计算机科学
COMPUTER SCIENCE

融合无监督SimCSE的短文本聚类研究

贺文灏, 吴春江, 周世杰, 何朝鑫

引用本文

贺文灏, 吴春江, 周世杰, 何朝鑫. 融合无监督SimCSE的短文本聚类研究[J]. 计算机科学, 2023, 50(11): 71-76.

HE Wenhao, WU Chunjiang, ZHOU Shijie, HE Chaoxin. Study on Short Text Clustering with Unsupervised SimCSE [J]. Computer Science, 2023, 50(11): 71-76.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220900214

融合无监督 SimCSE 的短文本聚类研究

贺文灏　吴春江　周世杰　何朝鑫
电子科技大学信息与软件学院　成都 610054
(202022090510@std.uestc.edu.cn)

摘　要　传统的浅层文本聚类方法在对短文本聚类时,面临上下文信息有限、用词不规范、实际意义词少等挑战,导致文本的嵌入表示稀疏、关键特征难以提取等问题。针对以上问题,文中提出一种融合简单数据增强方法的深度聚类模型 SSKU(SBERT SimCSE *K*-means Umap)。该模型采用 SBERT 对短文本进行嵌入表示,利用无监督 SimCSE 方法联合深度聚类 *K*-Means 算法对文本嵌入模型进行微调,改善短文本的嵌入表示使其适于聚类。使用 Umap 流形降维方法学习嵌入局部的流形结构来改善短文本特征稀疏问题,优化嵌入结果。最后使用 *K*-Means 算法对降维后嵌入进行聚类,得到聚类结果。在 StackOverFlow,Biomedical 等 4 个公开短文本数据集进行大量实验并与最新的深度聚类算法作对比,结果表明所提模型在准确度与标准互信息两个评价指标上均表现出良好的聚类性能。

关键词:短文本;深度聚类;预训练模型;降维方法;自然语言处理
中图法分类号　TP391

Study on Short Text Clustering with Unsupervised SimCSE

HE Wenhao,WU Chunjiang,ZHOU Shijie and HE Chaoxin
School of Information and Software Engineering,University of Electronic Science and Technology of China,Chengdu 610054,China

Abstract　Traditional shallow text clustering methods face challenges such as limited context information,irregular use of words,and few words with actual meaning when clustering short texts,resulting in sparse embedding representations of the text and difficulty in extracting key features. To address these issues,a deep clustering model SSKU(SBERT SimCSE Kmeans Umap) incorporating simple data augmentation methods is proposed in the paper. The model uses SBERT to embed short texts and fine-tunes the text embedding model using the unsupervised SimCSE method in conjunction with the deep clustering KMeans algorithm to improve the embedding representation of short texts to make them suitable for clustering. To improve the sparse features of short text and optimize the embedding results,Umap manifold dimension reduction method is used to learn the local manifold structure. Using *K*-Means algorithm to cluster the dimensionality-reduced embeddings,and the clustering results are obtained. Extensive experiments are carried out on four publicly available short text datasets,such as StackOverFlow and Biomedical, and compared with the latest deep clustering algorithms. The results show that the proposed model exhibits good clustering performance in terms of both accuracy and standard mutual information evaluation metrics.

Keywords　Short text,Deep clustering,Pre-training model,Dimension reduction,Natural language processing

1　引言

随着互联网信息技术在各行业领域不断应用,人们的日常活动愈加离不开互联网。例如网络社交平台、新闻报道等,每天都将产生大量的短文本数据,因此,对海量无标签数据进行智能分析与处理的需求日益增加。而文本聚类作为一种无监督的机器学习方法,已成为对文本信息进行有效地组织、摘要的重要手段,被广泛应用于自然语言处理的各类场景。

传统的文本聚类方法是先对文本进行特征提取再使用聚类模型对特征进行聚类。短文本存在上下文信息有限、不规范用语和缩略语多、实际意义词少等特点,其关键特征难以提取。传统的词嵌入方法对短文本进行向量化表示易出现特征稀疏、特征维度高、相似度量方法无效等问题。

使用当前最流行的 BERT(Bidirectional Encoder Representation from Transformer)[1] 预训练模型所获取的句嵌入向量,虽然能够表征整个句子的语义,但所得到的向量存在非光滑各向异性[2](Non-smooth Anisotropy)的问题。该问题会导致模型学习到的词嵌入向量在向量空间中占据一个狭窄的圆锥体形[2],即没有均匀地分布在向量空间上,彼此之间计算余弦相似度都很高。通过相似度量方法得到的结果,并不

