曹中华,黄欣,彭文忠,等. 基于词嵌入特性聚类的文本主题挖掘[J]. 江西师范大学学报(自然科学版),2022,46(5):468-474.
CAO Zhonghua,HUANG Xin,PENG Wenzhong,et al. The topic mining based on word embedding characteristics clustering[J]. Journal of Jiangxi Normal University(Natural Science),2022,46(5):468-474.

文章编号:1000-5862(2022)05-0468-07

基于词嵌入特性聚类的文本主题挖掘

曹中华[1],黄 欣[1],彭文忠[2],刘媛春[1]
(1. 江西师范大学软件学院,江西 南昌 330022;2. 江西财经大学信息管理学院,江西 南昌 330032)

摘要: 数据聚类是常用的无监督学习方法,通过词嵌入聚类能够挖掘文本主题,但现有研究大多数采用常规聚类算法挖掘词嵌入的簇类,缺少基于词嵌入特性设计实现词嵌入聚类的主题挖掘算法. 该文从语言模型通过建模词间相关信息来使相关及语义相似词的嵌入表示聚集在一起的特点出发,设计词嵌入聚类算法. 该算法首先计算中心词的簇类号,然后使该簇中心嵌入和相邻词嵌入的相似性增强,同时使其与负样本词嵌入远离,学习文本集词嵌入的簇类结构,并将其应用于文本主题挖掘. 在3种公开数据集上的实验表明:该算法在一些模型的词嵌入结果上能够挖掘出一致性和多样性更好的主题结果.

关键词: 词嵌入;聚类;语言模型;文本主题
中图分类号: TP 391 **文献标志码:** A **DOI:** 10.16357/j.cnki.issn1000-5862.2022.05.05

0 引言

主题模型常用于挖掘在大批量文本集中隐含的主题语义信息,获得表示文本集语义的主题词,用户通过主题词集能够了解每篇文本的主题分布和文本集的主要内容. 隐含狄利克雷分配(latent Dirichlet allocation,LDA)[1]是一种广泛使用的主题语义分析模型,它以概率方式描述文本内容的生成,通过近似求解的方法获得文本的主题信息,但是LDA使用词袋模型描述文本生成,忽略了词间序列等关键语义信息.

词嵌入表示[2-4]含有丰富的词间相关和相似语义,可以为主题模型提供词间序列等语义信息,目前已经出现了许多应用词嵌入的文本主题挖掘模型[5]. 研究者或是将词嵌入应用于概率模型中,求解词袋模型缺少词语义信息等问题[6-7];或是将词嵌入应用于神经网络结构的主题模型,通过生成或对抗学习挖掘出文本集的主题信息[8-9]. 由于词嵌入维度较高,所以神经网络结构主题模型会显著提高模型参数的复杂度,增加模型的计算成本.

近期的一些研究表明:常见词嵌入模型的词嵌入语义和文本主题一致性评价存在联系[10],通过对文本嵌入或词嵌入进行聚类分析能够挖掘出一致性较好的主题内容,但是这些研究还是使用常见的聚类算法来获取词嵌入的簇类信息,缺少根据词嵌入学习方式及语义特点来获取词嵌入的簇类信息的研究. 如Word2vec的SGNS(skip-gram with negative-sampling)模型采用负采样方法,使相关词嵌入相似性增强,目标词和负样本词嵌入相互远离;BERT(bidirectional encoder representations from transformers)等上下文内容词嵌入学习模型,通过多层注意力机制使重要相关词嵌入的相似性增强,其他词嵌入间的相似性减弱.

本文根据一些词嵌入模型通过预测相邻词的生成,使相关、相似语义词间的嵌入表示相似性较大,不相关词的嵌入表示相互远离,从而使隐含有相关、相似语义的词嵌入表示在特征空间中聚集在一起的特点,设计并实现了一种基于该特性的聚类算法(word embedding characteristics clustering,WECC),该算法充分利用词嵌入特性信息,能够较好地发现词嵌入聚类的簇中心,且能够适应大规模词嵌入聚类

收稿日期:2022-05-12
基金项目:江西省自然科学基金(20212BAB202016)和江西省教科基金(GJJ10091)资助项目.
作者简介:曹中华(1976—),男,江西鄱阳人,讲师,博士,主要从事文本挖掘和财政大数据处理的研究. E-mail:rjxy_czh@jxnu.edu.cn

任务,从而可实现挖掘大规模文本的主题信息. 实验表明: 在一些词嵌入模型的预训练词嵌入上,该算法能够挖掘到主题一致性和多样性更好的结果.

