International Journal of Machine Learning and Cybernetics
https://doi.org/10.1007/s13042-020-01133-3

ORIGINAL ARTICLE

Topic discovery by spectral decomposition and clustering with coordinated global and local contexts

Jian Wang[1] · Kejing He[1] · Min Yang[2]

Received: 29 June 2019 / Accepted: 15 April 2020

Abstract

Topic modeling is an active research field due to its broad applications such as information retrieval, opinion extraction and authorship identification. It aims to discover topic structures from a collection of documents. Significant progress have been made by the latent dirichlet allocation (LDA) and its variants. However, the "bag-of-words" assumption is usually made for the whole document by conventional methods, which ignores the semantics of local context that play crucial roles in topic modeling and document understanding. In this paper, we propose a novel coordinated embedding topic model (CETM), which incorporates spectral decomposition and clustering technique by leveraging both global and local context information to discover topics. In particular, CETM learns coordinated embeddings by using spectral decomposition, capturing the word semantic relations effectively. To infer the topic distribution, we employ a clustering algorithm to capture semantic centroids of coordinated embeddings and derive a fast algorithm to obtain the topic structures. We conduct extensive experiments on three real-world datasets to evaluate the effectiveness of CETM. Quantitatively, compared to state-of-the-art topic modeling approaches, CETM achieves significantly better performance in terms of topic coherence and text classification. Qualitatively, CETM is able to learn more coherent topics and more accurate word distributions for each topic.

Keywords Topic modeling · Spectral decomposition · Clustering · Global context · Local context

1 Introduction

With the growing of large collection of electronic texts, much attention has been given to topic modeling of textual corpora, which is designed to identify representations of data and learn thematic structure from large document collections without human supervision. Conventional topic models, such as Probabilistic Latent Semantic Analysis (PLSA) [15] and Latent Dirichlet Allocation (LDA) [4], can be viewed as graphical models with latent variables. Some non-parametric extensions to LDA have been successfully applied to characterize the contents of documents [31, 33]. However, the inference of those non-parametric models are computationally hard, such that inaccurate or slow approximations are resorted to calculate the posterior distributions over the topics. New undirected graphical model approaches, including the replicated softmax model [14], are also successfully used to explore the topics of documents, and in particular cases they outperform LDA [30].

A major limitation of these topic modeling approaches and many of their extensions is the "bag-of-words" assumption, which assumes that each document is characterized by the "bag-of-words" features. This assumption is favorable in the computational point of view, but ignores the word order and cannot capture the semantic regularities of documents. For example, the sentences "the chair department offers couches" and "the department chair couches offers" have the same unigram features, while they represent different meanings and topics. When deciding the word "chair" in the first sentence is generated by which topic, knowing that it is immediately preceded by the word "department" helps us to find that it is related to the university administration topic [35]. In addition, the conventional methods mainly use

✉ Kejing He
kejinghe@ieee.org

Jian Wang
cs_wangjian@mail.scut.edu.cn

Min Yang
min.yang@siat.ac.cn

[1] School of Computer Science and Engineering, South China University of Technology, Guangzhou, China

[2] Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences, Shenzhen, China

Published online: 16 May 2020

the global context information in document-level to discover topics. Such topics may not be meaningful or semantically coherent and even redundant sometimes.

In contrast, word embedding algorithms such as Neural Probabilistic Language Model (NPLM) [2] and Continuous Bag-of-Words (CBOW) [22], are based on local context information, and can capture the semantic and syntactic properties of words. They follow the distributional hypothesis [13] that the words occurring in the same context window tend to be semantically and syntactically similar. Thus, the semantically related words will appear closer to each other in the word embedding space. By leveraging the local context information, topic models can exploit informative topics more effectively. For example, Gaussian-LDA [7] performs topic inference with multivariate Gaussian distributions based on word embeddings pre-trained on external corpora such as Wikipedia. Latent Feature Topic Model (LFTM) [26] extends topic models by incorporating latent word embeddings trained on a large corpus to improve topic discovery on a smaller corpus. Despite the effectiveness of these models, there is still limitation of word embedding based topic modeling. The semantic regularities in different domains are usually distinct. For example, using the word embeddings learned from a general domain (e.g., Wikipedia) may deteriorate the performance of topic modeling in another specific domain (e.g., NIPS papers collection). The recently proposed Collaborative Language Model (CLM) [37] combines topic discovery and word embedding as collaborative tasks, which takes both global and local context information into consideration based on matrix factorization. However, the computation process is quite time consuming.

In this study, we propose a novel Coordinated Embedding Topic Model (CETM), which takes the aforementioned limitations into consideration. Different from the conventional probabilistic topic models (e.g., LDA) or matrix factorization based approaches (e.g., CLM), our model incorporates spectral decomposition and clustering technique by leveraging global context and local context information. Concretely, CETM leverages global context by exploiting document-level word coherence, and leverages local context based on the words that co-occur in a fixed context window. Then, spectral decomposition [1] is used to learn coordinated embeddings by coordinating global and local context information. Therefore, the learned latent topics are referred to as the semantic centroids (i.e., the most representative embeddings of clusters) when performing clustering on coordinated embeddings.

The main contributions of this work can be summarized as follows:

- We propose CETM, a novel topic model based on spectral decomposition and clustering, which understands semantics of documents effectively and discovers topic structures efficiently by leveraging the benefits of coordinated information.
- We derive a simple and fast mechanism to inference topic distribution. CETM requires lower computational cost and less hyper-parameters compared with other embedding-based topic models (e.g., Gaussian-LDA, CLM), since it almost only needs to tune the parameter controlling the weights of global context and local context.
- We conduct extensive experiments to justify the effectiveness of CETM in three widely-used text datasets. The experimental results indicate that CETM significantly outperforms the compared topic models from both quantitative and qualitative perspectives.

The rest of the paper is organized as follows. Section 2 provides a brief review of related works. In Sect. 3, we describe details of the proposed Coordinated Embedding Topic Model. Experimental results and qualitative analysis are shown in Sect. 4. Section 5 concludes our work.

2 Related work

In this section, we primarily review the related work in topic modeling and spectral decomposition algorithm.

2.1 Topic modeling

In the past few years, numerous topic modeling approaches have been proposed, which automatically discovered meaningful topics from documents [3, 4, 15]. Topic models are generally unsupervised methods to discover latent semantic structures from a corpus, and they are powerful for document analysis and information extraction. Latent Semantic Analysis (LSA) [8] was proposed to reduce dimensions of documents by projecting the document-word matrix into a lower dimensional space using Singular Value Decomposition (SVD). PLSA [15] extended LSA from probability perspective by introducing latent variables between documents and words, which could be viewed as topics. LDA [4] introduced Dirichlet priors at the document level. Each document is represented as a multinomial distribution over topics and each topic is represented as a multinomial distribution over words. Correlated Topic Model (CTM) [3] replaced Dirichlet priors of LDA with Logistic Normal priors, which modeled topic correlation better. These conventional topic models can be viewed as probabilistic topic modeling methods, and they are widely applied in real-life applications such as document classification and clustering [5], sentiment analysis [21], etc.