到稿日期:2022-09-23　返修日期:2023-02-27
通信作者:周世杰(sjzhou@uestc.edu.cn)

能代表句子间真实的语义相似性。

传统聚类方法将特征提取与聚类分开处理，容易导致聚类模型和特征不匹配，结合上述词嵌入的各向异性问题，将对聚类任务带来灾难性的影响。

针对以上问题，本文提出一种 SSKU（SBERT SimCSE Kmeans Umap）深度聚类模型。本文的主要贡献如下：

1）利用预训练模型获取短文本向量嵌入表示。使用数据增强方法对模型进行微调，改善文本的嵌入表示，增强模型的泛化能力，减弱光滑各向异性所带来的相似度无法度量问题。

2）使用 KMeans 深度聚类模型联合数据增强方法对预训练模型与深度聚类模型进行微调，进一步改善短文本的嵌入表示，并使其适于聚类。使用流式降维方法学习嵌入的局部流形结构，对训练后的文本表示进行重嵌入。

3）在 4 个公开数据集上进行了广泛实验并与最先进的深度聚类方法进行对比，来验证本文模型的有效性。此外，根据数据集的特点，探究了本文的数据增强方法的局限性，并进行消融实验验证了模型中各算法的必要性。

2 相关研究

2.1 文本嵌入方法

传统的文本聚类方法分为两步：1）提取文本特征，获取文本的嵌入表示；2）使用聚类方法对 1）中所获嵌入进行聚类。文本嵌入的质量将直接影响最终的聚类效果。

传统的向量嵌入方法是基于统计实现的，例如词袋模型（Bag of Words，BOW）和词频-逆文档频率（Term Frequency Inverse Document Frequency，TF-IDF）。由于短文本中的上下文信息有限，基于统计实现的嵌入表示往往非常稀疏，导致依赖于词重叠或高维向量间距离的传统相似性度量方法失效。许多研究者利用外部知识对短文本进行扩充，例如 Hu 等[3]、Banerjee 等[4]利用维基百科的相应数据对文本进行扩充，或者是将多个短文本合并为长的伪文本来增强文本表达能力。但这些方法需要额外的数据来源，各个领域也需要对应领域的知识，导致成本过高。

相比基于统计的向量嵌入方法，利用深度学习训练获得的文本嵌入的表示能力更强，具备不同程度的语义表示的能力。Word2Vec[5]是以词的上下文推断词含义的方式训练得到的词向量模型。该模型将每个词映射到一个稠密的向量，以获得对应的词嵌入表示。Word2Vec 训练得到的嵌入表示与单词是一对一关系，然而单词在不同上下文中具有不一样的含义，因此多义词的问题无法解决。

随着深度学习的发展，通过考虑上下文而选择不同语义的嵌入方法层出不穷。其中最出色的方法是使用以 Transformer 为基础架构的自监督学习方法来训练海量语料中词与词间语义关系的预训练模型。BERT 预训练模型是 Google 以无监督的方式利用大量无标注文本训练的语言模型，该模型所提取的特征向量比传统向量嵌入方法的语义信息更丰富，在多个 NLP 任务测试中有着优异表现。由于词频的差异，BERT 输出的词向量在向量空间中呈现锥形分布，高频的词都靠近原点（锥顶），低频词远离原点，因而词嵌入间距离不能很好地表达词间的语义相关性，在相似度任务上表现较差。

SBERT[6]预训练模型使用孪生（Siamese）和三级（Triplet）BERT 网络结构来获得句子嵌入，使用余弦相似度或欧氏、曼哈顿距离等进行比较，找到语义相似的句子。相比使用 BERT 对短文本进行编码，采用 SBERT 预训练模型编码在文本相似度等任务上的表现更好。

2.2 深度聚类方法

聚类是无监督学习领域最基本的挑战之一，几十年来一直受到广泛的关注。长期存在的聚类方法如 K-Means[7]以及高斯混合模型[8]依赖于在数据空间中的测量距离，这对于高维的数据来说往往是无效的。近年来，许多研究集中于通过优化表示空间中定义的聚类目标，将聚类与深度表示学习相结合，通过深度神经网络强大的表示能力来提升聚类算法性能，聚类结果反过来引导神经网络学习更好的特征。Xie 等[9]提出了基于 KL 散度的深度聚类算法，将 KL 散度损失用于微调一个预训练好的编码器，微调后的聚类结果较之前得到大幅度提升，他们将其称为深度嵌入聚类（Deep Embedding Clustering，DEC）。该方法是最早出现和最流行的深度聚类算法之一。Hadifar 等[10]实现了从数据空间到低维特征空间的映射，其研究表明低维的连续表示嵌入可以解决短文本向量稀疏性问题。

