计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200300087

基于稀疏表示的多文档自动摘要

钱玲龙　武　娇　王人锋　陆慧娟
中国计量大学　杭州 310018
(linglong.qian@kcl.uk)

摘　要　文档自动摘要是自然语言处理领域中的重要任务,受限于难以准确理解文档语义,大多通过词频、关键词等人工特征对文档句子进行重要程度排序,以此提取摘要。受稀疏表示理论启发,提出了一种基于稀疏表示的动态语义空间划分算法。算法对初始划分的语义子空间进行字典学习,利用所得字典对所有句向量进行稀疏重构,从而将各句向量动态调整至重构误差最小的划分,迭代地实现语义空间的重划分。对于划分后语义子空间内摘要句的提取,提出了一种基于稀疏相似度排序的自动摘要提取算法。将各语义子空间的所有句向量作为字典原子,通过稀疏重构,得到能体现句子对其他句子语义表征程度的稀疏相似度,以各句累积稀疏相似度作为衡量句子表征空间语义信息能力的指标,依据其排序来提取摘要句。在猫途鹰网站热门景点旅游评论数据集上进行了实验,结果表明语义空间重构误差快速迭代5次即可稳定收敛且平均有效降低重构误差约17%,且算法对数据维度不敏感,所提摘要避免了重复提取冗余度大、重复性高的文本,是一种有效的自动摘要方法。

关键词:自动摘要;字典学习;稀疏重构
中图法分类号　TP391.1

Multi-document Automatic Summarization Based on Sparse Representation

QIAN Ling-long, WU Jiao, WANG Ren-feng and LU Hui-juan
China Jiliang University, Hangzhou 310018, China

Abstract　Automatic document summary is an important task in the field of natural language processing. Limited by the difficulty of accurately understanding the semantics of documents, most of the documents are sorted by artificial features, such as word frequency and keywords, to extract the abstract. Inspired by the theory of sparse representation, a dynamic semantic space partition algorithm based on sparse representation is proposed. The algorithm performs dictionary learning on the initially divided semantic subspace, uses the obtained dictionary to sparsely reconstruct the sentence vector. Dynamically adjusts it to the division which has the smallest reconstruction error. Iteratively realizes the re-division of the semantic space. For abstracting sentences in the divided semantic subspace, an automatic extraction algorithm based on sparse similarity ranking is proposed. All sentence vectors in each semantic subspace are viewed as dictionary atoms. Through sparse reconstruction, the sparse similarity can be obtained which reflects the degree of semantic representation of one sentences to others. The cumulative sparse similarity of each sentence to other sentences is used as a metric to measure the ability of the sentence to represent the spatial semantic information. Ranking the cumulative sparse similarity, and then extract the required top N sentences. The experimental results on the travel review data set of popular attractions on the TripAdvisor website show that the semantic space reconstruction error can be rapidly reduced after 5 iterations, remain stable which shows the convergence. Except for effectively reduce the reconstruction error by nearly 17%, the algorithm is also not sensitive to data dimensions. The proposed summary avoids repeated abstraction of redundant and highly repetitive text, which is an effective multi-document automatic summarization method.

Keywords　Automatic summarization, Dictionary learning, Sparse reconstruction

1　引言

随着PC和移动设备的普及,互联网技术和移动通信技术的发展,人类社会进入大数据时代。随之出现的现象就是信息过载,或称为信息爆炸。据统计,世界上最大的搜索引擎公司Google所存储索引的网页数量超过了1万亿[1)],每天处理近100PB的数据。据数字宇宙的研究报告,未来8年里人类将产生超过40ZB(=40万亿GB)的数据量[1]。如何对互联网的海量数据进行有效的浏览并获取信息是自然语言处理领域的研究热点。

本文已加入开放科学计划(OSID),请扫描上方二维码获取补充信息。

基金项目:国家自然科学基金(61272315,61602431);浙江省自然科学基金(LQ20F030015);国家级大学生创新创业训练计划-基于自然语言处理的智能阅读模型(201810356020)
This work was supported by the National Natural Science Foundation of China (61272315, 61602431), Natural Science Foundation of Zhejiang province, China (LQ20F030015) and National College Student Innovation and Entrepreneurship Training Program (201810356020).

通信作者:陆慧娟(hjlu@cjlu.edu.cn)

1) http://infolab.stanford.edu/~backrub/google.html;中国互联网网络信息中心:数字宇宙研究报告

互联网用户的增加以及电子政务的快速发展，使得大量不同主题的电子文本在线可用，这些文本之间往往存在大量冗余信息。对于企业级用户来说，各类线上平台的评论数据是宝贵的财富，他们需要通过从大量评论数据中快速挖掘有效信息来调整产品结构、优化产品内容。自动文本摘要(Automatic document summarization，ADS)是快速获取信息的一种有效途径。

本文着重研究基于语义空间划分的多文档摘要(Multi-document summarization，MDS)提取方法。通过对语义空间进行划分，使得语义相近的句向量单独成簇。从语义子空间中抽取摘要句，形成能够最大程度表征文本语义信息的摘要。

本文第3节介绍多文档摘要提取技术的基本原理和步骤、文本表示模型，以及稀疏表示理论的相关理论；第4节提出基于语义空间动态划分的MDS方法，并详细介绍提出的基于稀疏表示的动态语义空间划分(Dynamic Semantic Spatial Partitioning based on Sparse Representation，SR-DSSP)算法和基于稀疏相似度排序(Sparse Similarity Ranking，SS-Rank)算法的数学模型构建及求解过程；第5节通过实验对本文所提算法的有效性进行验证；最后总结全文。

2 相关研究

自动摘要生成过程主要分为生成式摘要和抽取式摘要两种，前者通过对文档的深入理解，对原文作完整的浓缩，不仅要求有篇章理解能力，还要求生成的语句逻辑通顺，能包含原文信息，因此难度较大；后者则更偏重对文档内重要内容的识别，从而抽取出关键语句[2]。大多数文本摘要系统使用基于统计和代数的方法，提取文本关键词、高频词等人工特征，对句子标注重要程度，以此为依据提取摘要[3]。目前通用的摘要提取方法大致分为两种：无监督方法和监督方法。无监督方法通常使用排序模型从候选集中选择句子，包括基于语言模型的方法等；监督方法则将文档摘要视为分类问题或序列标记问题。然而，这些方法大多存在问题，即排名靠前的句子或标记的候选摘要语句通常共享很多冗余信息，并且不能同时区分显著性，大多数还单独评估句子并逐句选择句子。

Liu等[4]通过应用稀疏编码技术解决了从多文档集中提取摘要语句的问题，并为这一具有挑战性的问题提出了一个新的框架。基于数据重构和句子去噪假设，提出了一个两级稀疏表示模型来描述多文档摘要的过程。实验表明，提出的模型是有效的，并且优于当前最先进的算法。He等[5]从压缩感知的角度研究这个任务，将句子视为一种稀疏的，可压缩的信号，提出了一种新的群体稀疏学习多文档摘要框架，通过最小化逼近误差，最大限度地重构原始文档，并在句子间共同选择学习后的群组结构信息。通过近似约束的重构模型来对摘要相关性建模，使多个句子有共同的基础结构，从而形成摘要。He等[6]提出了一种基于数据重构的文档摘要框架，使得可以选择出能够最大程度重建整个文档的最具代表性的句子作为文档摘要。在上述工作的基础上，He等[7]还提出了一种基于数据重建的无监督文档摘要框架，其选择能够最好地重建整个文档的最具代表性的句子。通过线性重构和贪婪优化方法，提出非负重构和相应的迭代方法来获得全局最优。由于文档集句子之间存在较大的冗余，若将句子视为信号，虽然可以考虑通过提取能够最好地重构原始文档中所有句子的句子子集来实现摘要的提取，即将ADS问题转换为稀疏重构问题，还需要对选择将非结构化的文本数据结构化的表示方法、划分语义空间的方法，以及在语义空间中提取摘要所使用的稀疏重构的算法进行研究。

