TMSA: A Mutual Learning Model for Topic Discovery and Word Embedding

Dingcheng Li **Jingyuan Zhang** **Ping Li**
Cognitive Computing Lab (CCL), Baidu Research USA
{lidingcheng, zhangjingyuan03, liping11}@baidu.com

Abstract

Both topic modeling and word embedding map documents onto a low-dimensional space, with the former clustering words into a global topic space and the latter into a local continuous embedding space. In this study, we propose the TMSA framework to unify these two complementary patterns by the construction of a mutual learning mechanism between word-cooccurrence based topic modeling and auto-encoder. In our model, word topics generated with topic modeling are passed into auto-encoder to impose topic sparsity so that auto-encoder can learn topic-relevant word representations. In return, word embedding learned by auto-encoder is sent back to topic modeling to improve the quality of topic generations. Empirical studies show the effectiveness of the proposed TMSA model in discovering topics and embedding words.

1 Introduction

Both topic models [4] and word embedding [2, 5, 18] models play significant roles in modeling human languages and have become indispensable tools for natural language processing. In the past decade, topic modeling [4, 15], such as Probabilistic Latent Semantic Analysis (PLSA) or Latent Dirichlet Allocations (LDA) [4], has successfully revealed the thematic structure of collection of documents with exploring the patterns represented by word co-occurrence matrix. The advantage of topic modeling is its global clustering capacity across documents. When the corpus is large enough, semantic relatedness and coherent topics can be exposed without supervision. In contrast, word embedding models have been proved to be an effective approach to transform sparse discrete words into dense and low-dimensional continuous vectors [2, 18, 12, 21]. Since word embedding usually utilizes local word collocation patterns to construct an embedding link function, the semantic and syntactic relatedness represented is also more local, compared to topic modeling.

As they complement each other in language modeling, it motivates us to pursue constructing an integrated model which can make full use of their respective advantages. Two common characteristics for topic modeling and word embedding are the nature of dimensionality deduction and their semantic relatedness. Quite a few works have used word embeddings to improve topic modeling [20, 15]. CLM [27] and STE [23] are proposed to combine topic modeling and word embedding. CLM applies nonnegative matrix factorization to both topic modeling and word embedding. STE employs skip-gram to learn different topic-specific word embeddings to avoid polysemy. The existing methods, however, do not explicitly consider the mutual influence of global topics and local contexts in the learning process. Thus, the interaction between global topics and local contexts cannot be fully explored to boost the performance.

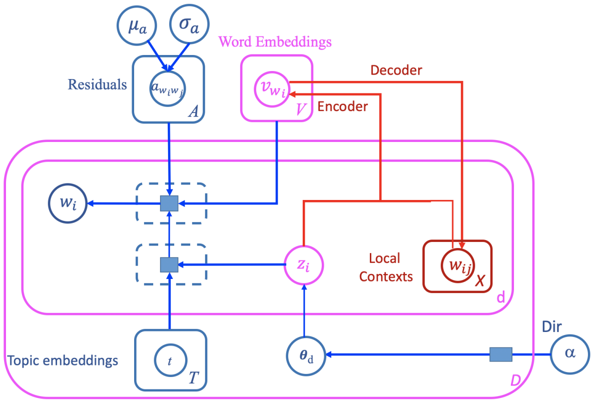

Figure 1: Graphic representation of TMSA, our proposed mutual learning framework. The part in blue is about the topic learning process via the TMSA$_{topic}$ component and the part in red is about the word embedding via the TMSA$_{word}$ component. The part in pink is shared by both components for the mutual learning procedure.

In this paper, we propose a unified framework TMSA (Topic Modeling and Sparse Autoencoder; see Figure 1) to explicitly incorporate the mutual influence of global topics and local contexts into the learning process. In TMSA, the influence of local word embeddings are integrated into the discovery of topics via the TMSA$_{topic}$ component named as *topic modeling boosted with sparse autoencoder*. The semantic information of word embedding helps TMSA$_{topic}$

learn topics in an effective way. In addition, the topics learned from TMSA$_{topic}$ will be further integrated into the word embedding process via the TMSA$_{word}$ component named as *sparse autoencoder sparsified with topic modeling*. Both topics and contexts will be reflected in the learned embeddings of words. The full architecture of the proposed TMSA is shown in Figure 1. With the mutual learning scheme, TMSA has the following advantages. First, parameter tuning and inferences can be done in a unified framework. Second, the mechanism of back propagation in sparse autoencoder can be utilized for fine tuning word embeddings. Third, extra layers can be easily added to handle other tasks, such as adding a softmax layer for classifications.