Non-negative Matrix Factorization (NMF) [17] is another major topic modeling approach. Ding et al. [11] proved that NMF is equal to PLSA as their objective functions to optimize are the same, and the non-negativity makes the matrix decomposition results easy to explain. L-EnsNMF [32] leveraged the idea of gradient boosting on the residual matrix of NMF and applied a local weighting scheme to discover high-quality local topics.

Word embedding, also known as distributed representations of words, is powerful to capture semantic regularities of documents by learning the information from local word co-occurrence contexts [23]. Continuous Bag of Words (CBOW) and Skip-Gram model proposed by Mikolov et al. [22] are two widely-used word embedding algorithms. Levy et al. [18] proved the equivalence between Skip-Gram with negative sampling and factorizing the Shifted Positive Pointwise Mutual Information (SPPMI) matrix of local word co-occurrence contexts.

Inspired by the recent success of word embedding in natural language processing, several recent topic modeling methods incorporated word embeddings to improve topic discovery. For example, Gaussian-LDA [7] leveraged word embeddings pre-trained on Wikipedia and modeled topics with multivariate Gaussian distributions on the embedding space. Latent Feature Topic Model (LFTM) [26] incorporated word embeddings trained on large corpora as latent features to improve topic discovery on a smaller corpus. TopicVec [20] generated topic embeddings by adding an embedding link function to model topic-word distribution. Collaborative Language Model (CLM) [37] modeled topics and word embeddings collaboratively by exploiting complementary global and local contexts with non-negative matrix factorization. In addition, some deep learning based methods further improve topic modeling using neural networks. For example, as a neural autoregressive topic model, iDocNADEe [12] incorporated word embeddings as a distributional prior. TMSA [19] proposed to unify both topic modeling and word embedding by the construction of a mutual learning mechanism, which simultaneously improve the quality of topic discovery and word embedding.

2.2 Spectral decomposition

Dimensionality reduction is a traditional task to find low-dimensional representations for high-dimensional data. It generally includes linear methods such as Principle Component Analysis (PCA) [28], and nonlinear methods such as Locally Linear Embedding (LLE) [27]. Spectral decomposition is a nonlinear method for dimensionality reduction by specially constructing weighted graph and using eigenvectors as low dimensional representations [34].

Representative algorithms of spectral decomposition include LLE [27], Isomap [34] and Laplacian eigenmaps [1]. Specifically, Laplacian eigenmaps algorithm is effective and widely-used. It reflects the intrinsic geometric structure of the manifold of the data. The Laplacian graph obtained from data is defined by the Laplacian-Beltrami operator, and the embedding maps for data are eigenmaps on the Laplacian. The locality preserving property of Laplacian eigenmaps algorithm makes it insensitive to noise and easy to be scalable with clustering.

Recently, spectral decomposition algorithms have gained increasing attention in word embedding. Dhillon et al. [9] proposed eigenwords algorithm for word embedding, which captured the meaning of words from their contexts by canonical correlation analysis and learned word embeddings pretty fast. Soleimani et al. [29] proposed Spectral Word Embedding with Negative Sampling (SENS) model, which provided a new view that the use of negative samples can improve the quality of spectral word embeddings. Overall, these algorithms have shown superior performance for learning word embeddings.

Nevertheless, to the best of our knowledge, none of existing methods have explored the connection between spectral decomposition and topic modeling. In this paper, we apply spectral decomposition into topic modeling by leveraging both global context and local context information, and thus derive a fast algorithm for topic discovery.

3 Coordinated embedding topic model

In this section, we first introduce notations and definitions mentioned in this paper. Second, we elaborate on Coordinated Embedding Topic Model (CETM) and give the optimization solutions.

3.1 Notations and definitions

Given a corpus D with N documents and the vocabulary size is V, topic models aim to discover topic distributions over documents and learn topic distributions with words. Generally, topic models find a lower-rank approximation given by

$$D \approx \Theta T^{\mathrm{T}}, \tag{1}$$

where $T \in \mathbb{R}_{+}^{V\times K}$ and $\Theta \in \mathbb{R}_{+}^{N\times K}$ are both non-negative factors and K is the number of topics.

The main notations used in this paper are summarized in Table 1 for clarity. In CETM, the document-word matrix D is calculated by tf-idf weights instead of raw frequency. The global context information is encoded in the global word co-occurrence matrix W^g and the local context information is encoded in the local word co-occurrence matrix W^l.

3.2 Coordinated embedding inference

CETM learns the coordinated embedding by leveraging global context and local context information, which is an essential step for further topic distribution inference.

Table 1 Table of main notations

Notation	Description
N	Number of documents
V	Vocabulary size
K	Number of topics
M	Dimensionality of embedding
U	Number of top words per topic
$D \in \mathbb{R}_{+}^{N \times V}$	Document-word matrix
$W^g \in \mathbb{R}_{+}^{V \times V}$	Global word co-occurrence matrix
$W^l \in \mathbb{R}_{+}^{V \times V}$	Local word co-occurrence matrix
$W \in \mathbb{R}_{+}^{V \times V}$	Adjacency matrix
$L \in \mathbb{R}^{V \times V}$	Laplacian matrix
$Y \in \mathbb{R}^{V \times M}$	Coordinated embedding matrix
$S \in \mathbb{R}^{K \times M}$	Topic embedding matrix
$T \in \mathbb{R}_{+}^{V \times K}$	Topic-word distribution matrix
$\Theta \in \mathbb{R}_{+}^{V \times K}$	Document-topic distribution matrix
$d_{ij}, w_{ij}, w^g_{ij}, w^l_{ij}$	The ij^{th} entry in matrix D, W, W^g, W^l respectively

3.2.1 Utilization of global context

As shown in Fig. 1a, given a collection of documents, the global context information refers to the document-level word co-occurrence. For example, the words “share” and “bonus” have a global context value of 213.1 while the words “propose” and “bonus” have a global context value of 30.9, indicating that the words “share” and “bonus” co-occur in more documents. Formally, the global word co-occurrence matrix W^g is constructed as below.

Given the document-word matrix D with tf-idf weights, the indicator that whether a word occurrs in a document can be easily defined as

$$\tilde{d}_{ij} = \begin{cases} 1, & \text{if } d_{ij} > 0 \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

To construct document-level word co-occurrence, $\tilde{w}^g_{ij}$ is defined as the dot-product of the indicator vectors of w_i and w_j, respectively.

$$\tilde{w}^g_{ij} = \tilde{d}_i^T \tilde{d}_j \tag{3}$$

Following similar idea of the tf-idf, a weight function on $\tilde{w}^g_{ij}$ is used in order to filter out high frequency patterns. By leveraging the inverse document frequency of w_i and w_j, w^g_{ij} is formed as

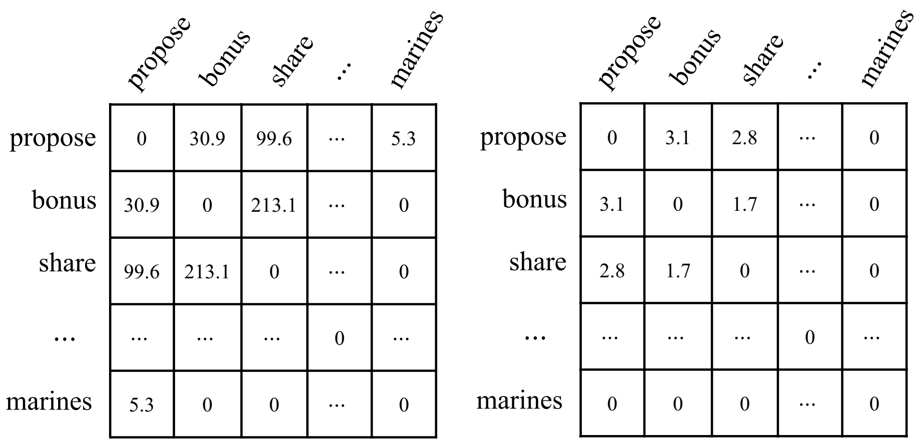
$$w^g_{ij} = (idf(w_i) idf(w_j)) \tilde{w}^g_{ij}, \tag{4}$$

where $idf(w_i) = \log \frac{N}{|\{j : w_i \in d_j\}| + 1}$.