以上方法通常使用词向量空间模型来表示文本，并利用聚类算法提取关键数据。对于词向量这种高维数据，传统方法很难获得好的聚类效果，使之在多文档自动摘要任务上略显疲态。文献[8]中大量的认知试验证明了高维数据存在较低的可表示其数据特征的维数，大多分布于高维空间的一个低维子流形上。受上述启发，本文将从子空间划分的角度入手研究多文档自动摘要任务。

3 文档结构化表示模型

3.1 词向量空间模型

非结构化数据如文本等不能直接被计算机所识别处理。在自然语言处理任务中，向量空间模型(Vector Space Model，VSM)以文档的字、词、句等基本语言单位为项，通过统计模型、语言模型、神经网络等方法，将文本数据结构化。其中以词为基本单位的表示方法称为词向量表示。词向量通常有两种形式：One-hot表示(One-hot representation)和分布式表示(Distribution Representation)[9]。

One-hot表示又称热独编码，在中文文本中，经过分词、停词、去噪等预处理，构建文本词典(又称词袋模型)。当语料库较大时，生成的词典中将含有数目相当庞大的词，基于One-hot表示的词的向量空间维度相当高且极其稀疏。向量的One-hot表示忽略了词序信息，且认为词与词之间是互相独立的，缺失了语义信息。

分布式表示被用来弥补上述表示的不足，其主要思想是基于某种规则将语料中的每一个词的One-hot表示全部映射为一个固定长度的向量，将所有向量从原始的高维稀疏空间映射到相对低维稠密的向量空间。该过程被称为词嵌入(Word embedding)，这一过程起到了维数约简的作用。在分布式表示下，词向量的维度可控，并且词之间所蕴含的语义可以由空间中点的距离等分布信息进行挖掘。在不同领域中，研究者提出了很多训练分布式词向量的方法，如早期机器学习中的Word2vec模型、深度学习中的ELMO[10]模型、Bert[11]模型等。其中，Google在2013年开源了一个获取基于Word2vec模型的词向量的工具包Word2vec，由于其简单、高效，引起了很多人的关注。

3.2 文档结构化表示

文档结构化表示的策略有很多，其中文档的内容可用语言基本单位的组合来表示，最常用的中文文本基本单位为词。例如，用 $\boldsymbol{d}$ 来表示一篇文档，那么 $\boldsymbol{d}$ 可以由表现词特征的词项所构成，即：

$$\boldsymbol{d}=(T_1,\cdots T_k,\cdots,T_L) \tag{1}$$

其中，T_k 是第 k 个词 w_k 的特征项，$1\leqslant k\leqslant L$，L 是语料包含的词的数量，文档 $\boldsymbol{d}$ 被表示为一个 L 维向量，其中每个元素均为特征项。

首先对文档进行分词处理，再用词的特征项来表示文档。本文采用常见的结巴分词工具[1]对文档进行分词。借助停用

[1] https://pypi.org/project/jieba/

词表去除语料中无意义的短语、虚词、标点符号等。通过词频、逆文档词频等统计特征项，可得到文档的一种常见的向量表示模型。

另一种流行的方法是使用词的分布式表示，即词向量来表示文档。假设在分词处理后，第 i 篇文档 $\boldsymbol{d}_i$ 的第 k 个词为 w_{ik}，将每个文档的词按顺序存入词袋。给定词向量维度，使用 Gensim 库中 Word2vec 工具，对词袋中的词进行训练可以得到每个词的词向量表示，记 w_{ik} 的词向量为 $\boldsymbol{v}_{ik}\in\mathbb{R}^M$，其维度为 M。那么，可以通过对文档 $\boldsymbol{d}_i$ 中所有词的词向量进行线性或非线性的加权平均来对其进行表示。下式给出词向量的线性加权形式：

$$\boldsymbol{d}_i=\sum_{k=1}^{L_i}\omega_{ik}\boldsymbol{v}_{ik} \tag{2}$$

其中，L_i 为文档 $\boldsymbol{d}_i$ 包含词的个数，ω_{ik} 为加权系数。为了简便，本文取 $\omega_{ik}=1/L_i$。由式(2)可知，文档 $\boldsymbol{d}_i$ 被表示为一个与词向量维度相同的 M-维向量。与基于词项特征的表示方式式(1)相比，当 $M\ll L$ 时，该表示是对文档较低维的表示形式，这对后续的文本处理任务，有降低算法复杂度和提高算法效率的作用。

3.3 稀疏表示

3.3.1 稀疏表示理论

稀疏表示是由 Candès 等[12]提出的一种新的理论框架，其目的是利用现有的信号构建冗余字典，并通过字典原子的线性组合来稀疏地表示目标信号。该理论表明：目标信号稀疏或可压缩时，可以通过观测得到的少量信息来重构目标信号[13]。给定字典 $\boldsymbol{D}=[\boldsymbol{d}_1,\cdots,\boldsymbol{d}_N]$，目标信号 $\boldsymbol{x}$ 的稀疏表示模型如下：

$$\boldsymbol{x}=\boldsymbol{D}\boldsymbol{\alpha}=\sum_{i=1}^{N}\alpha_i\boldsymbol{d}_i \tag{3}$$

其中，$\boldsymbol{\alpha}=[\alpha_1,\alpha_2,\cdots,\alpha_N]^{\mathrm{T}}$ 为 $\boldsymbol{x}$ 包含少量非零元素的稀疏系数向量。

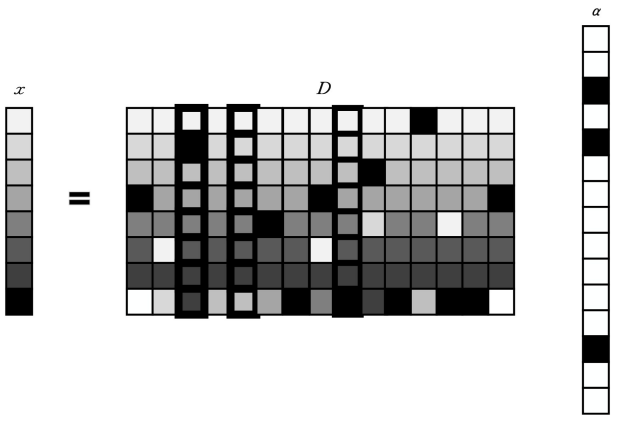


图 1 稀疏表示示意图

Fig. 1 Sparse representation

对于一个信号，当其有 k 个非零元素时，则称该信号为 k 稀疏信号。图 1 中，目标信号 $x\in\mathbb{R}^8$ 在冗余字典 $D\in\mathbb{R}^{8\times15}$ 表示下的变换系数向量 $\boldsymbol{\alpha}\in\mathbb{R}^{15}$ 只包含 3 个非零元素。这表示信号 $\boldsymbol{x}$ 可由字典中相应的 3 个原子的线性组合表示。称 $\boldsymbol{\alpha}$ 为 3-稀疏信号，$\boldsymbol{x}$ 在字典 $\boldsymbol{D}$ 下具有变换稀疏性[14]。

获取信号的稀疏表示包含两个重要的研究内容：字典学习和稀疏系数求解，如图 2 所示。

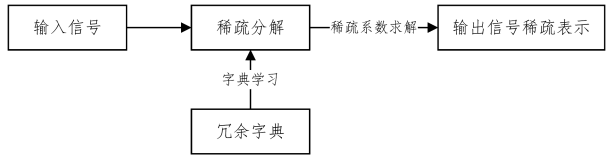


图 2 获取稀疏表示

Fig. 2 Obtain sparse representation

3.3.2 字典学习

在稀疏表示问题中，构建合适的字典是保证目标信号可被稀疏表示的基础和前提。近年来，出现了大量的学习自适应于信号样本的字典的研究。使用学习的字典，而不是固定字典中的原子的线性组合来稀疏表示信号，可以显著改善稀疏性，在图像处理等任务中可以取得更好的效果。