1　相关研究工作

聚类是常见的无监督学习方法,聚类分析首先需要提取数据特征,确定数据相似性度量方法,然后依据特征和距离度量方法将数据划分到不同簇中,使簇内数据相似性较大,簇间数据相似性较小. 词嵌入表示具有丰富的语义信息,也使得文本内容具有了形式多样的语义表征,通过对文本的各种表征信息进行无监督聚类,可以分析文本的各方面内容. 近期一些研究者提出了基于文本嵌入或词嵌入无监督聚类的主题挖掘算法,根据需要聚类的数据类别,这些研究主要可以分为以下 2 大类.

一类是通过文本嵌入聚类获得文本的簇中心嵌入,然后分析簇中心和词之间的关系,挖掘文本主题信息. D. Angelov [11] 首先使用 Doc2Vec 模型 [12] 学习文本和词的嵌入表示,然后通过文本嵌入聚类获得簇中心,并将其作为主题语义嵌入,最后计算词嵌入和主题嵌入的相似距离,挖掘文本主题信息. BERT-Topic [13] 首先使用 Sentence-BERT 框架 [14] 将句子或文本转换为密集的向量表示,然后将这些文本嵌入进行聚类操作后获得文本集的簇结果,最后使用类别 TF-IDF 方法提取每个簇的主题词. 但是这类模型方法在文本内容篇幅较大时所获得的文本嵌入语义并不能准确表示文本主要内容,从而会影响到主题词的提取.

另一类方法不需要获得文本嵌入,这类算法通过直接对词嵌入聚类来挖掘词嵌入的簇中心. S. Sia [15] 直接对词嵌入聚类获得簇中心,并将其作为主题嵌入,通过计算主题嵌入和词嵌入的相似性以及词在文档集中的相关权重选择主题词,从而挖掘文本主题结果. R. M. Guilherme 等 [16] 提出使用自组织映射对 Word2vec 方法生成的词嵌入进行聚类,从而挖掘文本集的主题内容. L. Laure 等 [17] 研究了具有上下文内容的词嵌入聚类,发现聚类结果能够自然地捕获词的多义性,主题内容能达到常见主题模型类似结果. 上述这些词嵌入聚类研究需要将所有的词类嵌入输入系统中,由于词嵌入维度较高,当需要聚类的词类数较大时,对机器的计算资源要求较高,所以常见算法在一般硬件环境下很难满足大规模词嵌入聚类任务需求. 因此一些词嵌入聚类算法会先对预训练词嵌入降维,然后使用 K-Means (KM)、高斯混合模型(GMM)等算法挖掘词嵌入的簇中心. 词嵌入降维会造成词语义丢失,使降维的数据特征不一定适合数据聚类目标,而且一些研究实验表明词嵌入降维对提高主题结果影响有限 [15,18].

针对近期各类词嵌入聚类挖掘文本主题研究,Zhang Zihan 等 [19] 比较了多种词嵌入聚类文本主题挖掘算法,并提出了一种新的主题词权重计算方法; Yu Meng 等 [13] 认为 BERT 的 MLM(masked language model)预训练学习目标的上下文词嵌入可看作是由与文本集词典大小的高斯模型混合生成的,提出文本嵌入和主题嵌入联合学习算法,挖掘文本集内的主题.

数据的特征表示会极大地影响聚类性能,词嵌入作为词语义的特征表示,含有词嵌入聚类的有关信息内容,上述这些工作都忽略了词嵌入聚类可以深入理解词嵌入模型学习机制,依据词嵌入数据的特征表示设计聚类算法,从而更好地挖掘词嵌入的簇类结果.

2　本文工作

2.1　词嵌入相关及相似性

设文本集 D 的词典集合是 W,词 $w_i \in W$ 为在文本内句子中某个词,w_i 的上下文内容词集为 $h(w_i)$,上下文内容词集可以像 BERT 的 MLM 模型一样是句子内其他未缺失词,或是像 SGNS 模型一样是在中心词 w_i 前后固定窗口范围内的词. 语言模型学习目标是最大化对数似然函数

$$L = \lg p(w_i \mid h(w_i)). \tag{1}$$

令 $\boldsymbol{v}_i \in \mathbf{R}^k$ 表示词 w_i 的嵌入表示,$\boldsymbol{v}_h \in \mathbf{R}^k$ 是在上下文内容 $h(w_i)$ 中有关词经过神经网络函数后的嵌入表示,如在 SGNS 模型中 $\boldsymbol{v}_h$ 表示上下文内容中的某个词嵌入,在 MLM 模型中 $\boldsymbol{v}_h$ 为上下文内所有词嵌入的加权表示,k 表示词嵌入维度,$k \ll |W|$,$pw_i \mid h(w_i))$ 使用词嵌入方式

$$p(w_i \mid h(w_i)) = \exp(\boldsymbol{v}_h^{\mathrm{T}}\boldsymbol{v}_i) / \left(\sum_{w_j \in W} \boldsymbol{v}_h^{\mathrm{T}}\boldsymbol{v}_j\right). \tag{2}$$