3.2.2 Utilization of local context

As shown in Fig. 1b, the local context information refers to the word co-occurrence in a fixed text slice (context window), which mainly focus on capturing local statistical features of given documents. For example, the words “marines” and “propose” have a local context value of 0, indicating that the word “marines”does not occur in preceding or following position of the word “propose” within a fixed slice length. Concretely, the local word co-occurrence matrix W^l is constructed as below.

Since each text slice consists of a target word and its neighboring context words within a fixed-size window centered at the target word, the value of entry w^l_{ij} should reveal the total number of times that words w_i and w_j co-occurred in the context window. The matrix W^l contains statistics of the primary source of local context information, but not all co-occurrences are significant. Pointwise Mutual



	propose	bonus	share	...	marines
propose	0	30.9	99.6	···	5.3
bonus	30.9	0	213.1	···	0
share	99.6	213.1	0	···	0
...	···	···	···	0	···
marines	5.3	0	0	···	0

(a) Global context

	propose	bonus	share	...	marines
propose	0	3.1	2.8	···	0
bonus	3.1	0	1.7	···	0
share	2.8	1.7	0	···	0
...	···	···	···	0	···
marines	0	0	0	···	0

(b) Local context

Fig. 1 Illustration of the utilization of **a** global context and **b** local context on Reuters dataset

Information (PMI) [6] is effective to detect insignificant co-occurrences and find associations between words. Given a word-context pair (w, c), PMI is defined as

$$PMI(w,c) = \log \frac{P(w,c)}{P(w)P(c)}. \tag{5}$$

Empirically, PMI can be estimated by the number of co-occurrences of word-context pair (w, c):

$$PMI(w,c) = \log \frac{\#(w,c)\cdot Y}{\#(w)\cdot\#(c)} \tag{6}$$

where $\#(w,c)$ is the number of times that word w and c co-occurred, $\#(w)=\sum_c \#(w,c)$, $\#(c)=\sum_w \#(w,c)$, and Y is the total number of word-context pairs. However, the non-negativity of each entry in PMI matrix can not be guaranteed. Based on the study of Levy et al. [18], implicitly factorizing the Shifted Positive Pointwise Mutual Information (SPPMI) matrix is equivalent to Skip-Gram with negative sampling, and the shift parameter is also equivalent to the negative sampling value k. The SPPMI matrix is given by

$$SPPMI_k(w,c) = \max(PMI(w,c) - \log k, 0). \tag{7}$$

Thus, CETM regards the SPPMI matrix as the local word co-occurrence matrix W^l.

3.2.3 Coordination

CETM coordinates global context and local context by spectral decomposition. Suppose given data points $X=[x_1,x_2,\cdots,x_n]$, we construct a weighted graph $G=(V,E)$ with n nodes, and then put an edge between nodes i and j if x_i and x_j are connected. The connectivity between nodes is measured by the weights of edges, formulating adjacency matrix W and each entry w_{ij} denotes the value of weight. Naturally, W is non-negative and symmetric. Consider multiview spectral embedding [36], given the ith view of adjacency matrix $W^{(i)}$, we put a Laplacian operator as the mapping function on the graph G:

$$L^{(i)} = Diag^{(i)} - W^{(i)} \tag{8}$$

where $Diag^{(i)}$ is diagonal and $Diag^{(i)}_{jj} = \sum_l W^{(i)}_{jl}$; $L^{(i)}$ is called un-normalized graph Laplacian matrix. In order to further balance the structure of graph G, a normalized graph Laplacian matrix is adopted and given by

$$\begin{aligned} L^{(i)}_n &= (Diag^{(i)})^{-1/2} L^{(i)} (Diag^{(i)})^{1/2} \\ &= I - (Diag^{(i)})^{-1/2} W^{(i)} (Diag^{(i)})^{-1/2} \end{aligned} \tag{9}$$

where $L^{(i)}_n$ is positive semidefinite and symmetric, the proof can be found in [36].

In CETM, the global word co-occurrence matrix W^g and the local word co-occurrence W^l can be thought as different views of adjacency matrices of the constructed graph on document-word distribution. Thus, the global view and local view of word co-occurrence matrix (i.e., W^g and W^l) can be collaboratively used to construct a normalized graph Laplacian matrix for topic modeling:

$$L_n = (1-\lambda)L^{(g)}_n + \lambda L^{(l)}_n \tag{10}$$

where $L^{(g)}_n$ and $L^{(l)}_n$ are the normalized graph Laplacian matrix of W^g and W^l respectively, λ is the parameter controlling weights of the global context information and local context information.

With Laplacian transformation of document-word distribution finished, we wish to obtain the low-dimensional coordinated embedding, denoted as $Y=[y_1,y_2,\cdots,y_M]\in\mathbb{R}^{V\times M}$, where M is the dimensionality of embedding. The optimization objective is to minimize the normalized cut [1]:

$$\begin{aligned} &\arg\min \sum_{i,j} ||y_i - y_j||^2 w_{ij} = \arg\min_{Y} tr(YL_nY^T) \\ &\quad = \arg\min_{Y,\lambda}(1-\lambda)tr(YL^{(g)}_nY^T) + \lambda tr(YL^{(l)}_nY^T) \\ &\quad\; s.t. \quad YY^T = I; 0 \le \lambda \le 1 \end{aligned} \tag{11}$$

where $tr(\cdot)$ is the trace operator. When λ is fixed, the global optimal solution of Y can be obtained based on the Theorem 1 described below.

Theorem 1 ***(Ky-Fan Theorem)*** *Given a symmetric matrix $A\in\mathbb{R}^{n\times n}$, let the smallest k eigenvalues be $\lambda_1\le\lambda_2\le\cdots\le\lambda_k$ and the corresponding eigenvectors be $E=[e_1,e_2,\cdots,e_k]$. Then given an arbitrary unitary matrix $X\in\mathbb{R}^{n\times k}$, the trace of X^TAX is minimized when X is an orthonormal basis for the eigenspace of A associated with its algebraically smallest eigenvalues, that is*

$$\min_{X\in\mathbb{R}^{n\times k}, X^TX=I_k} tr(X^TAX) = tr(E^TAE) = \sum_{i=1}^{k}\lambda_i \tag{12}$$

Therefore, the optimal X is given by EQ, where Q is an arbitrary orthogonal matrix.