本文主要采用经典的 KSVD 算法[15]，KSVD 算法通过 k 次矩阵奇异值分解，在更新字典原子的同时，更新与之相关的稀疏系数，很大程度上降低了时间复杂度。针对海量数据，这种高效率的字典训练方法较为适合。KSVD 算法的基本步骤如下：

Step1 初始化字典 $\boldsymbol{D}$，所有原子进行归一化处理，初始化迭代次数。

Step2 稀疏编码：首先固定字典 $\boldsymbol{D}$，求解训练样本在字典 $\boldsymbol{D}$ 下的稀疏系数的优化问题。

Step3 字典更新：固定稀疏系数矩阵，对字典 $\boldsymbol{D}$ 的原子进行更新。基于残差最小原则，每次仅更新某一列原子及其稀疏系数。

Step4 若迭代次数小于 k，则转到 Step2；否则，算法结束。

3.3.3 稀疏系数求解

稀疏系数求解问题研究的是如何在过完备字典下得到信号的稀疏系数，即获得式(3)所示的信号的稀疏表示。通常需要借助重构误差最小化，通过复杂的迭代计算来求解非线性的优化问题[8]。

稀疏系数求解问题可表示为如下的 ℓ_0-范数非凸优化问题：

$$\begin{cases}\min\limits_{\boldsymbol{\alpha}}\|\boldsymbol{\alpha}\|_0\\ \text{s.t. }\boldsymbol{x}=\boldsymbol{D}\boldsymbol{\alpha}\end{cases} \tag{4}$$

ℓ_0-问题是 NP-难问题，也就是说，要在 N 维空间中找到信号 $\boldsymbol{x}$ 在字典 $\boldsymbol{D}$ 下的 k-稀疏表示系数向量 $\boldsymbol{\alpha}$，需要在 C_N^k 个可行解中进行搜索，随着 N 的增大，可行解的个数呈指数增长。

Candès 等[12]的研究表示，在一定条件下，ℓ_0-范数优化问题可等价地转化为 ℓ_1-最小化问题：

$$\begin{cases}\min\limits_{\boldsymbol{\alpha}}\|\boldsymbol{\alpha}\|_1\\ \text{s.t. }\boldsymbol{x}=\boldsymbol{D}\boldsymbol{\alpha}\end{cases} \tag{5}$$

ℓ_1-范数最小化是一个凸优化问题，针对其已提出了许多基于凸松弛的算法，但此类算法的计算代价昂贵。

本文主要使用贪婪算法中的正交匹配追踪(Orthogonal matching pursuit，OMP)算法[16-17]进行稀疏系数求解，其保证了最佳的循环效果，同时还降低了迭代时间成本。其基本思想是，在每次迭代更新中对所有字典中的原子做正交投影变换，以保证所有原子间的相互正交性，并利用字典原子与残差之间的最大相关性，确定与字典原子相对应的系数，并对其进行修改来更新稀疏系数向量的当前估计，迭代直至收敛。其步骤如下：

Step1 输入语句信号 x、过完备字典 D 和稀疏度 k；

Step2 初始化各参数；

Step3 选择字典中的原子，找出于当前残差最为匹配的原子索引；

Step4 更新系数；

Step5 更新残差；

Step6 若满足迭代次数条件，算法终止；否则转 Step3；

Step7 输出稀疏表示系数。

稀疏表示的成功得益于自然信号内在的稀疏性和可压缩性。非结构化的文本数据在结构化之后，往往具有高维和稀疏的特性，故稀疏表示理论适用于文本任务处理。

4 基于语义空间动态划分的多文档摘要模型

在本节中，本文提出基于稀疏表示的动态语义空间划分(SR-DSSP)算法，迭代地实现语义空间的划分，并提出一种基于稀疏相似度排序(SS-Rank)的自动摘要提取算法，在划分后的语义子空间中进行摘要句的提取。最终形成多文档的摘要，具体的处理过程如图 3 所示。

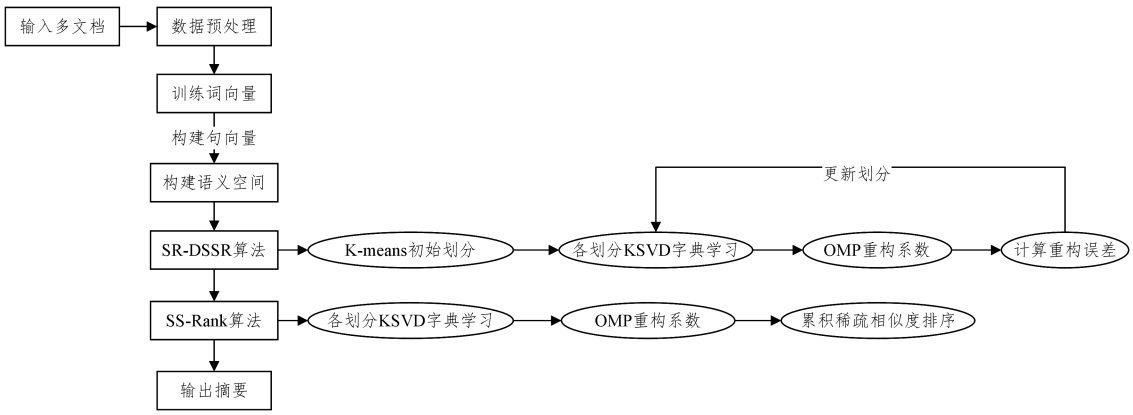


图 3 多文档自动摘要处理流程

Fig. 3 Processing flow of multi-document automatic summarization

4.1 句向量语义空间构建

对于自动摘要任务，摘要由句子组成，故从句子层面进行设计。类似文档的结构化表示，在多文档自动摘要任务中，需要先对文档中的句子进行结构化表示，再做进一步提取。以句号等常见句末标点为标志，对文档划分句。记文档 $\boldsymbol{d}_i$ 的第 j 个句子为 $\boldsymbol{x}_{ij}$，那么可以以句子为特征项对 d_i 进行表示：

$$\boldsymbol{d}_i=(\boldsymbol{x}_{i1},\cdots,\boldsymbol{x}_{ij},\cdots,\boldsymbol{x}_{iN_i}) \tag{6}$$

其中，N_i 是 $\boldsymbol{d}_i$ 中句子的个数。那么对于 I 个文档共有 $N=\sum_{i=1}^{I}N_i$ 个句子。

对句子进行分词处理，用词的特征来表示每个句子。根据 3.1 节中的讨论，使用 Word2vec 工具对大语料库进行训练，得到语料库中每个词的词向量之后，将词向量按一定权重线性或非线性组合得到句向量表示。本文采用类似于式(7)的加权平均得到句向量：

$$\boldsymbol{x}_{ij}=\sum_{k=1}^{L_{ij}}\omega_{ijk}\boldsymbol{v}_{ijk} \tag{7}$$

其中，L_{ij} 为文档 $\boldsymbol{d}_i$ 的第 j 个句子 $\boldsymbol{x}_{ij}$ 包含的词的个数，ω_{ijk} 为加权系数。本文中取 $\omega_{ijk}=1/L_{ij}$。由此，对需要进行摘要的 I 个文档($\boldsymbol{d}_1,\boldsymbol{d}_2,\cdots,\boldsymbol{d}_I$)，可以构建包含 N 个句子的句向量语义空间，记为 $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N)$，其中 $\boldsymbol{x}_j\in\mathbb{R}^M$ 是第 j 个句向量。