Li Bohan 等 [20] 认为词和上下文嵌入的点积可以表示为

$$\boldsymbol{v}_h^{\mathrm{T}}\boldsymbol{v}_i = \mathrm{pmi}(h(w_i), w_i) + \lg p(w_i) + \lambda_h, \tag{3}$$

其中 $\mathrm{pmi}(h(w_i), w_i)$ 表示词 w_i 和上下文 $h(w_i)$ 之间的点互信息,点互信息能够评价它们之间共现的重要性,$\lg p(w_i)$、λ_h 分别是中心词和上下文内容有关常量.

式(3)表明词嵌入学习会使相关的词和上下文内容嵌入相似性较大,此外一些模型(如 SGNS、BERT)还采用负采样或注意力机制等方法强化它们之间的相似性. 根据分布假设,若词 w_i、w_j 具有相同且重要的上下文内容,则词 w_i、w_j 的嵌入表示相似性也会较大. 在词嵌入训练后,这些具有相关和相似关系的词嵌入表示会在特征空间中聚集在一起,但是由于词嵌入维度较高,所以这种空间聚集关系不是很明显.

词嵌入聚类的目标是发现在空间中相似节点的聚集中心. 设词 w_i 属于簇类 c_j,则 $\boldsymbol{v}_i$ 在特征空间与簇中心嵌入 $\boldsymbol{v}_{c_j}(1 \leqslant j \leqslant l)$ 应该最相似,l 表示簇数. 由于词 w_i 的相关、相似语义词嵌入会聚集在一起,所以为增强簇中心 c_j 与词 w_i 的相似性,与语言模型训练词 w_i 的嵌入表示类似,构造簇中心 c_j 和词 w_i 及词的上下文 $h(w_i)$ 的相似关系,并且使簇中心 c_j 和其他负样本词相互远离,从而得到优化的簇中心嵌入.

图 1 描绘了 WECC 算法工作情况. 设实心圆表示当前中心词,空心实线圆表示它的相邻词,空心虚线圆表示负样本词,星型为当前中心词的最相似簇中心. 在训练过程中,中心词和相邻词会将簇中心拉向自己,而负样本词会将该簇中心推离自己,由此簇中心将到达词嵌入空间中的合适位置.

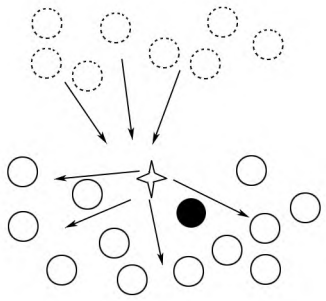

图 1 WECC 算法工作表示图

2.2 词嵌入特性聚类

设文本集的预训练词嵌入为$\{\boldsymbol{v}_1,\boldsymbol{v}_2,\cdots,\boldsymbol{v}_{|W|}\}$、簇类中心集合 $C=\{c_1,c_2,\cdots,c_l\}$、簇中心嵌入表示为$\{\boldsymbol{v}_{c_1},\boldsymbol{v}_{c_2},\cdots,\boldsymbol{v}_{c_l}\}$,$\boldsymbol{v}_{c_j}\in\mathbf{R}^k$,则词 w_i 属于第 j 个簇的概率为

$$p(c_j\mid w_i)=\exp(\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{v}_{c_j})/\Big(\sum_{1\leqslant n\leqslant l}\exp(\boldsymbol{v}_i^{\mathrm{T}}\boldsymbol{v}_{c_n})\Big), \quad (4)$$

词 w_i 概率值最大的簇类号 m 为

$$m=\operatorname{argmax}(p(c_1\mid w_i),p(c_2\mid w_i),\cdots,p(c_l\mid w_i)). \quad (5)$$