In summary, our key contributions are as follows:

- We propose a unified framework TMSA to improve topic discovery and word embedding simultaneously via a mutual learning mechanism.
- We introduce an efficient algorithm to boost topic learning by taking advantage of local context information from semantic word embedding.
- We design a unique topic based sparse autoencoder to improve the word representation learning by encoding both global topics and local context information into the learned embeddings.
- We demonstrate the effectiveness of TMSA by comparing it with several state-of-the-art methods on both topic modeling tasks and word embedding tasks.

2 Related Work

As it is discussed in the Introduction Section, the main theme of this work is to coordinate global topics and local contexts for better topic discovery and word embeddings. Therefore, most relevant works involve topic modeling and word embedding learning.

2.1 Topic modeling and its variations Topic modeling is a powerful unsupervised tool to discover latent semantic structure from a text corpus. The most representative one is the Latent Dirichlet Allocation (LDA) [4]. Typically, only a small number of topics are present in each document and only a small number of words have high probability in each topic. This pattern motivated [4] to deploy Dirichlet priors to regularize the topic distributions. Semantic centroids have the same nature as topics in LDA. The semantic relatedness exists in continuous embedding space while the topic related words exist in discrete space. This similarity leads explorations in common semantic centroids. For example, [20] proposed to improve topic models with latent feature word representations (Latent Feature Topic Modeling or LFTM for short). Specifically, they replace the topic-to-word Dirichlet multinomial component that generates words from topics with a two-component mixture of a topic-to-word Dirichlet multinomial component and a latent feature component. The latent feature component is a product of two matrices, pre-trained word embedding and updated topic embedding. In contrast, topic embedding, as topics in LDA, catches global context information while reflecting semantic centroids.

2.2 Word Embedding Current word embedding related works are usually based on neural probabilistic language model introduced by [2]. It has been proven to be able to capture semantic regularities in language by learning context information represented with the local word co-occurrences. Later, Mnih and Hinton [19] proposed three different embedding functions to model the conditional distribution of a word given its context (or vice versa). However, these methods are not scalable on large corpora due to the interaction matrices between the embeddings. [18] proposed Skip-Gram and Continuous Bag Of Words (CBOW) to improve the efficiency of word embeddings via direct interaction between two embeddings, which can be efficiently trained on large corpora and achieve good performance on various linguistic tasks. In particular, the skip-gram with negative sampling for training word embedding is discovered to implicitly factorize the point-wise mutual information matrix of the local word co-occurrence patterns [11].

2.3 Integrated Framework Besides above work, Topic Word Embedding (TWE) [16] was proposed to concatenate topic embedding with word embedding to form topical word embedding for each word. Li et al. [15] extended LDA to a model named as TopicVec. The extension partially follows LFTM by defining the probability function as a mixture of the conventional multinomial distribution and a link function between the embeddings of the focus words and topics. Furthermore, TopicVec treats pre-trained topic labels as special words and learns embeddings for topics by including the topic labels in the neural architecture. Another work along this line is Gaussian LDA [6]. It uses pre-trained word embeddings learned from large external corpora such as Wikipedia and then models topics with Gaussian distributions in the word embedding space. In addition, Skip-gram Topical word Embedding (STE) [23] was proposed to learn different topic-specific word embeddings to avoid the problem of polysemy. Recently, models like [29] and [14] construct informative and asymmetric Dirichlet priors with word embeddings as external knowledge. All of them somewhat make efforts to construct a channel between topic modeling and word embedding. Namely, they do not take into considerations much of the mutual influence of global topics and local contexts explicitly during the learning process.