由词向量构成的文档语义空间中，每一个词基于词向量的表示可看作空间中的一点，句子可看作空间点的有序集合，以空间曲线段的形式存在。语义相近的词向量在距离上更加接近，呈现出簇状分布。由式(7)可知，句向量由词向量组合而成，故句向量在句子的语义空间中也呈簇状分布。从而可以考虑将句子的语义空间进行划分，在各子空间中选取能够代表文本局部语义的句子作为摘要句，使摘要句能够全面地表征文本语义特征，避免对文本主题表征的偏移。

4.2 基于稀疏表示的动态语义空间划分

一般的聚类方法，如 K-means 等，对句向量的聚类可以获得语义空间的粗划分，但该划分无法有效区分低维子流型数据特征。因此，本文提出一种通过子空间语义字典学习和对句向量稀疏重构的动态语义空间划分算法——基于稀疏表示的动态语义空间划分(SR-DSSP)算法。SR-DSSP 算法主要包括以下两个步骤：1)子空间语义字典学习：通过字典学习获得每个子空间的语义字典；2)基于稀疏重构的语义空间重划分：通过计算句向量在每个子空间的语义字典下的稀疏重构误差，将句向量所属划分进行调整，获得语义空间划分的更新。

4.2.1 子空间语义字典学习

用 $\Omega=(\Omega_1,\cdots,\Omega_k)$ 表示对句向量语义空间划分，其中 K 为划分的个数，第 k 个划分 Ω_k 中的元素为被聚类到该划分的句向量的索引。利用划分 $\Omega=(\Omega_1,\cdots,\Omega_k)$，可以得到句向量语义空间 $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N)$ 的 K 个语义子空间，记第 k 个语义子空间为 $\boldsymbol{X}_{\Omega_k}$。对每个语义子空间，学习子空间的语义字典。对第 k 个语义子空间 $\boldsymbol{X}_{\Omega_k}$，字典学习的优化问题如下：

$$\min_{\boldsymbol{D}_k,\boldsymbol{A}_k}\|\boldsymbol{X}_{\Omega_k}-\boldsymbol{D}_k\boldsymbol{A}_k\|_F^2+\lambda\sum_{n=1}^{N_k}\|\boldsymbol{\alpha}_k\|_0 \tag{8}$$

其中，λ 为正则化参数，$\boldsymbol{D}_k\in\mathbb{R}^{M\times M}$ 为子空间语义字典，$\boldsymbol{A}_k=[\boldsymbol{\alpha}_{1,k},\cdots,\boldsymbol{\alpha}_{N_k,k}]\in\mathbb{R}^{M\times N_k}$ 为稀疏系数矩阵，其列向量 $\boldsymbol{\alpha}_{n,k}\in\mathbb{R}^M$ 为子空间 $\boldsymbol{X}_{\Omega_k}$ 中第 n 个句向量 $\boldsymbol{x}_{n,k}\in\mathbb{R}^M$ 在字典 $\boldsymbol{D}_k$ 下的稀疏系数向量。式(8)的优化问题表示，希望找到字典 $\boldsymbol{D}_k$ 和稀疏系数矩阵 $\boldsymbol{A}_k$，使得 $\boldsymbol{D}_k\boldsymbol{A}_k$ 能够尽可能地还原 $\boldsymbol{X}_{\Omega_k}$，并且 $\boldsymbol{A}_k$ 的每个列向量尽可能的稀疏。求解优化问题式(3)的一般方法是，固定其中一个变量，优化另一个变量，如此交替进行。本文中采用经典的 KSVD 算法求解式(9)，获得子空间的语义字典 $\boldsymbol{D}_k$。

4.2.2 基于稀疏重构的语义空间重划分

在得到 K 个子空间的语义字典后，通过计算每个句向量在语义字典下的稀疏重构误差对语义空间重新进行划分。假设在某次迭代中得到 K 个子空间的语义字典 $\boldsymbol{D}_1,\boldsymbol{D}_2,\cdots,\boldsymbol{D}_K$，对第 n 个句向量 $\boldsymbol{x}_n$，分别求解其在 K 个语义字典下的稀疏表示。$\boldsymbol{x}_n$ 在第 k 个字典 $\boldsymbol{D}_k$ 下的稀疏系数向量 $\hat{\boldsymbol{\alpha}}_{n,k}$ 可通过求解如下优化问题得到。

$$\hat{\boldsymbol{\alpha}}_{n,k}=\arg\min_{\boldsymbol{\alpha}_{n,k}}\|\boldsymbol{x}_n-\boldsymbol{D}_k\boldsymbol{\alpha}_{n,k}\|_2^2+\lambda_1\|\boldsymbol{\alpha}_{n,k}\|_0 \tag{9}$$

其中，λ_1 为正则化参数，$\boldsymbol{\alpha}_{n,k}\in\mathbb{R}^M$ 为稀疏系数向量。如 3.3.3 节所述，式(9)中的 ℓ_0-问题可转化为 ℓ_1-最小化，使用贪婪算法进行求解，本文使用 OMP 算法对式(9)进行求解得到 $\boldsymbol{x}_n$ 在 $\boldsymbol{D}_1,\boldsymbol{D}_2,\cdots,\boldsymbol{D}_K$ 下的稀疏系数向量 $\hat{\boldsymbol{\alpha}}_{n,1},\cdots,\hat{\boldsymbol{\alpha}}_{n,K}$ 后，计算由稀疏系数重构 x_n 的重构误差。由第 k 个稀疏系数向量 $\hat{\boldsymbol{\alpha}}_{n,k}$ 重构 x_n 的误差为：

$$r_{n,k}=\|\boldsymbol{x}_n-\boldsymbol{D}_k\hat{\boldsymbol{\alpha}}_{n,k}\|_2^2 \tag{10}$$

对 K 个重构误差进行排序，选择将 x_n 分配到具有最小重构误差的划分。

$$\boldsymbol{x}_n\to\Omega_{\hat{k}} \tag{11}$$

其中，

$$\hat{k}=\arg\min\{r_{n,k}\} \tag{12}$$

通过对所有的句向量的重分配，可以实现语义空间划分的更新。

在给定语义空间的初始划分下，SR-DSSP 算法在子空间语义字典学习和语义空间重划分间交替迭代，直到满足停止条件，得到语义空间的稳定划分。

SR-DSSP 算法流程如算法 1 所示。其中 $\boldsymbol{X}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_N)$ 表示待摘要的文档语料包含的句向量的集合，即句向量的语义空间，$\boldsymbol{x}_n$ 为第 n 个句向量，N 为语义空间中句向量的个数，$\Omega=(\Omega_1,\cdots,\Omega_k)$ 表示对句向量语义空间最终的划分。

以由 K-means 算法得到语义空间划分 $\Omega^0=(\Omega_1^0,\cdots,\Omega_k^0)$ 作为初始划分，$R^0=0$ 作为总重构误差的初始值输入算法。在第 i 次迭代中，首先在由划分 Ω^i 确定的每个语义子空间 $\boldsymbol{X}_{\Omega_k^i}$ 中学习字典 $\boldsymbol{D}_k^i(k=1,\cdots,K)$；其次，学习每个句子 $\boldsymbol{x}_n$ 在 K 个字典 $\boldsymbol{D}_1^i,\boldsymbol{D}_2^i,\cdots,\boldsymbol{D}_K^i$ 下的稀疏表示 $\boldsymbol{\alpha}_{n,k}^i(n=1,\cdots,N;k=1,\cdots,K)$，以及重构误差 $r_{n,k}^i$，并将 x_n 重新分配到具有最小重构误差的划分 Ω_k^{i+1}；最后，计算所有句向量的最小重构误差的总和 R^i。当前后两次迭代的总重构误差的变化大于阈值 ε 时，即 $|R^{i+1}-R^i|>\varepsilon$，或者小于最大的迭代次数 $I_{\max}$ 时，返回第 2 步，利用新划分下的句向量重新进行子空间语义字典学习和语义空间重划分，至满足停止条件。得到句向量语义空间的最终划分 Ω。