若将最大的簇类号 m 作为词 w_i 的所属簇,则由前述分析,词 w_i 相关、相似的其他词嵌入表示应该聚集在簇类 m 附近,从而词嵌入簇类优化目标是使簇中心 c_m 与 w_i、$h(w_i)$ 的相似性增加,因此模型目标是最小化对数似然函数

$$L=\Big(\lg p(c_m\mid w_i)+\sum_{w_o\in h(w_i)}\lg p(c_m\mid w_o)\Big)/(1+|h(w_i)|). \quad (6)$$

直接优化上述目标,会面临 Softmax 函数计算量较大和模型坍塌问题,因此本文采用和 SGNS 模型类似的优化方法,根据词 w_i 采样簇中心 c_m 的负样本词,使簇中心与当前词嵌入及上下文嵌入相近,并且使簇中心远离负样本词,该对数似然函数最小化目标为

$$L=\sum_{w_o\in w_i\bigcup h(w_i)}\lg\sigma(\boldsymbol{v}_0^{\mathrm{T}}\boldsymbol{v}_{c_m})+\sum_{w_n\in \mathrm{NEG}(w_i)}\lg\sigma(-\boldsymbol{v}_n^{\mathrm{T}}\boldsymbol{v}_{c_m}), \quad (7)$$

其中 σ 为 Sigmoid 函数,$\mathrm{NEG}(w_i)$ 为负样本词集,负样本词采样算法和 SGNS 模型负样本采样方法类似. 模型只需要优化簇中心嵌入,利用随机梯度下降对其优化求解,其梯度计算公式为

$$\partial L/\partial\boldsymbol{v}_{c_m}=\partial\Big(\sum_{w_o\in w_i\cup h(w_i)}\log\sigma(\boldsymbol{v}_0^{\mathrm{T}}\boldsymbol{v}_{c_m})+\sum_{w_n\in \mathrm{NEG}(w_i)}\log\sigma(-\boldsymbol{v}_n^{\mathrm{T}}\boldsymbol{v}_{c_m})\Big)/\partial\boldsymbol{v}_{c_m}=\sum_{w_o\in w_i\cup h(w_i)}(1-\sigma(\boldsymbol{v}_0^{\mathrm{T}}\boldsymbol{v}_{c_m}))\boldsymbol{v}_0+\sum_{w_n\in \mathrm{NEG}(w_i)}\sigma(\boldsymbol{v}_n^{\mathrm{T}}\boldsymbol{v}_{c_m})\boldsymbol{v}_n. \quad (8)$$

上述词嵌入聚类算法描述如算法 1 所示.

算法 1 词嵌入聚类训练算法.

输入: 预训练词嵌入、簇类数、训练文本.

输出: 簇中心嵌入.

(i) 文本数据预处理;
(ii) 提取文本内词间上下文关系数据;
(iii) 初始化簇中心嵌入,导入预训练词向量;
(iv) While not 收敛或迭代次数超过阈值;
(v) 训练数据混淆;
(vi) for 所有批数据;
(vii) 计算每个输入数据最大簇类号;
(viii) 采样该批数据的负样本;
(ix) 参照式(7)更新簇中心嵌入;
(x) 更新收敛条件数据.
(xi) 输出所有簇中心嵌入.

3 实验与分析

3.1 实验数据

本文实验选择用 3 种经典的文本数据集,它们分别是 20 个新闻组(20Newsgroups)、路透社新闻数

据(Reuters 21578)、搜狗新闻(Sogou).20Newsgroups 数据集由 20 个不同类别的新闻组数据组成,总共有 18 846 篇文本,每类数据大小几乎相同,实验使用 Sklearn 工具库获取该数据集,使用 NLTK 库去除了常见停用词、数字、特殊符号和低频词等,仅保留名词、动词、形容词,经过预处理后 20Newgroup 词典含有约 1.50 万个单词. 路透社数据集通过 NLTK 库获取,NLTK 库使用该数据的 ModApte 子集,总共包括 10 788 篇文本,该英文数据集也采用 20Newsgroups 相似的预处理方法,在经过预处理后,数据集词典含有约 0.58 万个单词. 搜狗新闻使用 2008 年 1 个月的精简版数据,该语料文本包含 15 种不同类别的新闻信息,如军事、商业、健康、体育运动、奥运等. 实验从每个类别中随机取 5 000 篇文本形成文本集,使用前先将其转码为简体中文编码,然后使用结巴分词工具去除非中文字符、停用词和低频词,同样也仅保留名词、动词、形容词,经过处理后,训练数据集词典含有约 3.50 万个单词. 本文算法在计算词嵌入聚类时,需要获得词间相邻关系,为减少噪声对簇结果的影响,在原始数据集上,实验只获取在词典内默认窗口大小等于 5 的词间相邻关系.