However, these composite models combine topic models and word embeddings in a separate and heuristic manner. Research which attempts to integrate both aspects into a framework comes from Collaborative Language Model

(CLM) [27] and Correlated Topic Modeling Using Word Embeddings [28]. CLM was proposed to formulate the topic modeling and word embedding into a co-factorization fashion. It employs non-negative matrix factorization (NMF) to obtain global topic matrix and utilizes the shifted positive point-wise mutual information matrix to generate word embedding vectors. The second one extends Gaussian LDA by modeling topic correlations with the help of word embeddings. Meanwhile, as their topic discovery process starts from learning the word embedding with semantic regularities, the model constructs a mutual learning mechanism. Yet, these models are to some degree constructed with topic modeling as the dominant so that word embedding plays less important roles. In contrast, our model aims at launching a mutual learning mechanism, explicitly enhancing the interactions of global topics and local contexts via two tightly correlated components TMSA$_{topic}$ and TMSA$_{word}$.

3 Problem Statement

Given a set of documents, the document-word matrix $\mathbf{D}$ represents the global context information. The topics for documents will be effectively discovered via the proposed topic modeling module TMSA$_{topic}$ by explicitly taking the word embedding information from local contexts into consideration. The local context information is represented by the word co-occurrence matrix $\mathbf{X}$, which is extracted from a sequence of words in documents within a fixed text window. Each word sequence has a focus word and its neighboring context words within a text window centered at the focus word. $x_{ij} \in \mathbf{X}$ records the times a word w_j appears in a word w_i's contexts. The word embeddings will be learned from $\mathbf{X}$ via the proposed TMSA$_{word}$ by incorporating the discovered topics into the embedding process. In accordance, word embedding learning and topics discovery form a mutual interactive cycle and continue till convergence.

4 Methodology

As is shown in Figure 1, our proposed TMSA framework consists of two components, the topic modeling module TMSA$_{topic}$ with the blue color in the figure and the word embedding module TMSA$_{word}$ with the red color in the figure. These two components closely interact with each other through the mutual learning mechanism with the shared part in pink in the figure. We will elaborate on these components.

4.1 Topic Modeling Boosted with Sparse Autoencoder

The topic modeling module, TMSA$_{topic}$, as shown in blue in Figure 1, is a generative process with word embeddings, topic embeddings and residuals for regularization. TMSA$_{topic}$, similar to LDA, represents each document d from a corpus as a probability distribution over topics, where each topic is modeled by a probability distribution over words in a fixed vocabulary. With the text corpus, the topic discovered can reflect the global semantic relatedness. The probability of a word is governed by such latent topics. TMSA$_{topic}$ is also a generative model. Differently, besides employing Dirichlet prior to generate document topic distributions, normal distributions are utilized to regulate the generations of topic embedding.

Here, we define the generative process and likelihood.

1. For each word, look up the word embedding v_{w_i} from the word embedding matrix, $\mathbf{V}$.
2. For each word co-occurrence of w_i and w_j, draw the residual a_{w_i,w_j} from $\mathcal{N}(0, \frac{1}{2g(\widetilde{p}(w_i,w_j))})$.
3. For the k-th topic, draw a topic embedding uniformly from a hyperball of radius γ as $\mathbf{t}_k \sim Unif(\beta_\gamma)$.
4. For each document d_i:
   (a) Draw the mixing topic proportions θ_i from the Dirichlet prior $Dir(\alpha)$.
   (b) For the j-th word:
      i. Draw topic assignment z_{ij} from the θ_i.
      ii. Draw word w_{ij} from W according to $p(w_{ij}|w_{i,j-c} : w_{i,j-1}, z_{ij}, d_i)$.

In this generative process, the word embedding matrix, $\mathbf{V}$ is updated in the TMSA$_{word}$ module. The residual a_{w_i,w_j} is a regulation of bigram w_i, w_j. p_{w_i,w_j} is a link function or a probability function for a bigram w_i, w_j, defined as,

$$p(w_i, w_j) = \exp\{\mathbf{v_{w_j}}^T\mathbf{v_{w_i}} + a_{w_i,w_j}\}p(w_i)p(w_j) \tag{4.1}$$

where $\mathbf{v_{w_j}}^T\mathbf{v_{w_i}}$ refers to the linear interactions between two word vectors and a_{w_i,w_j} is a residual information representing nonlinear or noisy interactions between two words.