Zhang 等[11]提出了一种基于对比学习的短文本聚类方法（SCCL），将实例对比学习与深度聚类相结合，联合优化自顶向下的聚类损失和自底向上的实例对比损失，在多个数据集上取得了优异的结果。其验证了对比学习在深度聚类应用的可行性。Wang 等[12]将多视图聚类与深度学习相结合，提出了一种基于 NMF 的广义深度学习多视图聚类方法 GDLMC，在不同类别的数据集上均有不错表现。Pugachev 等[13]的研究表明，由 Transformer 得到句向量结合不同的聚类方法可以成功应用于深度聚类任务。McConville 等[14]的研究表明，自动编码器嵌入的局部流形学习对发现更高质量的簇类是有效的，使用 Umap[15]（Uniform Manifold Approximation and Projection）方法捕捉数据局部的流形结构表现最佳。

3 SSKU 模型

如图 1 所示，SSKU 模型框架由文本嵌入、自监督训练、联合微调、重嵌入和聚类 5 部分组成，具体流程如下：1）文本嵌入，采用 SBERT 初始化短文本表示；2）自监督训练，利用无监督数据增强方法 SimCSE 对文本嵌入表示进行改善；3）联合微调，联合 K-Means 聚类算法与 SimCSE 方法对 SBERT 进行微调，增强嵌入的表征能力；4）Umap 重嵌入，使用 Umap 流式降维算法对文本进行重嵌入，得到文本的低维嵌入来对抗编码时向量稀疏的问题，提升聚类精度；5）聚类，使用 K-Means 算法对文本的最终嵌入进行聚类。