The detailed proof of the Ky-Fan theorem can be found in [16]. Based on the Ky-Fan theorem, the optimal solution of Y is given as the eigenvectors associated with the smallest k eigenvalues of the matrix L_n. The generation process of coordinated embedding is summarized in Algorithm 1.

Algorithm 1 Coordinated Embedding Inference for CETM

```
Input: Document-word matrix $D \in \mathbb{R}_{+}^{N \times V}$, embeddind size $M$, the parameter for controlling weights
    $\lambda$ and $0 \leq \lambda \leq 1$
Output: Coordinated embedding matrix $Y \in \mathbb{R}^{V \times M}$
Calculate $W^{g}$ as global view of context using Eq. (4).
Calculate $W^{l}$ as local view of context using Eq. (7).
Calculate $L_n^{(g)}$ and $L_n^{(l)}$ separately using Eq. (9).
$L_n = (1-\lambda)L_n^{(g)} + \lambda L_n^{(l)}$.
$Y = E^T$, where $E = [e_1, e_2, \cdots, e_M]$ are the eigenvectors associated with the smallest $M$ eigen-
    values of the matrix $L_n$.
return $Y$.
```

3.3 Topic distribution inference

Spectral decomposition of global and local contexts is a crucial prerequisite step for topic discovery. Then a fast and effective mechanism is derived to infer topic distribution. Concretely, we implement such mechanism based on K-means [10] clustering, which characterizes data by using K prototypes and centroids of clusters. Formally, given a data matrix $X = (x_1, x_2, \cdots, x_n)$, the objective function of K-means is to minimize the sum of squared errors:

$$
\begin{aligned}
J_k &= \sum_{k=1}^{K} \sum_{i \in C_k} ||x_i - m_k||^2 \\
&= \sum_i ||x_i||^2 - \sum_k \frac{1}{n_k} \sum_{i,j \in C_k} x_i^T x_j
\end{aligned}
\tag{13}
$$

where $m_k = \sum_{i \in C_k} x_i / n_k$ is the centroid of cluster C_k of n_k points.

In CETM, semantically related context words tend to be similar in embedding space, resulting in semantic clusters. The centroids of semantic clusters can be viewed as latent topic centroids. Thus, we utilize K-means clustering to obtain the centroids of clusters as topic embeddings. Given coordinated embedding matrix $Y \in \mathbb{R}^{V \times M}$, the centroids obtained by K-means are represented as:

$$
S = (s_1, s_2, \cdots, s_K)^T \tag{14}
$$

where $s_i \in \mathbb{R}^M (i \in [1, K])$ is a centroid vector corresponding to its cluster, and the number of clusters K is equivalent to the number of topics. Therefore, S is called the topic embedding matrix and $S \in \mathbb{R}^{K \times M}$.

Once the topic embedding is created, it is natural that the topic-word distribution can be measured by semantic similarity between words and topic centroids. Formally, given coordinated embedding matrix $Y \in \mathbb{R}^{V \times M}$ and topic embedding matrix $S \in \mathbb{R}^{K \times M}$, the topic-word distribution matrix T is formed as:

$$
T = \frac{Y \cdot S^T}{||Y||_F \cdot ||S^T||_F} \tag{15}
$$

where $|| \cdot ||_F$ is the Frobenius norm operator, each negative entry is set to zero to guarantee the non-negativity of matrix T. Similarly, document-topic distribution can also be inferred by semantic similarity. Given document-word matrix $D \in \mathbb{R}_{+}^{N \times V}$ and topic-word distribution matrix $T \in \mathbb{R}_{+}^{V \times K}$, the document-topic distribution matrix Θ is formed as:

$$
\Theta = \frac{D \cdot T}{||D||_F \cdot ||T||_F} \tag{16}
$$

where all negative entries in Θ are set to zeros, and thus $\Theta \in \mathbb{R}_{+}^{N \times K}$.

Therefore, according to the description above, CETM can conclude distinguishable topics in documents and discover representative words for each topic. The topic distribution inference process is summarized in Algorithm 2.

Algorithm 2 CETM Topic Distribution Inference

```
Input: Document-word matrix $D \in \mathbb{R}_{+}^{N \times V}$, coordinated embedding matrix $Y \in \mathbb{R}^{V \times M}$, number of
    topics $K$
Output: Topic-word distribution matrix $T \in \mathbb{R}_{+}^{V \times K}$, document-topic distribution matrix $\Theta \in \mathbb{R}_{+}^{N \times K}$
Initialize topic centroids $S = (s_1, s_2, \cdots, s_K)^T$ randomly, $s_j \in \mathbb{R}^M$, $j \in [1, K]$.
repeat
    $c^{(i)} = \arg\min_j ||Y^{(i)} - s_j||^2, Y^{(i)} \in \mathbb{R}^M, i \in [1, V]$.
    $s_j = \frac{\sum_{i=1}^{N} 1 \cdot \{c^{(i)}=j\} Y^{(i)}}{\sum_{i=1}^{N} 1 \cdot \{c^{(i)}=j\}}$.
until convergence
Calculate $T$ using Eq. (15).
Calculate $\Theta$ using Eq. (16).
return $T$, $\Theta$.
```

4 Experimental setup

4.1 Datasets

We evaluate our model on three widely-used datasets: 20 Newsgroups[1] (denoted as **20News**), Reuters-21578 corpus[2] (denoted as **Reuters**) and Stanford Question Answering Dataset[3] (denoted as **SQuAD**). 20News consists of 18,845 newsgroup documents partitioned into 20 different categories. Each category is related to a unique topic. Reuters contains about 10,000 newswire stories produced by Reuters journalists. Each document has been manually assigned to one or more categories. Note that the number of documents in different categories is highly imbalanced in Reuters. In order to evaluate different models on text classification, we remove the documents of Reuters appearing in more than one category, and only select the largest 8 categories, with 7,746 documents in total. SQuAD is a reading comprehension dataset collected from Wikipedia articles. To extract text for topic modeling, we regard each individual paragraph as a document which is assigned to a specific category. We obtain a total of 20,239 documents covering 13 different categories.

Following the strategy used in TopicVec [20] and CLM [37], we pre-process the raw data before performing topic modeling. We first convert all words into lowercase. Stop

[1] http://qwone.com/jason/20Newsgroups/~.

[2] http://www.daviddlewis.com/resources/testcollections/reuters21578/.

[3] https://rajpurkar.github.io/SQuAD-explorer/.

Table 2 Summary of the datasets

Dataset	#docs	#words	#categories/labels
20News	18,827	20,678	20
Reuters	7746	4759	8
SQuAD	20,239	14,515	13

words[4] and words with total frequency lower than 10 are then removed. Finally, the descriptions of the three datasets are summarized in Table 2.

