-KSD
20-news	26.71	23.29
StackOverflow	77.69	75.93
工业数据集	83.47	78.59

为了研究关键语义驱动模块的关注点，使用式(4)在 StackOverflow 上对每个单词的计算权重进行了可视化，可视化结果如图 6 所示(颜色越深代表权重越大)。

文本	类别
does spring provide 2 mvc platforms, grails and spring mvc? how do i inject a single property value into a string using spring 2.5.x ?	spring
what would i have to do to get the new html editor ajax control to work as a sharepoint content editor webpart?	sharepoint
what is the best way of generating a xslx file on a web site? possibly with millions of rows?	excel
synchronize data between frontend and backend	ajax
in haskell, how do i recursively manipulate a tuple and preappend a character to the first element in the tuple?	haskell
how do i respond to an internal drag-and-drop operation using a qlistwidget? how to decode "application/x-qabstractitemmodeldatalist" in qt for drag and drop?	qt

图 6 关键语义驱动模块权重的可视化

Fig. 6 Visualization of key semantics-driven module weights

可视化结果表明，该算法能有效地捕获关键词或短语，如"excel"类别中的"xslx file"，"ajax"类别中的"frontend"和"backend"等，这些 n-gram 是特定类别的关键词或短语，对聚类任务至关重要。

4)可适应多种特征提取器

为了探索本文方法是否能适应更多的特征提取器，选择 CNN，LSTM 和 BERT 来替换特征提取环节中的 AutoEncoder 进行实验(损失函数进行了相应的修改)。实验结果如表 5 所列，可以看出，本文提出的聚类框架在不同的特征提取器上均取得了良好的性能(ACC)，说明本文所提的聚类框架并不局限于单一的特征提取器，而是可以适应多种特征提取器的使用，具有一定的通用性和适应性，且能够与不同的特征提取方法相结合，更好地适应各种数据类型。

表 5 不同特征提取器的 ACC 对比

Table 5 Comparison of ACC of different feature extractors

(%)

数据集	CNN	LSTM	BERT	AutoEncoder
20-news	25.98	26.44	26.53	26.71
StackOverflow	77.34	76.94	76.93	77.69
AgNews	89.85	90.05	89.82	90.45
Tweet	85.52	86.03	86.62	87.53
GoogleNews-TS	90.96	91.87	91.77	91.87
GoogleNews-T	78.6	79.02	78.25	79.24
GoogleNews-S	85.53	85.32	85.75	85.93
工业数据集	83.07	82.66	82.85	83.47

5)可适应不同的数据增强方法

为了探索数据增强对模型性能的影响，本文对比了 WordNet Augmenter，Contextual Augmenter 和 Paraphrase via back translation 数据增强方法在不同数据集上的性能，结果如表 6 所列。其中 Paraphrase via back translation 技术明显优于其他两种技术，原因可能是 Paraphrase via back translation 采用来回翻译的方式，不会很大程度上改变句子的语义，而 WordNet Augmenter，Contextual Augmenter 均是对单词替换，这可能会使句子的语义发生变化甚至出现偏离原始语义的情况。

表 6 不同数据增强方法的对比

Table 6 Comparison of different data enhancement methods

(%)

数据集	*ACC*			*NMI*		
	Wnet	Ctxt	Para	Wnet	Ctxt	Para
工业数据集	81.69	80.71	**83.47**	62.02	60.37	**63.45**
20-news	25.47	24.63	**26.71**	33.54	31.24	**35.78**
StackOverflow	72.64	74.84	**77.69**	74.61	70.32	**76.92**
AgNews	86.65	88.54	**90.45**	62.11	58.94	**71.97**
Tweet	85.42	86.54	**87.53**	83.14	87.21	**90.32**
GoogleNews-TS	84.18	83.94	**91.87**	87.59	88.68	**93.94**
GoogleNews-T	75.96	77.19	**79.24**	90.21	88.73	**91.65**
GoogleNews-S	82.18	83.44	**85.93**	84.56	87.94	**89.22**

4 讨论

4.1 方法的有效性

本文方法主要从两个方面入手，采用关键语义驱动模块对文本表示的信息进行补足，引入对比学习以更好地促进类别的划分。接下来分析两个模块对方法有效性的提升。

1)关键语义驱动

基于词频和词频方差分布的方式提取的关键词,能排除停用词的影响,有效提取出文本中关键度高的词语,该关键词可以补足在信息抽取过程中丢失的关键语义信息,进而丰富文本的表示。此外,本文使用聚类模块同时学习重构损失和聚类损失(式(11)),由于重构损失直接关联到输入数据的重建,因此它通常更容易解释模型的行为和性能,有助于模型学习更泛化的特征表示。同时,它关注于数据的整体结构而不是序列的逐元素预测,且该部分将类簇信息特征引入了文本表示,使得最终的特征表示更符合聚类结果。本文 3.5 节的实验可以表明关键语义对聚类具有较强的指导作用。