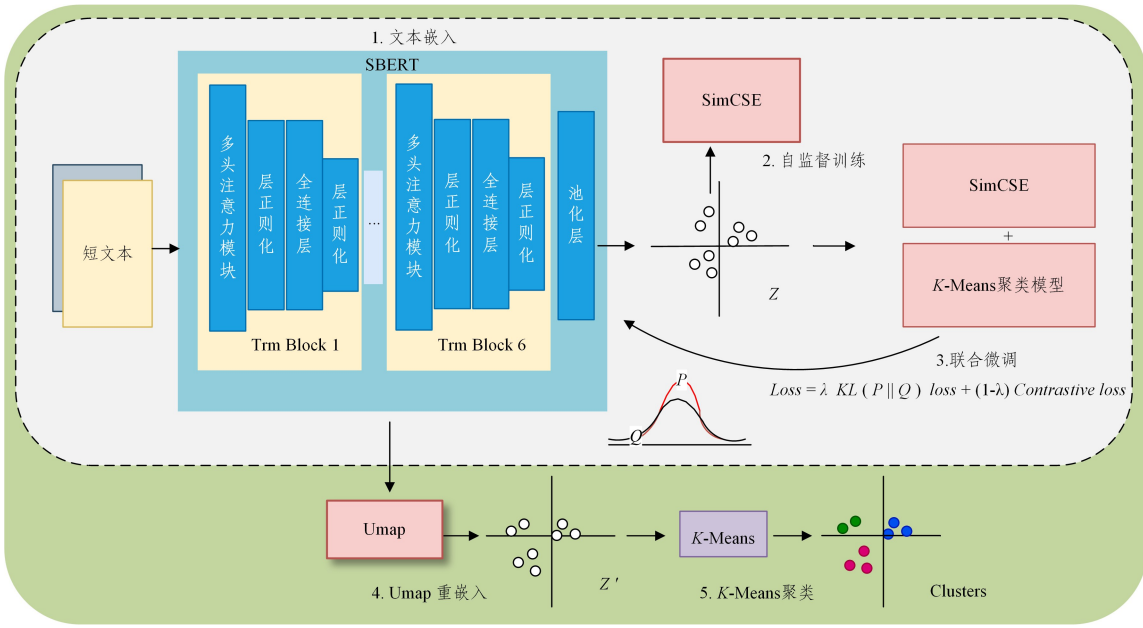


图 1 SSKU 模型

Fig. 1 SSKU model

3.1 短文本嵌入模型

本文采用 SBERT 作为短文本嵌入方法。该网络成功地结合了孪生网络在相似度度量任务上的优势与 BERT 在语义表达上的优势，使训练得到的句子嵌入能更好地用于相似度度量等任务。SBERT 模型有 3 种训练方式可对 BERT 嵌入结果进行微调，此处仅介绍回归任务。给定两个文本 A 和 B，首先将文本分别输入共享权重的孪生 BERT 进行特征提取与编码，经池化层转化为特征向量 $\boldsymbol{u}$ 和 $\boldsymbol{v}$，通过 Cosine 计算 $\boldsymbol{u}$ 和 $\boldsymbol{v}$ 的相似度，来得到文本 A 与文本 B 的相似性；使用 MSE(平方均误差)作为损失函数。其计算公式如式(1)所示：

$$MSE=\frac{1}{n}\sum_{i=1}^{n}(y_i-y_i')^2 \tag{1}$$

其中，y_i 与 y_i' 分别表示第 i 个样本的真实值与预测值，n 为样本个数。预训练获取的语义相近的句子嵌入向量距离较近，可用来进行相似度计算。

3.2 无监督 SimCSE 数据增强方法

BERT 类模型得到的词嵌入表示在空间分布不均匀，呈锥形分布，高频词都靠近原点(所有均值)，而低频词远离原点。高频词与低频词处于向量空间中的不同区域，则高频词与低频词间的空间距离失去了意义，不再适用。直接使用 K-Means 对 SBERT 句子嵌入聚类将导致模型的时间复杂度过高，聚类准确率低，并且后续的深度聚类模型效果非常依赖聚类簇心的初始化质量。

针对上述问题，本文采用 SimCSE 作为模型编码前的数据增强方法，以自监督方式对模型进行微调，改善句子嵌入表示。该方法利用对比学习的思想，通过大量正负样本进行对比的训练方式，使得相似样本在该空间的距离接近，而不相似样本间的距离较远。将相同的句子输入预训练模型 Transformer 编码两次，由于每次输入模型 dropout 的神经元不同，因此获得的两个相应的向量化表示也不同。而向量在语义上又是相似的，故将其构建为正样本数据。负样本是对同一个 batch 中的其他数据进行随机采样构成的。其训练损失函数公式如式(2)所示：

$$L_i=-\log\frac{e^x}{\sum_{j=1}^{N}e^x} \tag{2}$$
$$x=\mathrm{sim}(h_i,h_j^+)/\tau$$

其中，τ 为温度系数，作用是调节模型对与正样本更相似的负样本(困难样本)的关注程度，本文设置与文献[2]相同，$\mathrm{sim}(h_i,h_i^+)$是样本间的余弦相似度。

3.3 *K*-Means 深度聚类模型

为防止深度聚类方法出现退化解[16]，联合 SimCSE 与 K-Means 深度聚类模型进一步优化句子嵌入。在首次使用 K-Means 算法初始化簇心后，我们交替进行 3 个步骤：1)计算每个样本点属于各类簇的软分配；2)计算辅助概率分布并将其用作训练 SBERT 模型的目标；3)联合 SimCSE 与 K-Means 聚类模型迭代更新预训练网络权重与先前得到的聚类簇心嵌入。

K-Means 深度聚类模型的具体流程如下：

1)使用 K-Means 算法对经自监督训练后的句子嵌入进行聚类，初始化簇心。由于 K-Means 对初始化类簇中心非常敏感，因此，选择正确的簇心对聚类效果的好坏有很大的影响。为了减少该初始化影响，使用不同的初始簇心进行 100 次 K-Means 计算，将计算所得划分误差最小的簇心作为初始的聚类簇心。

2)对每个样本点与所划分的簇进行概率估计。使用学生 t-分布计算每个样本点 i 属于簇 j 的概率，计算公式如式(3)所示：

$$q_{ij}=\frac{(1+\|\boldsymbol{z}_i-\boldsymbol{\mu}_j\|^2)^{-1}}{\sum_{j'}((1+\|\boldsymbol{z}_i-\boldsymbol{\mu}_{j'}\|^2)^{-1})} \tag{3}$$

其中，$\boldsymbol{z}_i$ 为样本点的嵌入表示，$\boldsymbol{\mu}_j$ 表示簇心向量。本文将学生 t-分布的自由度设为 1[9]。

3)根据 Q 分布定义一个辅助目标分布 P[9]，来提高聚类簇的纯度以及置信度，防止大簇扭曲隐藏的特征空间。辅助

目标分布 P 由式(4) 计算。

$$p_{ij}=\frac{q_{ij}^2/\sum_{i'}q_{i'j}}{\sum_{j'}(q_{ij'}^2/\sum_{i'}q_{i'j'})} \tag{4}$$