4.2 Baseline methods

In the experiments, the proposed topic modeling approach CETM is compared with several strong baseline methods, which we describe below:

- **PLSA** [15], the traditional Probabilistic Latent Semantic Analysis model, which estimates parameters by Expectation Maximization (EM) method.
- **LDA** [4], the standard Latent Dirichlet Allocation in the genism library[5].
- **NMF** [17], the standard Non-negative Matrix Factorization in the scikit-learn library[6].
- **Gaussian-LDA** [7], an LDA-extended topic model which characterizes each topic as a multivariate Gaussian distribution and derives the posterior topic proportions for documents[7].
- **LFTM** [26], a latent feature topic model which extends LDA by incorporating word embeddings as latent features[8].
- **TopicVec** [20], a generative topic embedding model in which topics are depicted by embedding vectors[9].
- **CLM** [37], a collaborative language model which models topics and learns word embeddings collaboratively by matrix factorization[10].
- iDocNADEe [12], a neural autoregressive topic model which incorporates word embeddings as a distributional prior[11].

4.3 Implementation details

We implement our model using the widely-used toolkit scikit-learn[12]. The context window size is set to 10, as 5 for preceding words and 5 for following words for a given focus word when constructing local word co-occurrence matrix. The dimension of coordinated embedding is set to 250, the parameter for controlling weights λ is set to 0.1 for all datasets. For all the methods, the number of maximum iteration is set to 100. For LDA, the hyper-parameters α and β are set to $50/K$ and 0.01 respectively. For other baseline models, the hyper-parameters are set as the same with original settings. In our experiments, we evaluate the performances of CETM and the other baseline methods on two typical tasks. First, we evaluate these models on topic coherence, which depicts the quality of topics discovered by the models. Second, we evaluate the models on text classification task, which shows the abilities of the topic models in characterizing the documents.

5 Experimental results

In this section, we compare CETM with the baseline methods from both quantitative and qualitative perspectives.

5.1 Evaluation of topic coherence

Through topic modeling, each topic is represented by its word distribution. The quality of topics learned by different models can be evaluated by topic coherence.

5.1.1 Evaluation metrics

To quantitatively evaluate topic models, two widely used topic coherence metrics are adopted. One is the coherence score [24], which aims to automatically evaluate the coherence of topics. Given a topic t with top U words $V^{(t)} = \{v_1^t, v_2^t, \cdots, v_U^t\}$, the coherence score of this topic is defined as:

$$C(t; V^{(t)}) = \sum_{u=2}^{U} \sum_{l=1}^{u-1} \log \frac{D(v_u^{(t)}, v_l^{(t)}) + 1}{D(v_l^{(t)})}, \tag{17}$$

[4] Stop words list is from NLTK: http://www.nltk.org/nltk_data/.

[5] https://radimrehurek.com/gensim/models/ldamodel.html.

[6] http://scikit-learn.org/stable/modules/generated/sklearn.decomposition.NMF.html.

[7] https://github.com/rajarshd/Gaussian_LDA.

[8] https://github.com/datquocnguyen/LFTM.

[9] https://github.com/askerlee/topicvec.

[10] https://github.com/XunGuangxu/2in1.

[11] https://github.com/pgcool/iDocNADEe.

[12] https://scikit-learn.org/stable.

Table 3 Results of topic coherence on 20News dataset

Models	Coherence Score				PMI Score			
	$U = 5$	$U = 10$	$U = 20$	$U = 50$	$U = 5$	$U = 10$	$U = 20$	$U = 50$
PLSA	– 15.151	– 78.597	– 365.693	– 2684.952	1.976	1.955	1.801	1.613
LDA	– 15.308	– 80.482	– 368.820	– 2694.437	2.215	2.037	1.978	1.889
NMF	– 18.051	– 85.538	– 417.199	– 2796.776	1.761	1.633	1.62	1.548
Gaussian-LDA	– 19.450	– 94.523	– 435.903	– 3407.968	1.945	1.812	1.834	1.901
LFTM	– 16.589	– 78.541	– 385.734	– 2807.011	1.903	1.921	1.810	1.745
TopicVec	– 14.253	– 72.301	– 358.943	– 2612.486	2.251	2.225	2.201	2.016
CLM	– 11.624	– 60.303	– 282.799	– 2275.523	2.341	2.202	2.115	2.121
iDocNADEe	– 12.015	– 51.751	– 293.054	– 2075.153	2.287	2.213	2.152	**2.216**
CETM	**– 8.832**	**– 44.991**	**– 229.075**	**– 1862.196**	**2.382**	**2.316**	**2.217**	2.204

Table 4 Results of topic coherence on Reuters dataset

Models	Coherence Score				PMI Score			
	$U = 5$	$U = 10$	$U = 20$	$U = 50$	$U = 5$	$U = 10$	$U = 20$	$U = 50$
PLSA	– 13.226	– 70.078	– 333.570	– 2767.808	1.501	1.353	1.168	1.254
LDA	– 12.093	– 69.806	– 352.296	– 2840.746	1.624	1.388	1.365	1.306
NMF	– 11.281	– 66.412	– 335.619	– 2705.525	1.312	1.163	1.142	1.105
Gaussian-LDA	– 24.223	– 108.453	– 478.433	– 3688.172	1.821	1.615	1.531	1.051
LFTM	– 13.268	– 71.352	– 369.009	– 2982.395	1.723	1.715	1.562	1.024
TopicVec	– 14.208	– 68.367	– 342.108	– 2727.568	1.765	1.771	1.621	**1.321**
CLM	– 11.483	– 63.083	**– 313.459**	– 2683.163	1.950	1.851	1.603	1.215
iDocNADEe	– 10.261	– 64.032	– 320.168	– 2692.481	2.021	1.804	1.642	1.263
CETM	**– 9.523**	**– 53.849**	– 314.960	**– 2682.945**	**2.143**	**1.946**	**1.712**	1.316

where $D(v_l^{(t)})$ denotes the document frequency of word $v_l^{(t)}$, and $D(v_u^{(t)}, v_l^{(t)})$ denotes the number of documents in which words $v_u^{(t)}$ and $v_l^{(t)}$ co-occurred. It follows the intuition that most probable words in the same topic tend to co-occur in documents frequently. Note that the coherence score $C(t;V^{(t)})$ is a negative number. Thus, a higher value indicates a more coherent topic. In order to explore the overall quality of topics discovered by different models, we use the averaged coherence score:

$$\overline{C} = \frac{1}{K}\sum_{t=1}^{K} C(t;V^{(t)}) \tag{18}$$

Another metric is the PMI-Score proposed by [25]. Given a topic t with top U words $V^{(t)} = \{v_1^t, v_2^t, \cdots, v_U^t\}$, the PMI-Score of this topic is:

$$\text{PMI-Score}(t) = \frac{1}{U(U-1)} \sum_{1 \le i < j \le U} \log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}, \tag{19}$$

where $p(w_i)$ is the probability that word w_i occurs in a document, and $p(w_i, w_j)$ is the probability that words w_i and w_j occur in the same document. Note that an external corpus is needed to calculate PMI-Scores in Eq. (19), a higher PMI-Score indicates better learned topics for a model. In our experiments, we use 3 million English Wikipedia articles as external corpus and calculate the averaged PMI-Score for evaluating the overall topic coherence.

For 20News, Reuters and SQuAD, the number of topics K is separately set to 20, 8 and 13, as there are 20 groups, 8 categories and 13 categories, respectively.

5.1.2 Experimental results

The experimental results of topic coherence on 20News, Reuters and SQuAD are shown in Tables 3, 4 and 5 respectively. We vary the number of top words per topic $U = \{5, 10, 20, 50\}$, and the best results are highlighted in boldface.