2)对比学习引入

对比学习的具体思想是将样例与语义相似的例子(正样例)和语义不相似的例子(负样例)进行对比,使语义相近的例子对应的表示在表示空间更接近,语义不相近的例子对应的表示距离更远,这一思想与聚类思想一致。本文使用对比学习模块来更好地促进类别划分,3.5 节中的多个实验可以表明,联合优化聚类损失和对比学习损失模型可以获得更好的簇内和簇间距离,进而使聚类结果更优。

4.2 方法的耗时分析

本文方法的时间消耗主要来源于以下几个部分:

1)预训练:所提方法须在数据集上进行预训练,以供后续初始化模型使用。该阶段通常会消耗大量的时间。

2)数据处理:在数据处理过程中需要对原始数据进行关键词抽取和数据增强。该阶段时间消耗为正常时间开销。

3)损失函数:本文使用的损失函数联合优化了 3 部分的损失,相较于比较的方法更为复杂,但该部分的计算均为线性计算,不存在较复杂的数学运算,因此该阶段相比优化其他损失函数的时间消耗增加较少。

综上所述,本文方法在总体时间的消耗上比其他方法长,但从实际的工程应用角度看,许多领域已有预训练模型,因此预训练时间可忽略。在本文方法基础上去除预训练的时间后,在时间消耗上比对比方法更有优势,因此本文方法具有可行性。

结束语 通过对文本聚类问题的研究,本文提出了一种基于关键语义驱动和对比学习的文本聚类方法。在 8 个基准文本聚类数据集上对该方法进行了全面的评估,结果显示该方法表现最优。此外,本文还通过相关实验得出以下结论:

1)本文提出的关键语义驱动模块能够有效地捕获文本中的关键词或短语,在聚类任务上实现精准的指导。同时,联合优化聚类损失和对比学习损失可以更好地优化簇内和簇间距离,提高聚类表现。

2)本文提出的聚类方法不局限于单一的特征提取器,具有一定的通用性和适应性,可与多样化的特征提取方法相结合,更好地适应各种数据类型。

本文提出的方法在文本聚类问题上具有较好的性能,并且具备一定的通用性和适应性,为相关领域的研究和应用提供了有价值的参考。

未来工作可以从以下两个方面着手:1)本文的文本表示方法仅从语义的角度抽取得到,缺少词句之间的结构信息,图神经网络可以有效地捕捉数据中的结构特性,在后续工作中考虑将图神经网络引入聚类中,利用图神经网络提取文本数据中词句之间的结构信息,进一步增强文本的语义表示;2)本文提出的关键词抽取算法对于类别相差较大的数据还存在一些不足,在后续工作中考虑采用生成式的方式抽取文本中的关键词,对文本语义进行补足。

参 考 文 献

[1] SAEEDI EMADI H, MAZINANI S M. A novel anomaly detection algorithm using DBSCAN and SVM in wireless sensor networks[J]. Wireless Personal Communications, 2018, 98: 2025-2035.

[2] WIBISONO S, ANWAR M T, SUPRIYANTO A, et al. Multivariate weather anomaly detection using DBSCAN clustering algorithm[C]// Journal of Physics: Conference Series. IOP Publishing, 2021.

[3] LIU F, XUE S, WU J, et al. Deep learning for community detection: progress, challenges and opportunities[J]. arXiv: 2005.08225, 2020.

[4] MENG Y, ZHANG Y, HUANG J, et al. Hierarchical topic mining via joint spherical tree and text embedding[C]// Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2020: 1908-1917.

[5] MACQUEEN J. Some methods for classification and analysis of multivariate observations[C]// Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability. 1967: 281-297.

[6] CELEUX G, GOVAERT G. Gaussian parsimonious clustering models[J]. Pattern Recognition, 1995, 28(5): 781-793.

[7] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]// Proceedings of the Second International Conference on Knowledge Discovery and Data Mining(KDD'96). 1996: 226-231.

[8] DING C, HE X, SIMON H D. On the equivalence of nonnegative matrix factorization and spectral clustering[C]// Proceedings of the 2005 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2005: 606-610.

[9] NG A, JORDAN M, WEISS Y. On spectral clustering: Analysis and an algorithm[C]// NIPS. 2002.

[10] JIANG B, YE L Y, PAN W F, et al. Service Clustering Based on the Functional Semantics of Requirements. [J]. Chinese Journal of Computers, 2018, 41(6): 1035-1046.

[11] QIAO S J, HAN N, JIN C Q, et al. A Distributed Text Clustering Model Based on Multi-Agent[J]. Chinese Journal of Computers, 2018, 41(8): 1709-1721.

[12] XIE J, GIRSHICK R, FARHADI A. Unsupervised deep embedding for clustering analysis[C]// International Conference on Machine Learning. PMLR, 2016: 478-487.

[13] ZHANG D, SUN Y, ERIKSSON B, et al. Deep unsupervised clustering using mixture of autoencoders[J]. arXiv: 1712.07788, 2017.