Eq. (4.1) is actually the regularized pointwise mutual information between two word vectors. $\mathbf{t}_k$ is the topic embedding for k-th topic and β_γ is a hyperparameter. The fourth step is similar to LDA. Nonetheless, the generative process for each word w_{ij} is based on a link function $p(w_{ij}|w_{i,j-c} : w_{i,j-1}, z_{ij}, d_i)$ extended from (4.1) defined by [15], in which, an interaction function between the word vector and topic embedding is added. Corresponding to Figure 1, the above generative process can be summarized as a likelihood function for each document.

$$\begin{aligned}\mathcal{L}_{topic} &= p(\mathbf{D}, \mathbf{A}, \mathbf{V}, \mathbf{Z}, \mathbf{T}, \theta|\alpha, \mu) \\ &= \prod_{i=1}^{N} p(v_{w_i;}\mu_i) \prod_{i,j=1}^{N,N} p(a_{w_i,w_j}; g(\widetilde{p}(w_i, w_j))) \prod_{k}^{K} \\ &\quad Unif(\beta_\gamma) \prod_{d=1}^{M} p(\theta_d|\alpha)p(\mathbf{z}_d|\theta_d)p(\mathbf{w}_d|\mathbf{V}, \mathbf{A}, \mathbf{t}_d, \mathbf{z}_d)\end{aligned} \tag{4.2}$$

where $\mathbf{D}$, $\mathbf{A}$, $\mathbf{V}$, $\mathbf{Z}$, $\mathbf{T}$ refer to a document set, the residual matrix, the word embedding matrix, the topic matrix and topic embedding matrix respectively. In addition,

$p(v_{w_i};\mu_i)$ and $p(a_{w_i,w_j}; g(\widetilde{p}(w_i,w_j))$ are the two Gaussian priors for generating the word co-occurrences. The second term $g(\widetilde{p}(w_i,w_j))$ is a nonnegative monotonic transformation for $\widetilde{p}(w_i,w_j)$, aiming at penalizing the residual a_{w_i,w_j}

4.1.1 Optimization of TMSA$_{topic}$.

Following conventions, we optimize the regularized maximum likelihood function of $\mathcal{L}_{topic}$. Based on the distributions from the generative process, the complete-data likelihood of a corpus D can be factorized as follows:

$$
\begin{aligned}
(4.3)\quad & p(\mathbf{D},\mathbf{A},\mathbf{V},\mathbf{Z},\mathbf{T},\theta|\alpha,\mu,\gamma) \\
& = \frac{1}{Z(\boldsymbol{\Theta}), U_\gamma^K} \exp\big\{ - \sum_{i,j=1}^{N,N} g(\widetilde{p}(w_i,w_j)) a^2_{w_i,w_j} \\
& - \sum_{i=1}^{N} \mu_i ||v_{w_i}||^2 \big\} \prod_{d=1}^{M} \Big\{ \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{j=1}^{K} \theta_{ij}^{\alpha_j - 1} \\
& \prod_{j=1}^{L_i} (\theta_{i,z_{i,j}} P(w_{ij}) \exp\big\{ \mathbf{v}_{w_{ij}}^T \sum_{l=j-c}^{j-1} (\mathbf{v}_{w_{il}} \\
& + \mathbf{t}_{\mathbf{z}_{ij}}) + \sum_{l=j-c}^{j-1} a_{w_{il}w_{ij}} + r_{i,z_{i,j}} \big\}) \Big\}
\end{aligned}
$$

$\mathbf{V}$ in $p(v_{w_i};\mu_i)$ can be initialized by the pretrained word embedding and updated in TMSA$_{word}$. Among them, $\frac{1}{Z(\boldsymbol{\Theta}), U_\gamma^K}$ is the normalized term and Θ refers to all relevant parameters. Similar to LDA, the variational inference algorithm is employed to update corresponding parameters. The last term in (4.3),

$$
p(w_{i,j}) \exp\big\{ \mathbf{v}_{w_{ij}}^T (\sum_{l=j-c}^{j-1}) \mathbf{v}_{w_{il}} + \mathbf{t}_{\mathbf{z}_{ij}} + \sum_{l=j-c}^{j-1} a_{w_{il}w_{ij}} + r_{i,z_{i,j}} \big\}
$$

is the latent feature vector, $p(\mathbf{w}_d|\mathbf{V},\mathbf{A},\mathbf{t}_d,\mathbf{z}_d)$. The negative log-likelihood of the corpus factorizes topic-wise into factors L_t for each topic. With L_2 regularization for topic t, this term looks like,

$$
\begin{aligned}
(4.4)\quad L_{\mathbf{z_{ij}}} = & - \sum_{w \in W} \theta^{t,w} (\mathbf{t}_{z_{ij}} \mathbf{w_{ij}}) \\
& - \log(\sum_{w' \in W} \exp(\mathbf{t}_{z_{ij}} \mathbf{w}_{ij})) + \mu ||\mathbf{t}_{z_{ij}}||_2^2 .
\end{aligned}
$$