其中，q_{ij} 表示样本 i 属于簇心 j 的概率。

4)在辅助目标分布的帮助下，通过学习数据的高置信分配来迭代地改进聚类。模型通过软分配去匹配辅助目标分布来训练，因此，本文的损失函数定义为软分配 $\boldsymbol{q}_{ij}$ 和辅助分布 $\boldsymbol{p}_{ij}$ 之间的 KL 散度，如式(5)所示：

$$L_C=KL(P\parallel Q)=\sum_i\sum_j \boldsymbol{p}_{ij}\log\frac{\boldsymbol{p}_{ij}}{\boldsymbol{q}_{ij}} \tag{5}$$

SSKU 模型由无监督 SimCSE 与聚类模型联合训练，由式(2)可得，该模型损失函数公式如式(6)所示：

$$L=\boldsymbol{\lambda}L_C+(1-\boldsymbol{\lambda})L_i \tag{6}$$

其中，L_C 为聚类损失；L_i 为数据增强损失；$\boldsymbol{\lambda}$ 为控制嵌入空间畸变的系数，类似于多任务学习，用于平衡聚类任务与数据增强任务对模型学习的影响。因数据增强模型可改善数据的原有分布，使其均匀地分布在嵌入空间中，故在此基础上使用较小的聚类损失去改变嵌入空间不会导致其损坏。因此本文将 $\boldsymbol{\lambda}$ 设为 0.1 。

3.4 Umap 重嵌入方法

Umap 是一种基于流形学习技术和拓扑数据分析思想的降维算法。其依赖于 3 个假设，即数据均匀分布在黎曼流形上、黎曼度量是局部常数、流形是局部连接的。根据这些假设，可以用模糊拓扑结构对流形进行建模，通过搜索具有最接近的等效模糊拓扑结构的数据的低维投影来发现嵌入。相较于其他流式降维算法(如 Isomap[17])，Umap 在寻求准确地表示局部结构的同时保留了更多的全局结构，并且具有更好的运行性能。本文使用 Umap 对训练后的句嵌入表示进行重嵌入，以此来发现更适合聚类的嵌入表示。

4 实验结果与分析

4.1 实验数据集

SearchSnippets[18]：从 web 搜索词条中提取。其中包含 8 个类别共 12340 条短文本。

StackOverflow[19]：Kaggle 发布的数据的子集。Xu 等选择了其中 20 个不同类别相关的 20000 个问题题目。

Biomedical[19]：BioASQ3 发布的 PubMed 数据的子集。其中随机选择了 20 组共 20000 篇论文标题，所选论文标题最大长度为 53。

AgNews[20]：AG 新闻标题语料库的子数据集。包含 AG 语料库中 4 个最大的类别，分别是世界、体育、商业、科技。

数据集详细介绍如表 1 所列，其中 C 为各数据集的类别数，T 为各数据包含的短文本数量，L 表示各数据集的平均长度。

表 1 数据集详情

Table 1 Dataset details

Dataset	C	T	L
SearchSnippets	8	12340	18
StackOverflow	20	20000	8
AgNews	4	8000	23
Biomedical	20	20000	13

4.2 实验环境

本文实验使用 Ubuntu 18.04.5 LTS 操作系统，PyTorch 深度学习开发框架，用 Python 作为开发语言。实验采用的 CPU 为 Intel 酷睿 i7-7700K，GPU 为 NVIDIAGeForce RTX2080Ti。

4.3 实验步骤

在本文实验中，通过 SBERT 预训练模型的池化层获取一个 384 维向量作为短文本的嵌入表示。本文使用无监督 SimCSE 对 SBERT 模型进行自监督训练，将句子嵌入转换到一个各向同性且分布较均匀的空间，优化待聚类的嵌入。经迭代训练后，重新输入文本得到优化后的向量，再输入到聚类网络，利用聚类结果反向优化预训练模型与聚类模型参数。为防止聚类模型出现退化解，使用无监督 SimCSE 损失函数，联合 KL 散度作为该阶段的总体损失函数，通过不断优化模型输出的样本估计 Q 和辅助目标分布 P 间的距离，来提高聚类准确度。对联合训练后所得到的嵌入使用 Umap 进行重嵌入。最后使用 K-Means 算法对最终嵌入结果进行聚类。