As we can see from the results, the topic coherence scores of generative topic models such as PLSA and LDA are similar on all the datasets. Gaussian-LDA performs inferior to all other models in terms of coherence score but performs much better in PMI-Score. This may be because the top words in Gaussian-LDA are selected by Gaussian probabilities, the coherence score is calculated based on the given dataset while the PMI-Score is evaluated on a much larger external corpus. From Table 3, CETM significantly outperforms other baseline models on 20News dataset. We observe that iDocNADEe achieves the highest PMI score when $U = 50$,

Table 5 Results of topic coherence on SQuAD dataset

Models	Coherence Score				PMI Score			
	$U = 5$	$U = 10$	$U = 20$	$U = 50$	$U = 5$	$U = 10$	$U = 20$	$U = 50$
PLSA	– 20.124	– 81.301	– 389.361	– 3120.245	1.751	1.612	1.634	1.603
LDA	– 19.892	– 80.254	– 378.250	– 3122.014	1.778	1.631	1.635	1.612
NMF	– 19.035	– 77.854	– 386.014	– 3121.847	1.774	1.634	1.641	1.625
Gaussian-LDA	– 21.298	– 90.364	– 392.014	– 3302.418	1.833	1.781	1.804	1.706
LFTM	– 17.364	– 71.661	– 381.021	– 3250.145	1.824	1.815	1.806	1.801
TopicVec	– 15.351	– 70.304	– 375.698	– 3127.452	1.955	1.934	1.923	1.827
CLM	– 14.684	– 68.697	– 361.045	– 2997.351	2.032	1.963	1.934	1.816
iDocNADEe	– 15.022	– 69.364	– 370.361	– 3004.218	2.321	2.204	**2.134**	**2.124**
CETM	**– 13.523**	**– 66.849**	**– 353.561**	**– 2936.881**	**2.325**	**2.219**	2.038	2.117

and similar results are also observed on SQuAD dataset. It is due to that the pre-trained word embeddings are used as prior knowledge in iDocNADEe, which benefits the evaluation when using a large external corpus. From Table 4, we observe CETM performs much better than PLSA, LDA, NMF, Gaussian-LDA and LFTM. In addition, CLM achieves the highest coherence score when $U = 20$ and TopicVec achieves the highest PMI score when $U = 50$. This may be because all the topics (i.e., categories) in Reuters dataset are semantically close, which makes it hard for CETM to distinguish different topic centroids. Furthermore, from experimental results of all the datasets, we find that both coherence score and PMI score decrease with the number of top words U increasing, which is due to the fact that a word with lower ranking in the topic-word distribution contributes less to the representation of the topic it belongs to. When enlarging U, more words with lower ranking are included in calculating word co-occurrence, resulting in topic coherence inferior. In summary, compared with other baseline methods, our CETM achieves better ability to discover topic structures.

Table 6 Text classification results on 20News dataset

Models	Precision	Recall	F1
PLSA	0.713	0.702	0.707
LDA	0.730	0.723	0.717
NMF	0.686	0.687	0.697
Gaussian-LDA	0.411	0.415	0.420
LFTM	0.716	0.715	0.712
TopicVec	0.723	0.716	0.711
CLM	0.809	0.801	0.799
iDocNADEe	0.825	**0.842**	0.821
CETM	**0.841**	0.838	**0.838**

5.2 Evaluation of text classification

Through topic modeling, each document can be represented by its topic distribution. Thus, the quality of a topic model can be evaluated by the performance of document classification.

5.2.1 Evaluation metrics

We use the topic distribution over documents as the features of documents, and then perform a classification task for all the methods. For easy comparison to baseline methods, we set the number of topics K to 280 for 20News, and 110 for Reuters respectively, following CLM [37]. Similarly, we set the number of topics K to 180 for SQuAD dataset.

For each dataset, we randomly split the topic-level representations of documents into training set and testing set with the proportion of testing set being 0.3. Then we employ a linear SVM classifier with $\ell 1$ regularization using scikit-learn toolbox[13]. In order to evaluate the overall performance of all the models, we adopt the averaged results on testing set over 10 runs. On this task, we adopt macro-averaged precision, recall and F1 value as the evaluation metrics. In addition, we report the micro-averaged ROC curves and AUC values of our model on different datasets.

5.2.2 Experimental results

The classification results of different methods on 20News, Reuters and SQuAD are presented in Tables 6, 7 and 8 respectively, and the best results are highlighted in boldface. Note that the results for CLM we report in Tables 6 and 7 are different from the original paper. It is because by checking their released code, we discover that the results in the original paper are based on training set not on testing set. From the results on different datasets, we can observe that CETM consistently outperforms the baselines by a noticeable margin. In addition, Gaussian-LDA performs poorly on all the datasets. After checking the topic embeddings generated

[13] http://scikit-learn.org/stable/modules/svm.html.

Table 7 Text classification results on Reuters dataset

Models	Precision	Recall	F1
PLSA	0.910	0.906	0.903
LDA	0.892	0.865	0.882
NMF	0.901	0.867	0.884
Gaussian-LDA	0.462	0.315	0.353
LFTM	0.876	0.667	0.682
TopicVec	0.918	0.909	0.913
CLM	0.915	0.899	0.906
iDocNADEe	0.885	0.901	0.889
CETM	**0.937**	**0.916**	**0.925**

Table 8 Text classification results on SQuAD dataset

Models	Precision	Recall	F1
PLSA	0.530	0.529	0.530
LDA	0.562	0.541	0.546
NMF	0.559	0.541	0.545
Gaussian-LDA	0.457	0.404	0.438
LFTM	0.590	0.584	0.588
TopicVec	0.584	0.509	0.533
CLM	0.651	0.648	0.648
iDocNADEe	0.658	0.664	0.659
CETM	**0.664**	**0.677**	**0.661**

by Gaussian-LDA, we observe that different topics share similar Gaussian distributions and hence the document-topic representations are almost non-discriminative. The generative models, e.g., PLSA, LDA and LFTM stably exceed the Gaussian-LDA and achieve similar results. By combining word embedding and LDA, TopicVec achieves better results on Reuters, but worse results on 20News and SQuAD, which is similar to the aforementioned evaluation of topic coherence, indicating that TopicVec is more suitable for small-scale datasets. Note that iDocNADEe is a strong baseline and achieves the highest recall on 20News dataset, but our CETM still performs best in F1 value. According to Table 8, we find that all models perform inferior on SQuAD dataset than on the other two datasets. This is because the number of documents in different categories (labels) is highly imbalanced in SQuAD, and we do not eliminate those categories with few documents, thus making it difficult for document-topic distribution representing all documents well. Figure 2 depicts the ROC curves and AUC scores of CETM on different datasets. CETM achieves AUC scores higher than 0.85 on all the datasets, which shows significant effectiveness for classification. In conclusion, CETM achieves better performance than other models on all the datasets, which verifies impressive ability in characterizing documents.