3.2　实验设置

实验主要通过主题的一致性(topic coherence, TC) 和主题词的多样性(topic diversity,TD) 2 个部分内容评估主题质量. 主题的一致性使用归一化的点互信息(normalized pointwise mutual information, NPMI) 评价主题词之间的共现紧密程度,主题一致性值越大表示主题词内容更容易被人理解[10];主题多样性[9] 衡量在所有主题中单词的唯一性,多样性值越大表示主题内容涵盖面更广泛. 基准模型使用 S. Sia 等[15] 提出的词嵌入聚类主题挖掘算法、M. Grootendorst[13] 提出的先文本嵌入聚类,然后寻找簇内主题词的方法以及其他概率模型、神经网络模型方法,这与本文直接词嵌入聚类技术线路不同,所以未列入比较.

实验使用 SGNS、GloVe、BERT 等 3 种主流模型的预训练词嵌入来比较不同模型的词嵌入使用各种聚类算法的主题挖掘质量. 在词嵌入聚类前,先用相关工具获得文本集的词嵌入,SGNS、GloVe 是使用较广的静态词嵌入模型,SGNS 模型的词嵌入使用 Gensim 工具训练得到,GloVe 词嵌入使用 Pennington 提供的程序训练得到,它们的词嵌入维度 $k = 300$,其他模型参数都使用默认值. BERT 是具有多层结构的动态词嵌入模型,模型会根据上下文内容动态产生不同的词嵌入,实验使用它的最后层输出作为词嵌入,并计算每类词嵌入的平均值作为聚类输入数据. 英文 BERT 词嵌入使用 Transformers 的 bert-base-uncased 预训练模型库获得,中文使用 WoBERT Plus 模型[21] 获得动态词嵌入,BERT 模型词嵌入维度 $k = 768$,本文簇类中心维度和当前使用的词嵌入维度大小一样.

由于词嵌入只含有词语义信息,而文本主题需要挖掘文本集的总体语义信息,因此只依据聚类的簇中心很难全面地反映词在文本集的重要性. 算法没有采用式(5) 将每个词分配给概率最大的簇,根据簇内词的概率值大小选择它的代表词,而是依据簇中心嵌入和词嵌入的余弦相似性大小获得每个主题的初始主题代表词,这样可以和 LDA 模型一样,每个主题都含有所有词的生成概率. 每个主题的初始代表词将按照词的文本权重进行调整,选择前 10 个词作为主题代表词,相关权重设计方法为

$$\mathrm{tf}(w_i) = n_i / \sum_j n_j, \tag{9}$$

$$\mathrm{tfdf}(w_i) = \mathrm{tf}(w_i) \mid \{d \in D \mid w_i \in d\} \mid / \mid D \mid, \tag{10}$$

$$\mathrm{tfidf}(w_i) = \mathrm{tf}(w_i) \lg(\mid D \mid / (\mid \{d \in D \mid w_i \in d\} \mid + 1)), \tag{11}$$

其中 n_i 为词 w_i 在数据集 D 中的总词频数,d 为数据集 D 的某篇文本. 若不依据词权重进行调整,则直接取最相似的 10 个词作为主题词,本文称其为原始(original) 方式.

3.3　实验结果与分析

实验获取了本文算法在多个不同簇类数下的主题评价结果,当用于文本主题挖掘时,簇类数即为主题数作为超参数预先设定. 和 S. Sia 等[15] 实验一样,在这 3 类数据集下,当主题数等于 20 时得到最优的 NPMI 值. 表 1 列出了当主题数等于 20 时各类算法的最优实验结果,其中 20Newsgroups 和 Reuters 数据集使用 S. Sia 等研究的最优结果,S. Sia 等未进行 Sogou 中文数据实验,在其工作基础上,本文补充了其在 Sogou 新闻数据集上的主题结果.

从表 1 实验结果可以看出: 本文算法在 SGNS 模型词嵌入的主题一致性结果上都优于其他所有情况. 在词的多样性方面,除 WECC 算法在 GloVe 词嵌入的情况外,其他算法的结果表现都较好; 对于同种预训练词嵌入,本文算法在 SGNS 词嵌入的主题一致性值上提高了 0.10; BERT 模型词嵌入在英文数据集下的主题一致性提高了 0.05,在中文数据集下的 BERT 词嵌入比一些算法的有所降低. 这是由

于中文分词会影响 BERT 模型的词嵌入表示. 对于 GloVe 词嵌入,本文算法得到的主题结果则不理想,这可能是由于在 SGNS、Bert 模型训练词嵌入时能够增强词嵌入的相关性,而 GloVe 词嵌入模型只是拟合词间全局相关性值,没有去调节词间的重要相关性. 本文的词嵌入聚类算法和 SGNS 模型学习词嵌入过程非常相似,更能够捕获该模型的词嵌入簇类结构信息.