在实验过程中选用 SGD 作为优化器，动量 (Momentum) 设置为 0.99。batchsize 设置为 64，初始学习率设为 0.003，自监督训练 epoch 为 80，联合训练 epoch 为 30。

4.4 评价指标

本文采用聚类精确度(Accuracy，ACC)与标准互信息(Normalized Mutual Information，NMI)作为模型评价指标。聚类准确率 ACC 计算公式如下：

$$ACC=\frac{\sum_{i=1}^{n}\delta(S_i,map(r_i))}{N} \tag{7}$$

其中，r_i 为聚类后的标签；S_i 为真实标签；N 为数据总的个数；δ 表示指示函数；map 表示最佳类标的重现分配，即将模型得到的粗标映射为数据真实标签。其具体函数如式(8)所示：

$$\delta(a,b)=\begin{cases}1, & \text{if } a=b\\ 0, & \text{otherwise}\end{cases} \tag{8}$$

NMI 常用于聚类度量两个聚类结果的相近程度。NMI 计算公式如下：

$$NMI(U,V)=\frac{MI(U,V)}{F(H(U),H(V))} \tag{9}$$

其中，MI 为互信息(Mutual Information)，$H(\cdot)$ 为信息熵，$F(\cdot)$ 可以使用几何平均，如式(10)所示，亦可以使用算术平均，如式(11)所示。

$$F(x_1,x_2)=\sqrt{x_1x_2} \tag{10}$$

$$F(x_1,x_2)=\frac{x_1+x_2}{2} \tag{11}$$

为了对比实验的准确性和可信度，在实验中我们选取算术平均而非几何平均，与文献[9-10]中的对比方法保持一致，NMI 计算式可表示为：

$$NMI(U,V)=2\frac{MI(U,V)}{H(U)+H(V)} \tag{12}$$

4.5 对比方法

为证明本文方法在短文本聚类上表现出良好的性能，选取以下方法作为对比方法。

BoW 与 TF-IDF：在 1500 维相关特征上应用 K-Means 算法对 BoW 和 TF-IDF 方法进行评价。

STCC[19]：对每个数据集使用 Word2Vec 方法训练短文本嵌入。优化一个卷积神经网络，以进一步丰富输入 *K*-Means 进行最后阶段聚类的表示。

Self-Train[10]：使用 SIF[21]增强预训练词嵌入。逐层预训练自编码器，再通过文献[9]中的训练方法，对自编码器进行微调。

HAC-SD[20]：在稀疏成对相似性矩阵上应用层次聚类，该相似性矩阵通过将低于选定阈值的相似性得分归零而获得。

SCCL[11]：将实例对比学习与 Xie 等[9]提出的深度聚类相结合，联合优化自顶向下的聚类损失与自底向上的实例对比损失，以实现更好的类间距离和类内距离。

4.6 实验结果分析

本实验使用聚类精确度与标准互信息作为评价指标。取 5 次实验的平均值作为最终结果，在相同的数据集上做了 5 组对比实验，表 2 列出了实验结果。

表 2 对比实验的准确率与互信息

Table 2 Accuracy and mutual information of comparative experiments

methods	StackOverflow		AgNews		SearchSnippets		Biomedical	
	ACC	*NMI*	*ACC*	*NMI*	*ACC*	*NMI*	*ACC*	*NMI*
BoW	18.5	14.0	27.6	2.6	24.3	9.3	14.3	9.2
TF-IDF	58.4	58.7	34.5	11.9	31.5	19.2	28.3	23.2
STCC	51.1	49.0	—	—	77.0	63.2	43.6	38.1
Self-Train	59.8	54.8	—	—	77.1	56.7	54.8	47.1
HAC-SD	64.8	59.5	81.9	54.6	82.7	63.8	40.1	33.5
SCCL	75.5	74.5	**88.2**	**68.2**	**85.2**	**71.1**	46.2	41.5
SSKU	**85.6**	**76.5**	84.9	64.1	81.2	69.6	**55.4**	**49.3**

从表 2 中可以得出：

1）SSKU 在 4 个短文本数据集中均表现出良好的聚类性能。该方法在 StackOverFlow 与 Biomedical 数据集上在聚类精度指标上比 SCCL 方法分别高出了 10.1%，9.2%，在标准互信息指标上分别高出 2%，7.8%。