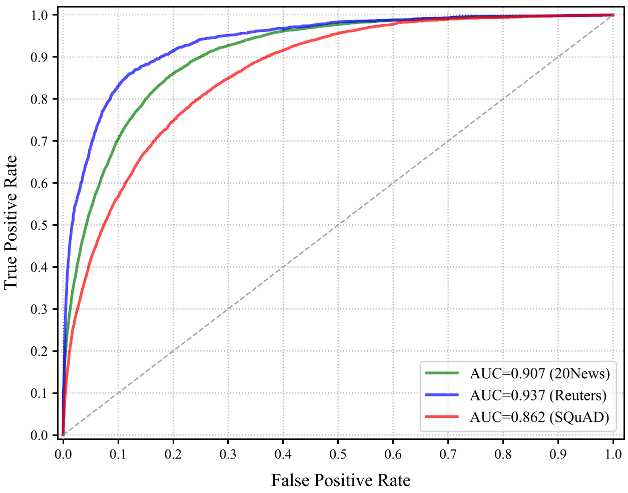


Fig. 2 ROC curves and AUC scores of CETM on different datasets

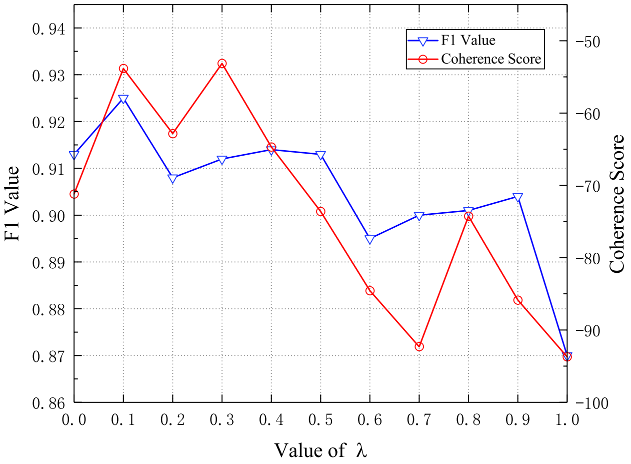


Fig. 3 Effect of parameter λ with the setting of $K = 8$ and top words $U = 10$ on Reuters

5.3 Parameter and complexity analysis

In order to further verify that the coordination of global and local contexts in CETM is effective for topic modeling, we investigate the impact of the parameter setting of λ which controls the weight of global context information and local context information. Due to the limited space, we conduct the experiment on Reuters dataset, setting the number of topics $K = 8$ and the number of top words per topic $U = 10$.

Figure 3 reports the classification F1 value and topic coherence score of CETM by varying λ from 0 to 1. As to the F1 value, CETM can achieve much better performance when λ is set to 0.1. For topic coherence, CETM obtains almost the highest score when λ is set to 0.1 and 0.3. Generally, when $\lambda > 0.5$, further increasing the value of λ results

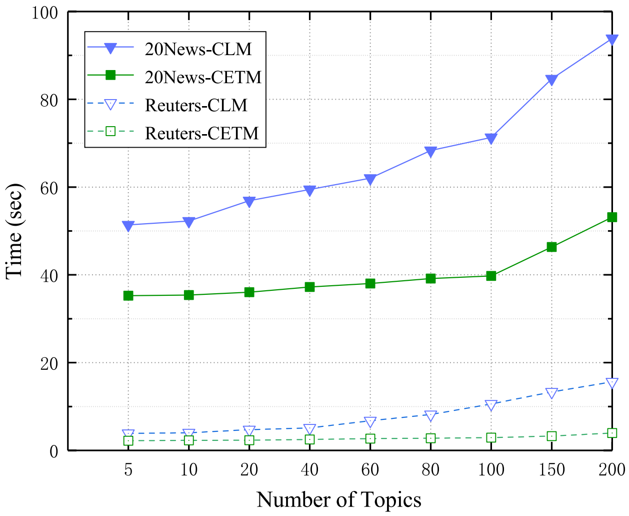


Fig. 4 Comparison of average training time per iteration

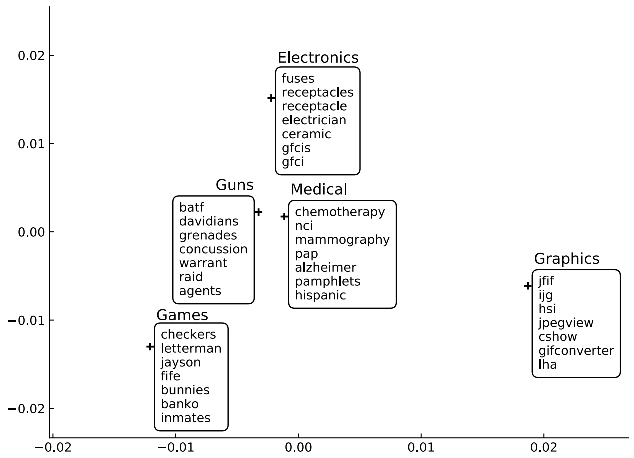


Fig. 5 2-D PCA projection of the topic distribution on 20News

in performance degradation. It should be mentioned that when λ is set to 0 or 1, neither classification precision nor topic coherence score achieves best result. When $\lambda = 0.1$, the overall performance is almost best. This suggests that, by leveraging global context information and local context information, CETM is more effective for topic discovery. Thus, we set $\lambda = 0.1$ in our experiments.

In addition, to evaluate the efficiency of our model, we compare the time complexity between our model and CLM, since both are based on matrix factorization and implemented using Python. We calculate the average running time for each iteration during training process using one CPU only. The results are shown in Fig. 4. For Reuters dataset, time cost per iteration of CETM and CLM is close, while for a larger dataset 20News, time cost of CETM is much lower. Generally, CETM performs faster than CLM with the number of topics increasing. Moreover, there are more hyper-parameters to be set in CLM, including the parameters controlling the weights and regularization λ_d, λ_w and λ_s. However, in CETM, no additional hyper-parameter is required except the parameter λ. Therefore, CETM requires less model parameters and has lower computational cost compared with CLM.

5.4 Qualitative analysis of topics

CETM discovers topic structures by applying clustering on coordinated embeddings, and the semantic relations between the topic distributions and coordinated embeddings is modeled in Eq. refcomputespsT. To evaluate the proposed model qualitatively, we report the topics discovered by our model and analyze whether meaningful semantics have been captured. We apply classical PCA technique to visualize topic structures.

Due to space limitation, we only report five topics discovered by our model for 20News. For each topic, we visualize it with the top 7 words which have the highest probabilities under the topic. Figure 5 shows the 2-dimensional (2-D) PCA projection of the selected topic distributions learned by our model. It can be observed that the semantic differences between topics are correlated with the distances between topic distributions in 2D embedding space. For example, the topic “Gaphics” is far from the topic “Guns”, while the topic “Medical” is much nearer from the topic “Guns”. This accords with human intuition. Furthermore, the five selected topics can be easily interpreted with the top 7 words. For example, the top 7 words in “Electronics” topic are “fuses”, “receptacles”, “receptacle”, “electrician”, “ceramic”, “gfcis” and “gfci”, where “gfci” is short for “Ground Fault Circuit Interrupter”. The words in a topic are semantically related, verifying that CETM can significantly discover coherent topics.

6 Conclusion and future work

In this paper, we proposed a Coordinated Embedding Topic Model (CETM) for topic discovery, with the aim to ensemble spectral decomposition and clustering by leveraging the benefits of global and local context information. In contrast to existing work, CETM could discover more coherent topics and characterize documents more effectively. In addition, lower computational cost and less hyper-parameters were required. The extensive experiments on three widely-used datasets showed that our model achieved better performance than other baseline methods. Moreover, we provided parameter analysis and showed time cost to prove the effectiveness of our model.