表 1 各模型最优一致性和多样性值

Model	TC/TD		
	20Newsgroups	Reuters	Sogou
SGNS(KM)	0.190/0.937	0.003/0.949	0.282/0.853
SGNS(GMM)	0.200/0.852	0.008/0.986	0.163/0.835
SGNS(WECC)	0.303/0.955	0.211/0.845	0.338/0.980
GloVe(KM)	0.230/0.952	0.001/0.937	0.257/0.823
GloVe(GMM)	0.240/0.955	0.005/0.995	0.239/0.812
GloVe(WECC)	-0.870/0.395	-0.718/0.485	-0.131/0.310
Bert(KM)	0.250/0.952	0.120/0.973	0.181/0.940
Bert(GMM)	0.250/0.965	0.150/0.975	0.142/0.872
Bert(WECC)	0.300/0.985	0.186/0.960	0.169/0.960

表 2 ~ 表 4 分别为本文算法运行于 20Newsgroups、Reuters 、Sogou 数据和主题数为(20,50,80,100,120)的实验结果. 从每类数据的实验结果可以看出:当主题数增加时,主题一致性值都呈现递减趋势. 对于主题词的多样性,未调整的 original 结果都保持有较好的结果. 其他经过调整的结果也随着主题数的增加,主题词的多样性呈递减趋势.

在相同主题值时,通过词的文本权重调整后的主题一致值比直接采用 original 方式的主题一致性值都有所增加,而主题词的多样性则有所降低. 这说明词嵌入在特征空间上的分布间隔并不明显,使得一些簇间的边界并不清晰. 当主题词调整后能够成为主题词时,降低了主题词的多样性. tf、tfidf 和 tfdf 这 3 种词的文本权重方法通过调整聚类的初始主题词都能够提高主题的一致性. 从结果来看,很难判别出哪种方法更好.

综合比较这 3 类数据集结果,在数据预处理方法一样的情况下,随着数据规模的增加主题一致性也逐渐变大,这体现了数据量的增加词间的共现密集程度也增加、词嵌入的重要聚集关系表现更加明显. 同时可以看出:虽然各类词嵌入模型的理论分析上和词嵌入的最相似词计算结果等方面都比较接近,但不同模型的词嵌入还是具有各自特点,它们的词嵌入特征在空间上的分布并不一样,这需要后续持续研究.

表 2 20Newsgroups 一致性和多样性值

Topic number	model	TC/TD			
		original	tfidf	tf	tfdf
20	SGNS	0.198/1.000	0.282/0.960	0.303/0.955	0.299/0.965
	GloVe	-0.949/0.380	-0.950/0.400	-0.884/0.420	-0.870/0.395
	Bert	-0.124/0.995	0.261/0.985	0.290/0.975	0.300/0.985
50	SGNS	0.127/0.992	0.258/0.816	0.277/0.808	0.269/0.814
	GloVe	-0.961/0.414	-0.932/0.388	-0.832/0.432	-0.832/0.448
	Bert	-0.177/0.948	0.132/0.932	0.212/0.934	0.212/0.938
80	SGNS	0.101/0.998	0.225/0.738	0.241/0.739	0.238/0.746
	GloVe	-0.952/0.348	-0.934/0.367	-0.877/0.415	-0.881/0.420
	Bert	-0.232/0.891	0.112/0.832	0.179/0.835	0.154/0.840
100	SGNS	0.068/0.994	0.218/0.673	0.243/0.670	0.237/0.688
	GloVe	-0.955/0.351	-0.935/0.316	-0.848/0.370	-0.858/0.373
	Bert	-0.232/0.868	0.088/0.775	0.155/0.799	0.161/0.798
120	SGNS	0.098/0.984	0.227/0.643	0.248/0.632	0.239/0.640
	GloVe	-0.948/0.330	-0.920/0.300	-0.823/0.349	-0.836/0.356
	Bert	-0.287/0.814	-0.004/0.754	0.093/0.770	0.087/0.775