2）将 BOW，TF-IDF，SBERT，SimCSE_SBERT 这 4 种不同的嵌入表示方法与 *K*-Means 结合，验证文本嵌入的质量对文本聚类的重要性。从表中可以看到，4 个实验中，在短文本上直接利用基于词频的语言模型获得的文本嵌入聚类效果最差，经数据增强后的文本嵌入效果最好。从实验结果来看，由大量语料训练得到的高质量文本嵌入可以提高下游聚类任务的效果。

4.7 消融实验

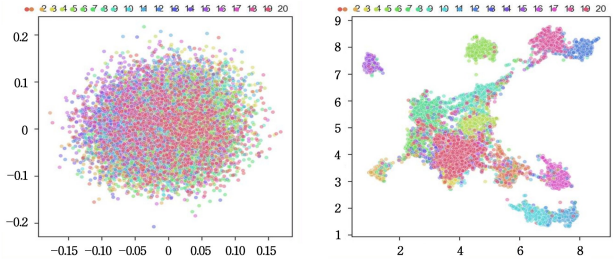
为了更好地验证本文提出的聚类框架的有效性，我们对是否经过 SimCSE 与 Umap 改善的句子嵌入进行了对比实验。从消融实验中，我们可以得到如下信息：

1）从表 3 可知，SSKU 模型的聚类准确率与互信息两个评价指标相比其他方法组合均有提升。通过该框架优化得到的文本嵌入更适合聚类任务。这说明了本文框架对聚类嵌入优化的可行性，以及低维度嵌入能有效提高短文本的表示能力。为了更直观地体现 SSKU 方法的有效性，对 SBERT，Umap 和 SSKU 这 3 种嵌入方式在 StackOverFlow 数据集上使用 *K*-Means 的聚类结果进行可视化。从图 2 与图 3 中可以看出，经 SSKU 模型得到的嵌入表示簇与簇间的划分与边界更加明显，聚类效果更好。

表 3 消融实验(1)的结果

Table 3 Results of ablation experiment(1)

methods	StackOverflow		SearchSnippets	
	ACC	*NMI*	*ACC*	*NMI*
SBert	72.0	70.0	74.5	59.5
SBert-Umap	76.7	71.6	78.3	68.8
Sbert-SimCse	75.2	71.4.	73.4	59.5
SSK	80.8	74.4	74.6	59.5
SSKU	**85.6**	**76.5**	**81.2**	**69.6**



(a)SBERT (b)SBERT_Umap

图 2 SBERT 与 SBERT_Umap 在 StackOverFlow 数据集上的可视化结果

Fig. 2 Visualization of SBERT and SBERT_Umap on StackOverFlow dataset

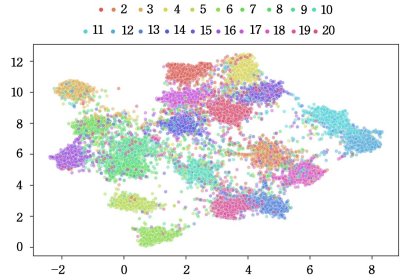


图 3 SSKU 方法在 StackOverFlow 数据集上的可视化结果

Fig. 3 Visualisation results of SSKU method on StackOverFlow dataset

2）在 SearchSnippets 数据集上对经 SimCSE 得到的嵌入进行聚类比直接对 SBERT 获取的句子嵌入聚类结果低 1.1%。由于 SimCSE 将同一句子输入模型两次来构建对比学习的正样本，因此正样本对间包含长度相似的信息，导致模型对相似长度样本进行错误判断。这一点在文献[22]中得到了印证。

本文为探究经 SimCSE 训练得到的句子嵌入负优化结果进行以下实验。在 Search_Snippets 每一个簇中分别抽取句子长度为 17～19（平均长度±1）（Search_Snippets(Avg)）和长度各不相同（Search_Snippets(Dif)）的数据各 1 500 条。首先使用 SimCSE 对句子嵌入进行优化，再使用 *K*-Means 算法得到聚类结果。实验结果如表 4 所列。从表中可以看出，对于长度相似的数据进行数据增强得到负优化的结果，相比直接聚类的方法的 ACC 低 5.5%，而在不同长度的数据集上提高了 2.4%。相比整个数据集，使用 SSKU 方法在长度不同的数据集上表现更优。Search_Snippets 与 AgNews 在不同的簇类中平均长度较为相似，导致本文方法在该类数据集中提升不大，在数据增强方法上，SimCSE 仍有改进空间。