In the future, we plan to extend our model to support parallel computing to handle the topic discovery of large-scale documents. Another research direction is to apply our model into more natural language processing tasks, such as document retrieval and text summarization.

Acknowledgements This work was supported by the Science and Technology Planning Project of Guangdong Province, China (2017B030306016), the Pearl River Science & Technology Star Project (201610010046), the Special Support Program of Guangdong Province (201528004), and the National Natural Science Foundation of China (NSFC) (No. 61272200). This work was partially supported by the Natural Science Foundation of Guangdong Province of China (No. 2019A1515011705), the SIAT Innovation Program for Excellent Young Researchers, and the project AWS13C008.

References

1. Belkin M, Niyogi P (2002) Laplacian eigenmaps and spectral techniques for embedding and clustering. In: Advances in neural information processing systems, pp. 585–591
2. Bengio Y, Ducharme R, Vincent P, Jauvin C (2003) A neural probabilistic language model. J Mach Learn Res 3:1137–1155
3. Blei D, Lafferty J (2006) Correlated topic models. Adv Neural Inf Process Syst 18:147
4. Blei DM, Ng AY, Jordan MI (2003) Latent dirichlet allocation. J Mach Learn Res 3:993–1022
5. Cai D, Mei Q, Han J, Zhai C (2008) Modeling hidden topics on document manifold. In: Proceedings of the 17th ACM conference on information and knowledge management, pp. 911–920. ACM
6. Church KW, Hanks P (1990) Word association norms, mutual information, and lexicography. Comput Linguist 16(1):22–29
7. Das R, Zaheer M, Dyer C (2015) Gaussian lda for topic models with word embeddings. In: Proceedings of the 53rd annual meeting of the association for computational linguistics and the 7th international joint conference on natural language processing, (Volume 1: Long Papers)
8. Deerwester S, Dumais ST, Furnas GW, Landauer TK, Harshman R (1990) Indexing by latent semantic analysis. J Am Soc Inf sci 41(6):391–407
9. Dhillon PS, Foster DP, Ungar LH (2015) Eigenwords: spectral word embeddings. J Mach Learn Res 16(1):3035–3078
10. Ding C, He X (2004) K-means clustering via principal component analysis. In: Proceedings of the twenty-first international conference on machine learning, p 29. ACM
11. Ding C, Li T, Peng W (2006) Nonnegative matrix factorization and probabilistic latent semantic indexing: equivalence chi-square statistic, and a hybrid method. AAAI 42:137–143
12. Gupta P, Chaudhary Y, Buettner F, Schütze H (2019) Document informed neural autoregressive topic models with distributional prior. Proc AAAI Conf Artif Intell 33:6505–6512
13. Harris ZS (1954) Distributional structure. Word 10(2–3):146–162
14. Hinton GE, Salakhutdinov R (2009) Replicated softmax: an undirected topic model. In: Advances in neural information processing systems, pp 1607–1614
15. Thomas H (1999) Probabilistic latent semantic analysis. In: Proceedings of the fifteenth conference on uncertainty in artificial intelligence, pp. 289–296. Morgan Kaufmann Publishers Inc
16. Horn RA, Horn RA, Johnson CR (1990) Matrix analysis. Cambridge University Press, Cambridge
17. Lee DD, Seung HS (2001) Algorithms for non-negative matrix factorization. In: Advances in neural information processing systems, pp 556–562
18. Levy O, Goldberg Y (2014) Neural word embedding as implicit matrix factorization. In: Advances in neural information processing systems, pp 2177–2185
19. Li D., Zhang J, Li P (2019) Tmsa: a mutual learning model for topic discovery and word embedding. In: Proceedings of the SIAM international conference on data mining, pp 684–692
20. Li S, Chua T-S, Zhu J, Miao C (2016) Generative topic embedding: a continuous representation of documents. In: Proceedings of the 54th annual meeting of the association for computational linguistics (Volume 1: Long Papers)
21. Lin C, He Y (2009) Joint sentiment/topic model for sentiment analysis. In: Proceedings of the 18th ACM conference on information and knowledge management, pp 375–384. ACM
22. Mikolov T, Sutskever I, Chen K, Corrado GS, Dean J (2013) Distributed representations of words and phrases and their compositionality. In: Advances in neural information processing systems, pp 3111–3119
23. Mikolov T, Yih W-T, Zweig G (2013) Linguistic regularities in continuous space word representations. In: Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies, pp 746–751
24. Mimno D, Wallach HM, Talley E, Leenders M, McCallum A (2011) Optimizing semantic coherence in topic models. In: Proceedings of the conference on empirical methods in natural language processing, pp. 262–272. Association for Computational Linguistics
25. Newman D, Lau JH, Grieser K, Baldwin T (2010) Automatic evaluation of topic coherence. In: Human language technologies: the 2010 annual conference of the North American chapter of the association for computational linguistics, pp. 100–108. Association for Computational Linguistics
26. Nguyen DQ, Billingsley R, Du L, Johnson M (2015) Improving topic models with latent feature word representations. Trans Assoc Comput Linguist 3:299–313
27. Roweis ST, Saul LK (2000) Nonlinear dimensionality reduction by locally linear embedding. Science 290(5500):2323–2326
28. Shlens J (2014) A tutorial on principal component analysis. arXiv preprint arXiv:1404.1100
29. Soleimani BH, Matwin S (2018) Spectral word embedding with negative sampling. In: Proceedings of the thirty-second AAAI conference on artificial intelligence, New Orleans, Louisiana, USA, February 2–7, 2018, pp 5481–5487
30. Srivastava N, Salakhutdinov RR, Hinton GE (2013) Modeling documents with deep boltzmann machines. arXiv preprint arXiv:1309.6865
31. Steyvers M, Griffiths T (2007) Probabilistic topic models. Handb Latent Semant Anal 427(7):424–440
32. Suh S, Choo J, Lee J, Reddy CK (2016) L-ensnmf: boosted local topic discovery via ensemble of nonnegative matrix factorization. In: 2016 IEEE 16th international conference on data mining (ICDM), IEEE, pp 479–488.
33. Teh YW (2006) A hierarchical bayesian language model based on pitman-yor processes. In: Proceedings of the 21st international conference on computational linguistics and the 44th annual meeting of the association for computational linguistics, pp. 985–992. Association for Computational Linguistics
34. Tenenbaum JB, De Silva V, Langford JC (2000) A global geometric framework for nonlinear dimensionality reduction. Science 290(5500):2319–2323
35. Wallach HM (2006) Topic modeling: beyond bag-of-words. In: Proceedings of the 23rd international conference on machine learning, ACM, pp 977–984
36. Xia T, Tao D, Mei T, Zhang Y (2010) Multiview spectral embedding. IEEE Trans Syst Man Cybernet Part B (Cybernet) 40(6):1438–1446

37. Xun G, Li Y, Gao J, Zhang A (2017) Collaboratively improving topic discovery and word embeddings by coordinating global and local contexts. In: Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining, ACM, pp 535–543