表 3　Reuters 一致性和多样性值

Topic number	model	TC/TD			
		original	tfidf	tf	tfdf
20	SGNS	0.178/0.995	0.204/0.850	0.194/0.820	0.211/0.845
	GloVe	-0.922/0.520	-0.845/0.530	-0.742/0.505	-0.718/0.485
	Bert	-0.176/0.975	0.145/0.960	0.174/0.970	0.186/0.960
50	SGNS	0.137/0.986	0.206/0.678	0.205/0.664	0.206/0.680
	GloVe	-0.936/0.340	-0.799/0.392	-0.722/0.366	-0.715/0.364
	Bert	-0.121/0.938	0.187/0.758	0.202/0.760	0.209/0.772
80	SGNS	0.066/0.966	0.182/0.545	0.189/0.547	0.190/0.552
	GloVe	-0.924/0.323	-0.756/0.365	-0.690/0.345	-0.672/0.335
	Bert	-0.184/0.852	0.143/0.658	0.167/0.655	0.172/0.658
100	SGNS	0.060/0.958	0.179/0.462	0.186/0.471	0.188/0.479
	GloVe	-0.915/0.285	-0.790/0.314	-0.712/0.291	-0.699/0.281
	Bert	-0.216/0.819	0.117/0.600	0.153/0.585	0.148/0.597
120	SGNS	0.031/0.926	0.168/0.415	0.182/0.415	0.188/0.422
	GloVe	-0.932/0.262	-0.803/0.307	-0.739/0.295	-0.727/0.276
	Bert	-0.207/0.756	0.032/0.530	0.070/0.526	0.085/0.525

表 4　Sogou 一致性和多样性值

Topic number	model	TC/TD			
		original	tfidf	tf	tfdf
20	SGNS	0.220/1.000	0.338/0.980	0.316/0.980	0.326/0.980
	GloVe	-0.764/0.310	-0.164/0.290	-0.151/0.310	-0.131/0.310
	Bert	-0.605/0.960	0.169/0.960	0.167/0.960	0.159/0.960
50	SGNS	0.185/0.998	0.292/0.914	0.289/0.912	0.286/0.912
	GloVe	-0.806/0.248	-0.327/0.204	-0.331/0.220	-0.267/0.196
	Bert	-0.654/0.910	0.180/0.966	0.182/0.968	0.183/0.968
80	SGNS	0.185/0.996	0.294/0.847	0.285/0.851	0.283/0.862
	GloVe	-0.802/0.162	-0.323/0.165	-0.290/0.185	-0.252/0.173
	Bert	-0.643/0.907	0.183/0.961	0.192/0.962	0.201/0.965
100	SGNS	0.152/0.989	0.277/0.796	0.270/0.803	0.277/0.806
	GloVe	-0.776/0.119	-0.219/0.129	-0.198/0.132	-0.159/0.127
	Bert	-0.615/0.905	0.221/0.946	0.225/0.951	0.225/0.951
120	SGNS	0.151/0.981	0.277/0.750	0.274/0.751	0.272/0.757
	GloVe	-0.791/0.109	-0.270/0.118	-0.230/0.127	-0.201/0.114
	Bert	-0.641/0.905	0.208/0.937	0.216/0.939	0.224/0.941

4　结束语

本文设计并实现了一种基于词嵌入特性的聚类算法,该算法通过增强中心词的簇中心与相邻词的相似性,并且使其远离负样本词,可以从 SGNS、BERT 模型词嵌入中学习到更好的词嵌入簇类结构信息,并将其应用于文本主题挖掘. 实验表明本文词嵌入聚类算法挖掘到的主题词能够显著提高主题一致性值,并且算法采用批量文本训练方式,能够适用于大规模词嵌入聚类任务需求.

5　参考文献

[1] BLEI D M,NG A Y,JORDAN M I. Latent Dirichlet allocation [J]. Journal of Machine Learning Research,2003,3(4/5):993-1022.

[2] TOMAS M,ILYA S,CHEN Kai,et al. Distributed representations of words and phrases and their compositionality [EB/OL]. [2013-10-16]. https://arxiv.org/abs/1310.4546.

[3] PENNINGTON J,SOCHER R,MANNING C D,et al. Glove: global vectors for word representation [EB/OL]. [2014-10-01]. https://aclanthology.org/D14-1162/.

[4] JACOB D,CHANG Mingwei,KENTON L,et al. Bert: pre-training of deep bidirectional transformers for language understanding [EB/OL]. [2018-10-11]. https://arxiv.org/abs/1810.04805.