The MAP estimate of topic vector $\mathbf{t}_{z_{ij}}$ is obtained by minimizing the regularized negative log-likelihood. The derivative with respect to the j-th element of the vector for topic z_{ij} is,

$$
(4.5)\quad \frac{\partial L_{\mathbf{z}_{ij}}}{\partial \mathbf{t}_{z_{ij}}} = - \sum w \in \mathbf{W} \theta^{z_{ij}} (\mathbf{w}_{ij} - \sum_{l \in \mathbf{W}} w_{lj} v_{w_{lj}} t_{z_{lj}})
$$

4.2 Sparse Autoencoder (SA) Sparsified with Topic Modeling

To learn embeddings of words, we adopt the classic sparse autoencoder (SA) using the self-reconstruction criterion [3, 26]. Autoencoder is an unsupervised feedforward neural network that applies backpropagation by fitting the input using the reconstructed output. It is often used to handle high-dimensional features and pre-train deep learning models. Word embeddings can also be trained via autoencoder [9, 13]. Before training autoencoder for word embedding, we first construct the word co-occurrence probabilities by counting the number of times each context word occurs around its focus word divided by the frequency of the focus word. The square root of the probabilities, denoted as $\mathbf{X}$, are considered as the input of autoencoder as in [9].

With word co-occurrence information, SA encodes the word co-occurrence $\mathbf{x}_i$ of the i-th input word to an embedding representation $\mathbf{v}_i \in \mathbb{R}^N$ by a feedforward propagation

$$
\mathbf{v}_i = f(\Phi \mathbf{x}_i + \mathbf{b}).
$$

$\Phi \in \mathbb{R}^{N \times S}$ is a weight matrix and $\mathbf{b} \in \mathbb{R}^N$ is an embedding bias vector. $f(\cdot)$ is called the activation function, e.g., the sigmoid function

$$
f(x) = \frac{1}{1 + \exp(x)}.
$$

After the feedforward pass, $\mathbf{v}_i$ is decoded to a representation

$$
\hat{\mathbf{x}}_i = f(\Phi^\top \mathbf{v}_i + \mathbf{c}).
$$

$\mathbf{c} \in \mathbb{R}^N$ is a bias vector for the decoder. A sparsity constraint is imposed on the embedding representation of $\mathbf{v}_i$ to reduce noise in SA. The overall cost function of SA is

$$
\begin{aligned}
(4.6)\quad \mathcal{L}_{SA}(\Phi,\mathbf{b}) = & \frac{1}{M} \sum_{i=1}^{M} \frac{1}{2} ||\hat{\mathbf{x}}_i - \mathbf{x}_i||^2 + \\
& \frac{\lambda}{2} ||\Phi||^2 + \xi \sum_{j=1}^{N} KL(\rho||\hat{\rho}_j),
\end{aligned}
$$

where the first term is the average of reconstruction loss on all word co-occurrences with sum-of-squares. The second term is a regularization term to prevent over-fitting. λ is the regularization parameter. The third term is the Kullback-Leibler (KL) divergence between two Bernoulli random variables with mean ρ and $\hat{\rho}_j$, respectively. It aims to control the sparsity of the weight and bias parameters Φ and $\mathbf{b}$. ρ is the sparsity parameter that specifies the level of sparsity. ξ is the weight of the sparsity term in the cost function. We use

$$
(4.7)\quad KL(\rho||\hat{\rho}_j) = \rho \log \frac{\rho}{\hat{\rho}_j} + (1-\rho) \log \frac{1-\rho}{1-\hat{\rho}_j}
$$