算法 1 SR-DSSP

输入：句向量集合 $\mathbf{X}=(\mathbf{x}_1,\mathbf{x}_2,\cdots,\mathbf{x}_N)$

输出：语义空间划分 $\Omega=(\Omega_1,\cdots,\Omega_k)$

1. $i=0$，$R^0=0$，使用 k-means 算法对句向量集 X 进行聚类，得到初始划分 $\Omega^0=(\Omega_1^0,\cdots,\Omega_k^0)$。
2. 对划分 $\Omega_k^i\neq\varphi$ 相应的句向量子集 $\mathbf{X}_{\Omega_k^i}$，通过 KSVD 算法求解优化问题式(3)，计算相应的字典 $\mathbf{D}_k^i$，$k=1,\cdots,K$。
3. 令 $\Omega^{i+1}=(\Omega_1^{i+1},\cdots,\Omega_K^{i+1})$，$\Omega_k^i\neq\varnothing$，$k=1,\cdots,K$。
4. $\forall n\in(1,\cdots,N)$，对句向量 $\mathbf{x}_n$。
   4.1. 利用 OMP 算法求解优化问题式(5)，得到 $\mathbf{x}_n$ 在每个字典 $\mathbf{D}_k^i$ $(k=1,\cdots,K)$ 下的稀疏表示系数 $\hat{\boldsymbol{\alpha}}_{n,k}^i$；
   4.2. 计算重构误差 $r_{n,k}^i=\|\mathbf{x}_n-\mathbf{D}_k^i\hat{\boldsymbol{\alpha}}_{n,k}^i\|_2^2$，$k=1,\cdots,K$；
   4.3. 更新划分：$\mathbf{x}_n\to\Omega_k^{i+1}$，其中 $\hat{k}=\arg\min_k\{r_{n,k}^i\}$。
5. 对所有的 $n\in\{1,\cdots,N\}$，计算句向量 $\mathbf{x}_n$ 在新划分 Ω^{i+1} 下的总重构误差 $R^i=\sum_n\{\min_k\{r_{n,k}^i\}\}$。
6. 如果 $|R^{i+1}-R^i|>\varepsilon$，或者 $i<I_{\max}$，$i=i+1$，转到第 2 步。
7. $\Omega=\Omega^i$。

4.3 基于稀疏相似度排序的摘要提取

对划分后语义子空间中摘要句的提取，本文提出一种基于稀疏相似度排序(SS-Rank)的自动摘要提取算法。SS-Rank 算法在每个语义子空间中学习句向量的稀疏表示，计算各句向量对其余句向量的累积稀疏相似度，并依据其排序来提取摘要句。具体的算法如下：

记第 k 个划分 Ω_k 对应的句向量子空间为 $\boldsymbol{X}_{\Omega_k}=(\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_{N_k})$，对 $\boldsymbol{X}_{\Omega_k}$ 中的第 n 个句向量 $\boldsymbol{x}_n$，选取字典 $\boldsymbol{B}_{n,k}=[\boldsymbol{x}_1,\cdots,\boldsymbol{x}_{n-1},\boldsymbol{x}_{n+1},\cdots,\boldsymbol{x}_{N_k}]$，即 $\boldsymbol{B}_{n,k}$ 是以子空间中不同于 x_n 的其他句向量作为字典原子构成。$\boldsymbol{x}_n$ 在字典 $\boldsymbol{B}_{n,k}$ 下的稀疏表示可以由 ℓ_0-范数最小化问题得到：

$$\min_{\boldsymbol{\beta}_n}\|\boldsymbol{x}_n-\boldsymbol{B}_{n,k}\boldsymbol{\beta}_n\|_2^2+\lambda_2\|\boldsymbol{\beta}_n\|_0 \tag{13}$$

其中，λ_2 为正则化参数，$\boldsymbol{\beta}_n=(\boldsymbol{\beta}_{n,1},\cdots,\boldsymbol{\beta}_{n,n-1},\boldsymbol{\beta}_{n,n+1},\cdots,\boldsymbol{\beta}_{n,N_k})^{\mathrm{T}}$ 是稀疏系数向量。对式(13)的优化问题本文仍使用 OMP 算法求解。

文献[18]提出一种基于稀疏表示的相似性测量，研究表明，稀疏系数的幅值越大，字典中相应的原子与信号的相似性越强。例如，$|\beta_{n,j}|$ 越大，代表 $\boldsymbol{B}_{n,k}$ 中的 $\boldsymbol{x}_j$ 与 $\boldsymbol{x}_n$ 越相似。换句话说，$|\boldsymbol{\beta}_{n,j}|$ 越大，$\boldsymbol{x}_j$ 对 $\boldsymbol{x}_n$ 的表示能力越强。在文本摘要任务中，希望提取的摘要句能对其他句子在语义上有很强的表示能力，因此，本文将 $\boldsymbol{x}_j$ 对 $\boldsymbol{x}_n$ 的表示能力定义为 $\boldsymbol{x}_j$ 对 $\boldsymbol{x}_n$ 的稀疏相似度：

$$s(\boldsymbol{x}_j,\boldsymbol{x}_n)=|\boldsymbol{\beta}_{n,j}| \tag{14}$$

需要注意的是，与文献[18]定义的稀疏相似度不同，在本文中，$s(\boldsymbol{x}_j,\boldsymbol{x}_n)\neq s(\boldsymbol{x}_n,\boldsymbol{x}_j)$。其中，$s(\boldsymbol{x}_j,\boldsymbol{x}_n)$ 代表 $\boldsymbol{x}_j$ 对 $\boldsymbol{x}_n$ 的表示能力，而 $s(\boldsymbol{x}_n,\boldsymbol{x}_j)$ 代表 $\boldsymbol{x}_n$ 对 $\boldsymbol{x}_j$ 的表示能力，分别通过求解两个不同的稀疏优化问题得到。

进一步，希望摘要句对其所属的语义子空间中的其他句子都具有较强的表示能力。因此，将 $\boldsymbol{x}_n$ 对其所属语义子空间 $\boldsymbol{X}_{\Omega_k}$ 的语义表示能力定义为 $\boldsymbol{x}_n$ 对其他句向量的累积稀疏相似度。

$$s_T(\boldsymbol{X}_{\Omega_k},\boldsymbol{x}_n)=\sum_{j\neq n}s(\boldsymbol{x}_j,\boldsymbol{x}_n)=\sum_{j\neq n}|\beta_{n,j}| \tag{15}$$

其中，$j=1,2,\cdots,N_k$，且 $j\neq n$。

最后，对 $\boldsymbol{X}_{\Omega_k}$ 中每个句向量的累积稀疏相似度进行排序：

$$\operatorname*{Rank}_n\{s_T(\boldsymbol{X}_{\Omega_k},\boldsymbol{x}_n),n=1,\cdots,N_k\} \tag{16}$$

依据此排序，可以选取满足 $s_T(\boldsymbol{X}_{\Omega_k},\boldsymbol{x}_n)>\varepsilon_s$ 的句向量对应的句子作为语义子空间 $\boldsymbol{X}_{\Omega_k}$ 的摘要句，其中 ε_s 为相似度阈值。SS-Rank 算法流程如算法 2 所示。

算法 2 SS-Rank

输入：语义子空间 $\mathbf{X}_{\Omega_1},\mathbf{X}_{\Omega_2},\cdots,\mathbf{X}_{\Omega_k}$

输出：多文档摘要 $\widetilde{\mathbf{X}}=(\widetilde{\mathbf{X}}_1,\cdots,\widetilde{\mathbf{X}}_K)$

1. $\forall k\in(1,\cdots,K)$，对语义子空间 $\mathbf{X}_{\Omega_k}$ 提取摘要句 $\widetilde{\mathbf{X}}_k=(\widetilde{\mathbf{x}}_1,\cdots,\widetilde{\mathbf{x}}_{L_k})$。