表 4 消融实验(2)的结果

Table 4 Results of ablation experiment(2)

methods	StackOverflow		SearchSnippets	
	ACC	*NMI*	*ACC*	*NMI*
SBert	74.8	60.3	66.7	60.4
Sbert-SimCse	69.3	55.6	69.1	61.1
SSKU	78.5	67.3	83.2	70.6

结束语 针对传统浅层聚类方法的聚类、降维效果差等问题,本文提出了一种融合数据增强的深度聚类模型。首先利用 SBERT 提取短文本的向量表示,然后使用无监督 SimCSE 作为短文本数据增强方法,改善短文本嵌入表示。再联合深度聚类模型对 SBERT 进行微调,使预训练模型提取的嵌入表示更符合聚类任务。最后使用 Umap 流式降维方法学习优化的嵌入表示局部流式结构,从而抵抗短文本稀疏向量的问题。本文模型在 StackOverflow 数据集上的聚类准确率 ACC 和标准互信息 NMI 远高于当前最先进的方法。

未来我们将探究对模型进行微调的方法来优化计算成本较高的问题。此外,在实际应用中 K-Means 聚类算法的 k 值是未知的,而选择合适的 k 值对算法的影响非常大,而常用的确定 k 的值方法对于本文来说开销过大,我们今后也将在此方向努力。

参 考 文 献

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

[2] GAO T, YAO X, CHEN D. Simcse: Simple contrastive learning of sentence embeddings[J]. arXiv: 2104.08821, 2021.

[3] HU X, ZHANG X, LU C, et al. Exploiting wikipedia as external knowledge for document clustering[C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery Anddata Mining. 2009: 389-396.

[4] BANERJEE S, RAMANATHAN K, GUPTA A. Clustering short texts using Wikipedia[C]// Proceedings of the 30th Annual International ACM SIGIR Conference on Research and development in Information Retrieval. 2007: 787-788.

[5] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient Estimation of Word Representations in Vector Space[J]. arXiv: 1301.3781, 2013.

[6] REIMERS N, GUREVYCH I. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing(EMNLP-IJCNLP). 2019: 3982-3992.

[7] MACQUEEN J. Classification and analysis of multivariate observations[C]// 5th Berkeley Symposium on Mathematical Statistics and Probability. 1967: 281-297.

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

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

[10] 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.

[11] ZHANG D, NAN F, WEI X, et al. Supporting Clustering with Contrastive Learning[C]// NAACL-HLT. 2021.

[12] WANG D, LI T, DENG P, et al. A Generalized Deep Learning Algorithm based on NMF for Multi-view Clustering[J]. IEEE Transactions on Big Data, 2022.

[13] PUGACHEV L, BURTSEV M. Short text clustering with transformers[J]. arXiv: 2102.00541, 2021.

[14] MCCONVILLE R, SANTOS-RODRIGUEZ R, PIECHOCKI R J, et al. N2d: (not too) deep clustering via clustering the local manifold of an autoencoded embedding[C]// 2020 25th International Conference on Pattern Recognition(ICPR). IEEE, 2021: 5145-5152.

[15] MCINNES L, HEALY J, MELVILLE J. Umap: Uniform manifold approximation and projection for dimension reduction[J]. arXiv: 1802.03426, 2018.

[16] GUO X F. A Study on Image Clustering Algorithms with Deep Neural Networks[D]. Changsha: National University of Defense Technology, 2020.

[17] TENENBAUM J B, SILVA V, LANGFORD J C. A global geometric framework for nonlinear dimensionality reduction[J]. Science, 2000, 290(5500): 2319-2323.

[18] PHAN X H, NGUYEN L M, HORIGUCHI S. Learning to classify short and sparse text & web with hidden topics from large-scale data collections[C]// Proceedings of the 17th International Conference on World Wide Web. 2008: 91-100.

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

[20] RAKIB M R H, ZEH N, JANKOWSKA M, et al. Enhancement of short text clustering by iterative classification[C]// International Conference on Applications of Natural Language to Information Systems. Cham: Springer, 2020: 105-117.

[21] ARORA S, LIANG Y, MA T. A simple but tough-to-beat baseline for sentence embeddings[C]// International Conference on Learning Representations. 2017.

[22] WU X, GAO C, ZANG L, et al. Esimcse: Enhanced sample building method for contrastive learning of unsupervised sentence embedding[J]. arXiv: 2109.04380, 2021.

HE Wenhao, born in 1997, postgraduate. His main research interests include natural language processing and machine learning.

ZHOU Shijie, born in 1970, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. His main research interests include artificial intelligence and network security.

(责任编辑:何杨)