[14] SHAHAM U, STANTON K, LI H, et al. Spectralnet: Spectral

clustering using deep neural networks[J]. arXiv:1801.01587, 2018.

[15] ZHOU S, XU H, ZHENG Z, et al. A comprehensive survey on deep clustering: Taxonomy, challenges, and future directions [J]. arXiv:2206.07579, 2022.

[16] CAI X Y, HUANG J J, BIAN Y C, et al. Isotropy in the Contextual Embedding Space: Clusters and Manifolds[C]// International Conference on Learning Representations. 2021.

[17] JIANG Z, ZHENG Y, TAN H, et al. Variational deep embedding: A generative approach to clustering [J]. arXiv: 1611.05145, 2016.

[18] HADIFAR A, STERCKX L, DEMEESTER T, et al. A self-training approach for short text clustering[C]// Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019). 2019: 194-199.

[19] ZHANG W, DONG C, YIN J, et al. Attentive representation learning with adversarial training for short text clustering[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 34(11): 5196-5210.

[20] ZHANG D, NAN F, WEI X, et al. Supporting clustering with contrastive learning[J]. arXiv: 2103.12953, 2021.

[21] BAI R N, HUANGR Z, ZHENG L Y, et al. Structure enhanced deep clustering network via a weighted neighbourhood auto-encoder[J]. Neural Networks, 2022(155): 144-154.

[22] MIHALCEA R, TARAU P. Textrank: Bringing order into text [C]// Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. 2004: 404-411.

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

[24] WANG D, LIU P, ZHENG Y, et al. Heterogeneous graph neural networks for extractive document summarization [J]. arXiv: 2004.12393, 2020.

[25] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Procedings of the 31st International Conference on Neural Information Processing Systems. 2017: 6000-6010.

[26] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv: 1810.04805, 2018.

[27] TAN J Y, DIAO Y F, QI R H, et al. Automatic summary generation of Chinese news text based on BERT-PGN mode[J]. Journal of Computer Applications, 2021, 41(1): 127-132.

[28] LEWIS M, LIU Y, GOYAL N, et al. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension[J]. arXiv: 1910.13461, 2019.

[29] RAFFEL C, SHAZEER N, ROBERTS A, et al. Exploring the limits of transfer learning with a unified text-to-text transformer [J]. The Journal of Machine Learning Research, 2020, 21(1): 5485-5551.

[30] YU W, LU N, QI X, et al. PICK: processing key information extraction from documents using improved graph learning-convolutional networks[C]// 2020 25th International Conference on Pattern Recognition (ICPR). IEEE, 2021: 4363-4370.

[31] YI Z L, ZHANG H L, NA R L, el al. Deep text clustering algorithm based on key Semantic Information [J]. Application Research of Computers, 2023, 40(6): 1653-1659.

[32] ROSE S, ENGEL D, CRAMER N, et al. Automatic Keyword Extraction from Individual Documents[J]. text Mining: Application and Theory, 2010, 4: 1-20.

[33] MAATEN L V D, HINTON G. Visualizing data using t-SNE [J]. Journal of Machine Learning Research, 2008, 9(86): 2579-2605.

[34] REB S, DENG Y, HE K, et al. Generating natural language adversarial examples through probability weighted word saliency [C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 1085-1097.

[35] KOBAYASHI S. Contextual augmentation: Data augmentation by words with paradigmatic relations[J]. arXiv: 1805.06201, 2018.

[36] SHEN T, OTT M, AULI M, et al. Mixture models for diverse machine translation: Tricks of the trade[C]// International Conference on Machine Learning. PMLR, 2019: 5719-5728.

[37] LUO C J, ZHAN J F, WANG L, et al. Cosine normalization: Using cosine similarity instead of dot product in neural networks [C]// Artificial Neural Networks and Machine Learning-ICANN 2018. Springer International Publishing, 2018: 382-391.

[38] XU J, XU B, WANG P, et al. Self-taught convolutional neural networks for short text clustering[J]. Neural Networks, 2017, 88: 22-31.

[39] ZHANG X, LECUN Y. Text understanding from scratch[J]. arXiv: 1502.01710, 2015.

[40] YIN J, WANG J. A model-based approach for text clustering with outlier detection[C]// 2016 IEEE 32nd International Conference on Data Engineering (ICDE). IEEE, 2016: 625-636.

[41] RASHADUL H R M, ZEH N, JANKOWSKA M, et al. Enhancement of Short Text Clustering by Iterative Classification [J]. arXiv: 2001.11631, 2020.

[42] LI H. Statistical learning methods (VersionII) [M]. Beijing: Tsinghua University Press, 2019.

ZHANG Shiju, born in 1997, postgraduate, is a member of CCF (No. I8208G). His main research interest is natural language processing.

YANG Fengyu, born in 1980, associate professor, is a member of CCF (No. 37982S). His main research interest is analysis and mining of physical quality data for aviation products.

（责任编辑：何杨）