1.1. $\forall n\in(1,\cdots,N_k)$，计算句向量 $\mathbf{x}_n$ 对 $\mathbf{X}_{\Omega_k}$ 的累积稀疏相似度。

1.1.1. 利用 OMP 算法求解式(13)，得到 $\mathbf{x}_n$ 在字典 $\mathbf{B}_{n,k}$ 下的稀疏系数 $\boldsymbol{\beta}_n$。

1.1.2. 由式(15)计算 $\mathbf{x}_n$ 累积稀疏相似度 $s_T(\mathbf{X}_{\Omega_k},\mathbf{x}_n)$。

1.2. $\forall n\in(1,\cdots,N_k)$，由式(16)对 $\mathbf{X}_{\Omega_k}$ 中所有句向量的累积稀疏相似度排序。

1.3. 依据排序提取满足 $s_T(\mathbf{X}_{\Omega_k},\boldsymbol{x}_n)>\varepsilon_s$ 的句子作为 $\mathbf{X}_{\Omega_k}$ 的摘要句 $\widetilde{\mathbf{X}}_k=(\widetilde{x}_1,\cdots,\widetilde{x}_{L_k})$。

2. 形成多文档的摘要 $\widetilde{\mathbf{X}}=(\widetilde{\mathbf{X}}_1,\cdots,\widetilde{\mathbf{X}}_K)$。

5 仿真实验

5.1 实验环境

实验各算法均基于 Python 3.6.3 环境实现，使用 Spyder3 IDE 进行编译与运行。程序主要依赖的第三方库及相应的版本为：Scikit-learn (0.19.1)，Numpy (1.13.3)，Jieba (0.39)。实验计算机硬件配置为：处理器 Intel(R) Xeon(R) E5-2623 v4，内存 16GB，操作系统为 Windows 10 企业版。

5.2 实验语料

复旦中文语料[1)]是一个多类别中文语料库。本文采用其部分数据进行实验，共 19637 篇文档，48010861 个有效词。实验在该语料库的基础上，爬取猫途鹰网站热门景点旅游评论数据，进行词向量训练。对各景点评论数据进行实验，提取景点旅游评论的摘要。各景点旅游评论数据统计如表 1、表 2 所列。

表 1 景点评论数据统计

Table 1 Scenic spots comment statistics

编号	景点名称	评论条数	有效词数
1	安徽-九华山	594	101917
2	北京-故宫博物院	3874	14429
3	浙江-杭州西湖	6818	179008
4	福建-厦门鼓浪屿	8575	224332
…	…	…	…

表 2 景点评论数据统计

Table 2 Scenic spots comment statistics

景点数量	91
平均评论数/个	803
平均有效词数/个	21895.26
平均有效词数/条	27.27
总词数	1992469

实验对 91 个景点评论数据进行自动提取摘要。将单条评论视为一个单文档，因其均对同一景点进行评价，故存在相似语义内容，可用于验证所提算法的有效性。实验摘要结果分析只列举表 4 个景点的部分摘要。

实验共分为 3 部分：1)训练不同维度的词向量；2)迭代优化子空间次数；3)与传统摘要算法作比较。

5.3 结果分析

5.3.1 重构误差分析

在本节中，对使用 SR-DSSP 算法进行语义空间动态划分中的重构误差的变化进行分析。

为使算法更稳定，以每个划分 $\Omega_k(k=1,2,\cdots,K)$ 下的语义子空间 $\boldsymbol{X}_{\Omega_k}$ 所包含的句向量个数 N_k 为约束，要求词向量 $\boldsymbol{x}_n\in\mathbb{R}^M$ 的维度 $M\leqslant N_k$。因此，本节试验分别训练 5，10，15，20，30，40，50，100 维度的词向量，以此对 SR-DSSP 算法基于不同维度词向量的语义空间进行划分时的重构误差进行分析比较。其中，Word2vec 库参数设定如表 3 所列。

表 3 模型参数

Table 3 Model parameters

window	4
max_vocab_size	None
min_count	1
sg	0
workers	4

其中，window 表示滑窗大小，含义为每个词向量将包含滑窗内词的语义信息；max_vocab_size 表示训练最大内存限制；min_count 表示最低词频阈值，小于词频阈值的词将被忽略；sg 为训练模型的选择，1 为 skip-gram 模型，0 为选取 CBOW 模型；workers 为训练进程数。

将复旦中文预料库连同所有评论数据一同作为训练语料训练词向量。

第 1 次的重构误差是指利用 K-means 算法得到的初始划分进行字典学习重构误差，可以看到在词向量的各个维度下，总重构误差都比较大，这说明由 K-means 算法得到的语义子空间中句向量之间的语义相关度不够显著，因而句向量在子空间语义字典下的稀疏重构的整体精度不够高。由表 4 可知，在词向量维度为 100 维的情况下，语义空间重划分可有效降低重构误差，平均压缩率达 17.3%。特别是，在第 2 次迭代时总重构误差大幅度下降，说明使用 SR-DSSP 算法对初始划分进行调整能够有效提高划分的合理性，加强子空间句向量之间的语义相关度，使得学习的语义字典能更好地刻画子空间的局部结构，从而得到句向量在语义字典下更优的稀疏表示。

表 4 SR-DSSP 结果分析(100 维)

Table 4 Analysis of SR-DSSP results(100-dimension)

重构误差	第 1 次	第 2 次	第 3 次	压缩率/%
1	106.848	68.8578	68.7924	35.61657682
2	126.051	112.716	117.012	7.170907014
3	210.364	158.057	160.151	23.86957844
…	…	…	…	…
91	131.502	134.827	134.658	−2.399963
总计平均	152.888	126.274	122.717	17.388129

由表 5 可知，不同维度的词向量均可在有限次迭代后有效降低重构误差，平均压缩率约为 16.2%，且均保持收敛的特性。证明了该方法对数据维度不敏感，能有效避免因数据维度不同而效果相差较大的问题。证明该算法达到了预期优化语义空间划分的效果。

表 5 SR-DSSP 结果分析

Table 5 Analysis of SR-DSSP results

平均重构误差(91 景区)	第 1 次	第 2 次	第 3 次	平均压缩率/%
5 维	27.7667	22.73899	22.7722	17.82983
50 维	108.572	92.47944	90.9626	15.28957
100 维	152.888	126.274	122.717	17.388129
总计平均	125.985	105.5482	103.555	16.21755

1) http://www.nlpir.org/wordpress/download/tc-corpus-answer.rar

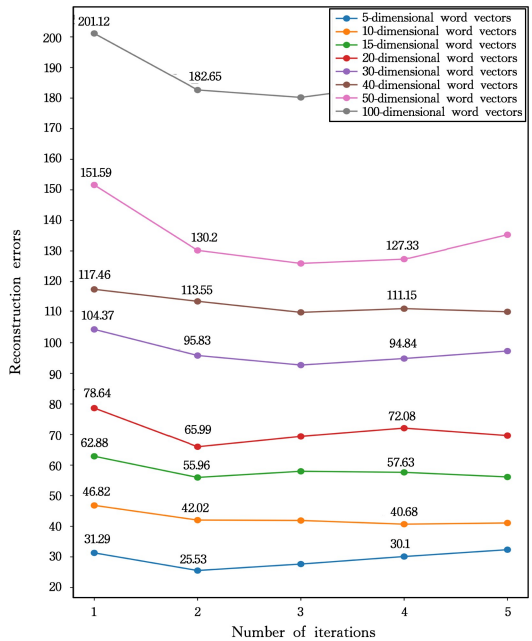
表 6 结果从另一方面进一步表明了该算法对同一数据的高维度表示和低维度表示不敏感，可有效针对数据内在信息降低重构误差。