[5] 黄佳佳,李鹏伟,彭敏,等. 基于深度学习的主题模型研究 [J]. 计算机学报,2020,43(5):827-855.

[6] DAS R,ZAHEER M,DYER C,et al. Gaussian LDA for topic models with word embeddings [EB/OL]. [2015-07-19]. https://aclanthology.org/P15-1077.

[7] ADJI B D,FRANCISCO J R,DAVID M B. Topic modeling in embedding spaces [EB/OL]. [2019-07-08]. https://arxiv.org/abs/1907.04907.

[8] MIAO Yishu,YU Lei,PHIL B,et al. Neural variational inference for text processing [EB/OL]. [2015-11-19]. https://arxiv.org/abs/1511.06038.

[9] FENG Nan,RAN Ding,RAMESH N,et al. Topic modeling with wasserstein autoencoders [EB/OL]. [2019-07-25]. https://aclanthology.org/P19-1640.

[10] 夏家莉,曹中华,彭文忠,等. Skip-Gram 结构和词嵌入特性的文本主题建模 [J]. 小型微型计算机系统,2020,41(7):1400-1405.

[11] ANGELOV D. Top2vec: distributed representations of topics [EB/OL]. [2020-08-20]. https://arxiv.org/abs/2008.09470.

[12] QUOC V L,TOMAS M. Distributed representations of sentences and documents [EB/OL]. [2014-05-17]. https://arxiv.org/abs/1405.4053.

[13] GROOTENDORST M. BERTopic: neural topic modeling with a class-based TF-IDF procedure [EB/OL]. [2022-03-11]. https://arxiv.org/abs/2203.05794.

[14] NILS R,IRYNA G. Sentence-bert: sentence embeddings using siamese bert-networks [EB/OL]. [2019-08-27]. https://aclanthology.org/D19-1410.

[15] SIA S,AYUSH D,SABRINA J M. Tired of topic models? clusters of pretrained word embeddings make for fast and good topics too! [EB/OL]. [2020-04-30]. https://aclanthology.org/2020.emnlp-main.135.

[16] GUILHERME R M,RODRIGO P,LEANDRO N C. Detecting topics in documents by clustering word vectors [EB/OL]. [2019-06-22]. https://link.springer.com/chapter/10.1007/978-3-030-23887-2_27.

[17] LAURE T,DAVID M. Topic modeling with contextualized word representation clusters [EB/OL]. [2020-10-24]. https://arxiv.org/abs/2010.12626.

[18] YU Meng,Yunyi ZHANG,HUANG Jiaxin,et al. Topic discovery via latent space clustering of pretrained language model representations [EB/OL]. [2022-02-09]. https://arxiv.org/abs/2202.04582.

[19] ZHANG Zihan,FANG Meng,CHEN Ling,et al. Is Neural Topic Modelling better than Clustering? An empirical study on clustering with contextual embeddings for Topics [EB/OL]. [2022-04-21]. https://aclanthology.org/2022.naacl-main.285.

[20] LI Bohan,ZHOU Hao,HE Junxian,et al. On the sentence embeddings from pre-trained language models [EB/OL]. [2020-11-02]. https://aclanthology.org/2020.emnlp-main.733.

[21] 苏剑林. 提速不掉点: 基于词颗粒度的中文 WoBERT [EB/OL]. [2020-09-18]. https://www.spaces.ac.cn/archives/7758.

The Topic Mining Based on Word Embedding Characteristics Clustering

CAO Zhonghua[1],HUANG Xin[1],PENG Wenzhong[2],LIU Yuanchun[1]

(1. School of Software,Jiangxi Normal University,Nanchang Jiangxi 330022,China;
2. School of Information Management,Jiangxi University of Finance and Economics,Nanchang Jiangxi 330032,China)

Abstract: Data clustering is a common unsupervised learning method,which can be used to mine topics through word embedding clustering. However,most researchers used conventional clustering algorithms to mine the cluster of word embedding. There is still a lack of research on the design of clustering algorithm based on word embedding characteristics to mine text topics. In the paper,a word embedding clustering algorithm is designed based on the language model,which gathers the embedding representations of related and similar words by learning the relevant information. The algorithm first calculates the cluster number of the central word,and then enhances the similarity between the cluster central and the related words,at the same time keeps it away from the negative sample word. Therefore,it can learn the word embedding cluster structure of the text set,and be used to mine text topics. Experiments on three public datasets show that the algorithm can mine topic results with better coherence and diversity in some models of word embedding.

Key words: word embedding; clustering; language model; text topics

(责任编辑: 冉小晓)