图 4 给出 SR-DSSP 算法在词向量的不同维度下语义空间划分时的部分重构误差 R（见 SR-DSSP 算法第 5 步）随迭代次数变化的趋势图。

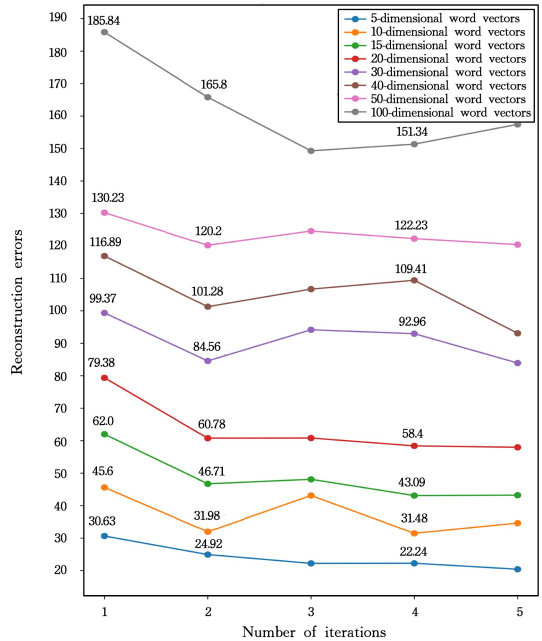
表 6 SR-DSSP 结果分析

Table 6 Analysis of SR-DSSP results

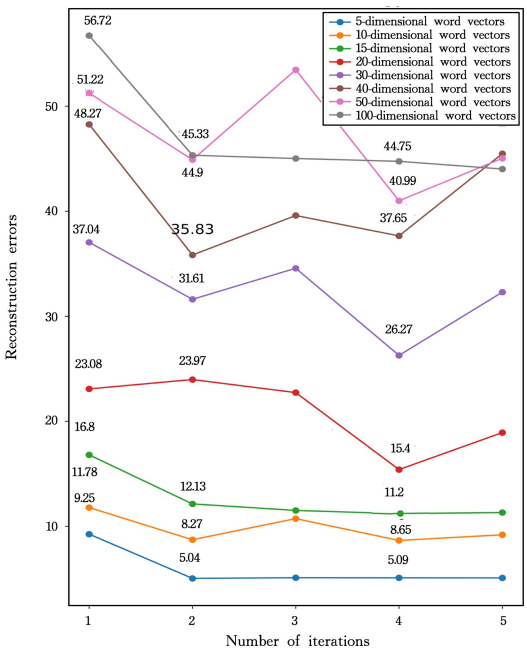
平均重构误差	5 维	50 维度	100 维	平均压缩率/%
1	24.2206	6.500042	35.6166	22.1124
2	8.5066	18.29124	7.17091	11.32292
3	24.2871	25.8823	23.8696	24.67964
…	…	…	…	…
91	0.99686	0.419054	55.2694	18.23054
总计平均	17.8298	15.28957	17.3881	16.83584



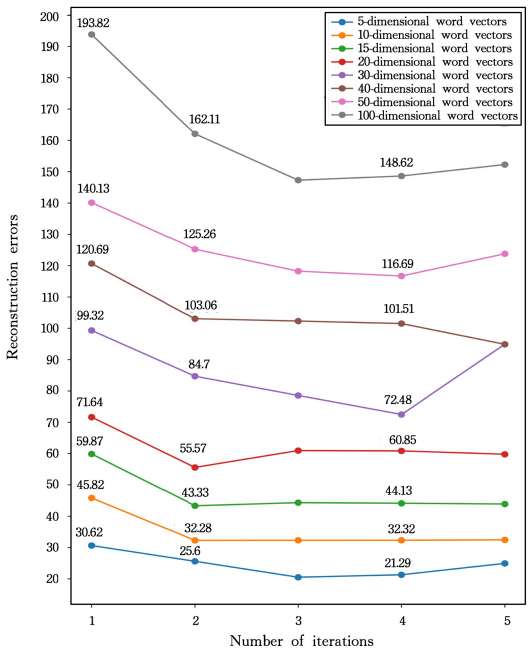
(a)北京故宫



(b)安徽九华山



(c)厦门鼓浪屿



(d)杭州西湖

图 4 SR-DSSP 算法迭代过程中的重构误差变化

Fig. 4 Changes of reconstruction error during SR-DSSP algorithm iteration

图 5(a)－图 5(d)分别为对 4 个旅游景点评论数据实验的结果。图中不同颜色的曲线表示词向量的不同维度；横坐标为迭代次数变化，实验给出了 SR-DSSP 算法在第 1－5 次迭代中的总重构误差。随着迭代次数的增大，可以看到在词向量的各个维度下的总重构误差逐步下降。在第 2－3 次迭代之后，重构误差呈现出小幅震荡，语义空间的划分已趋于稳定，表明该算法可快速收敛。词向量维度越高，整体重构误差均较大，表明同一数据的结构特征因表示维度的不同而差异较大。可根据经验从实验结果中选择算法最大迭代次数 $I_{\max}=5$。

5.3.2 摘要分析

在本节中，使用提出的基于语义空间动态划分的自动摘

要方法对各景区的旅游评论提取摘要，并与传统的摘要提取算法的结果进行比较。

实验中，令语义子空间的个数 $K=5$，SR-DSSP 算法的最大迭代次数 $I_{max}=5$，其中 KSVD 算法的参数为其默认值；OMP 算法中稀疏度为其默认值；对 SS-Rank 算法，可通过设置相似度阈值 ε_s 来控制提取摘要句的个数；在本节实验中，设置在每个语义子空间中提取前 5 个摘要句。本文所提摘要提取算法主要基于关键词的词序距离阈值对文章分族，根据句子得分排序，提取摘要。旅游评论自动摘要部分结果如表 7、表 8 所列。

表 7 九华山摘要结果分析

Table 7 Analysis of Jiuhuashan abstracted result

安徽-九华山	
基于语义空间动态划分的多文档摘要提取	'风景壮丽。','九华山景区是一个非常漂亮的地方五星级。'……
基于关键词得分的多文档摘要提取	'去了黄山，就不得不去九华山，安徽最出名的两座山之一，风景很美，大自然让九华山的大石头变得奇形怪状，美不胜收，一年四季，去了都值得看，另外还是佛教的朝拜的地方，据说很灵，一定要拜拜。'……

表 8 西湖摘要结果分析

Table 8 Analysis of the West Lake abstracted result

浙江-杭州西湖	
基于语义空间动态划分的多文档摘要提取	'景色很好，很漂亮。','西湖还是有雾的好，初去西湖便遇到雾天，西湖山水被雾笼罩起来，当真美不胜收。'……
基于关键词得分的多文档摘要提取	'西湖有三潭映月、雷峰塔等等美景，那一汪美丽的湖水映照着无数的人文墨客的美好时光，这的水美，景美，人更美漫步在这里有一种回归古时的感动，很美妙。'……

由以上摘要结果可以发现，传统的以关键词对句子进行标注打分的策略对句子长度不敏感，往往越长的句子得分越高，难以高效地提取摘要，而基于稀疏表示的多文档摘要提取策略不易忽略语义信息表达强烈的短句子，从而避免了遗漏重要信息句。并且该算法在多文档的情况下，避免了重复提取冗余度大、重复性高的文本内容，效果更佳。

由于摘要没有唯一的答案，结果评估太过主观。不同的人群对摘要的要求也各不相同，所以很难用统一的评价标准去衡量所有的摘要。针对信息抽取这类特殊情况，答案需由人工判定摘要信息是否"相关"或是否"重要"，是否表征了文档的主要信息。对于已经提供摘要的数据集，可以计算获得的摘要与标准摘要的语义相似度，以此评估算法的有效性。但对于类似评论等开放域数据集来说，目前仍没有特别合适的自动摘要评估方法，因此构建一种合理的评估方法也是该领域的一个热点问题。

结束语 目前，稀疏表示理论正在逐步的成熟和完善，针对字典学习算法和信号重构的讨论也有不小的进展。但目前其针自然语言处理中的各文本任务的研究还相对较少。尽管现在基于机器学习、深度学习的文本摘要技术效果较好，但其低解释性及高计算成本仍需要不断的改进。本文希望通过稀疏表示等理论结合目前的自然语言处理技术，对下游任务做出一些探索，为进一步改进算法提供思路。因此，基于稀疏表示的多文档自动摘要算法的研究具有很大的现实意义：1）空间划分理论可帮助算法进一步对语义空间有更抽象的理解，其重构误差作为训练词向量的一个实验方向；2）自由文本摘要类任务受各种因素的影响较大，引入带有常识性语义的词向量表示可对语义空间进行更多的探索。受设备等限制，算法可从以下 4 个方面加以改进：1）在 ELMO，Bert 等高效预训练词向量模型的基础上训练词向量，进一步采用其他模型进行指代消解、语义传递等处理；2）结合句子结构等语言模型进一步优化改进句向量表示；3）子空间初始化可采用超球 k-means 聚类等算法进行优化；4）现有稀疏系数求解算法运行效率较低，可结合神经网络模型等进一步设计针对高维数据的高效求解算法。

参 考 文 献

[1] ZHANG C. Text summary algorithm based on semantic reconstruction[D]. Nanjing: Nanjing University, 2016.

[2] ALLAHYARI M, POURIYEH S, ASSEFI M, et al. Text Summarization Techniques: A Brief Survey[J]. International Journal of Advanced Computer Science & Applications, 2017, 8(10): 397-405.

[3] FERILLI S, PAZIENZA A. An Abstract Argumentation-Based Approach to Automatic Extractive Text Summarization [C]// Italian Research Conference on Digital Libraries. Springer, Cham, 2018.

[4] LIU H, YU H, DENG Z H. Multi-document summarization based on two-level sparse representation model[C] // AAAI. 2015: 196-202.

[5] HE R, TANG J, GONG P, et al. Multi-document summarization via group sparse learning[J]. Information Sciences, 2016, 349: 12-24.

[6] HE Z, CHEN C, BU J, et al. Document summarization based on data reconstruction[C]// AAAI. 2012: 620-626.

[7] HE Z, CHEN C, BU J, et al. Unsupervised document summarization from data reconstruction perspective [J]. Neurocomputing, 2015, 157: 356-366.

[8] JIAO L C. Sparse learning, classification and recognition[M]. SCIENCE PRESS, 2017.

[9] XIONG X. Research on Extractive Answer Fusion for Q & A Community[D]. Harbin: Harbin Institute of Technology, 2018.

[10] PETERS M E, NEUMANN M, IYYER M, et al. Deep contextualized word representations[J]. arXiv: 1802.05365, 2018.

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

[12] CANDES E J, ROMBERG J K. Practical signal recovery from random projection [J]. Proc Spie, 2005, 5674: 76-86.

[13] PENG S. Sparse representation coding model and its application in text classification [D]. Tianjin: Tianjin University, 2015.

[14] DAVENPORT M A, DUARTE M F, ELDAR Y C, et al. Introduction to Compressed Sensing[M]// Compressed Sensing: Theory and Applications. Cambridge: Cambridge University Press, 2012.

[15] AHARON M, ELAD M, BRUCKSTEIN A. K-SVD: an algorithm for designing overcomplete dictionaries for sparse repre-

sentation[J]. IEEE Transactions on Image Processing, 2006, 54(11):4311-4322.
[16] PATI Y C, REZAIITAR R, KRISHNAPRASAD P S. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition[C]// Proceeding of the 27thAsilomar Conference on Signals, Systems and Computers. 1993:40-44.
[17] DAVIS G, MALLAT S, AVELLANEDAM. Adaptive greedy approximation[J]. Constructive Approximation, 1997, 13 (1): 57-98.
[18] CHENG H, LIU Z, HOU L, et al. Sparsity-Induced Similarity Measure and Its Applications[J]. IEEE Transactions on Circuits & Systems for Video Technology, 2016, 26(4):613-626.

QIAN Ling-long, born in 1996, postgraduate, is a member of China Computer Federation. His main research interest include national language processing, knowledge graph and explainable AI.

LU Hui-juan, born in 1962, professor, is director and an outstanding member of China Computer Federation. Her main research interest include machine learning, deep learning and big data.

(上接第 85 页)

标准矛盾体添加新子句及向其相关子句添加相关文字后，其最小标准矛盾体呈现的规律。以上结果不仅丰富了标准矛盾体演绎中增加文字的策略，而且理论上使得完全标准矛盾体可通过增加新子句或向相关子句添加相关文字完成与最小标准矛盾体的相互转换。后一性质为标准矛盾体分离演绎理论进一步应用于计算机求解问题提供了一定的理论基础。

参 考 文 献

[1] ROBINSON J A. A Machine-Oriented Logic Based on the Resolution Principle[J]. Journal of theAcm, 1965, 12(1):23-41.
[2] HARRISON J. Handbook of practical logic and automated reasoning[M]. New York: Cambridge University Press, 2009:631-667.
[3] PLAISTED D A. History and Prospects for First-Order Automated Deduction[C]// International Conference on Automated Deduction. Springer Verlag, 2015, 9195:3-28.
[4] ROBINSON J A. Handbook of Automated Reasoning[M]. New York: MIT Press, 2001.
[5] BURESH-OPPENHEIM J, PITASSI T. The Complexity of Resolution Refinements[J]. Journal of Symbolic Logic, 2007, 72(4):1336-1352.
[6] CHANG C L, RICHARD C T. Symbolic Logic and Mechanical Theorem Proving[M]. New York: Academic Press, 1973.
[7] RUBIN N, HARRISON M C. Another Generalization of Resolution[J]. Journal of the ACM(JACM), 1978, 25(3):341-351.
[8] REGER G, SUDA M, VORONKOV A, et al. Selecting the selection[C]// The 8th International Joint Conference on Automated Reasoning(IJCAR2016). Coimbra, Portugal, 2016:313-329.
[9] SCHULZ S, MOHRMANN M. Performance of clause selection heuristics for saturation-based theorem proving[C]// The 8th International Joint Conference on Automated Reasoning (IJCAR2016). 2016:330-345.
[10] SUTCLIFFE G. The TPTP problem library and associated infrastructure: the FOF and CNF parts, v2. 10. 0[J]. Journal of Automated Reasoning, 2009, 43(4):337-362.
[11] KALISZYK C, SCHULZ S, URBAN J, et al. System Description: E. T. 0. 1[C]// Proceedings of the 25th International Conference on AutomatedDeduction (CADE-25). Berlin: Springer International Publishing, 2015:389-398.
[12] GOZNY J, PALEO B W. Towards the compression of first-order resolution proofs by lowering unit clauses[C]// Automated Deduction(CADE-25). Berlin: Springer International Publishing, 2015:356-366.
[13] XU Y, LIU J, CHEN S W, et al. Contradiction separation based dynamic multi-clause synergized automated deduction[J]. Information Sciences, 2018, 462:93-113.
[14] CAO F, XU Y, CHEN S W, et al. Application of Multi-Clause Synergized Deduction in First-Order Logic Automated Theorem Proving[J]. Chinese Journal of Southwest Jiaotong University, 2020, 55(2):401-408.
[15] CAO F, XU Y, WU G F, et al. Application of multi-clause dynamic deduction in Prover9[J]. Chinese Computer Engineering & Science, 2019, 41(9):1686-1692.

TANG Lei-ming, born in 1996, master. His main research interests include propositional logic and SAT problem.

HE Xing-xing, born in 1982, Ph.D, associate professor, is a member of China Computer Federation. His main research interests include automated reasoning based on logic and so on.