to penalize $\hat{\rho}_j$ deviating from the sparsity parameter ρ, with

$$
\hat{\rho}_j = \frac{1}{M} \sum_{i=1}^{M} v_{ij}
$$

being the average activation of the j-th embedding representation. $v_{ij} \in \mathbf{v}_i$ is the j-th embedding value for the i-th word. The word co-occurrences in SA encode the local context information only. In this paper, we incorporate global topical information into SA and propose TMSA$_{word}$, the sparse autoencoder sparsified with topic modeling, to improve the word embeddings. Our aim is to encapsulate topical information into the overall cost function of SA so that the learned word embeddings also reflect the topic distributions of words. In order to achieve this goal, we propose to add the fourth term as a topic guidance term and the goal of TMSA$_{word}$ is to minimize the following objective function:

$$\mathcal{L}_{word}(\Phi, \mathbf{b}) = \frac{1}{M}\sum_{i=1}^{M}\frac{1}{2}||\hat{\mathbf{x}}_i - \mathbf{x}_i||^2 + \frac{\lambda}{2}||\Phi||^2 + \xi\sum_{j=1}^{N} KL(\rho||\hat{\rho}_j) + \tau\sum_{k=1}^{K} KL(\varphi||\hat{\varphi}_k), \tag{4.8}$$

where φ is the topic sparsity parameter for the embeddings and τ is the weight of the topic guidance term in the overall objective function. $\hat{\varphi}_k$ is the average activation of the embeddings for the k-th topic:

$$\hat{\varphi}_k = \frac{1}{MD_t}\sum_{i=1}^{M}\sum_{j=1}^{D_t}||h^i_{jk}||^2, \tag{4.9}$$

where $h^i_{jk} \in \mathbf{h}_i \in \mathbb{R}^{N\times K}$ is the topic distribution of the $j-$th embedding value over the k-th topic for the i-th word.

$$\mathbf{h}_i = \mathbf{v}_i\mathbf{x}_i^{\top}\mathbf{z}_i$$

is the topic distribution for the embedding $\mathbf{v}_i$ and $h_i \in H$, the total of h_i. The topic guidance term is designed to help the learned embeddings $\mathbf{v}$ reflect the global topical information of words. Here the KL divergence $KL(\varphi||\hat{\varphi}_k)$ helps reconstruct the input with the activations that are related to the most discriminative topics.

4.3 Full Architecture With the semantic word embedding information extracted from local contexts, we can better discover topics from texts; and by exploiting the global topical information, topic-related information will be effectively reflected in word embeddings. These two processes interact closely with each other to boost the performance of both topic discovery and word embedding. The overall objective function can be defined as

$$\mathcal{L} = \arg\min_{f} \mathcal{L}_{topic} + \mathcal{L}_{word}. \tag{4.10}$$

We first fix word embeddings in TMSA$_{word}$ to update topic modeling TMSA$_{topic}$. With the updated topics, we then run TMSA$_{word}$ to learn better word embeddings. This iterative process continues until converge is achieved. The whole procedure is illustrated in Algorithm 1. The proposed TMSA has several advantages. Firstly, parameter tuning and inferences can be done in a unified framework. Secondly, the mechanism of back propagation in Sparse autoencoder can be utilized for fine tuning word embeddings. Thirdly, extra layers can be easily added to handle other tasks, such as adding a softmax layer for classifications. In Algorithm 1, the complete mutual learning procedure is summarized. The input includes word co-occurrence matrix and topic Dirichlet prior α. After initializations of needed weights and embedding matrices, topic modeling and word encoder are updated in return until the topic difference is smaller than the pre-defined ϵ or the given epoch number is reached.

Algorithm 1 The mutual learning algorithm TMSA for topic modeling and word embedding

```
Input:    $\mathbf{D}$, $\mathbf{X}$, $\alpha$
Initialization: $\mathbf{Z}$, $\mathbf{T}$, $\mathbf{A}$, $\mathbf{c}$, $\mathbf{\Phi}$, $\mathbf{b}$
while topic difference $< \epsilon$
or iteration $< total\ epoch\ number$
    /* topic modeling step: */
    1. update $\theta$ with $\alpha$ and $\mathbf{T}$
    2. update $\mathbf{Z}$ with $\theta$
    3. update $p(\mathbf{w_d})$ with $\mathbf{T}$, $\mathbf{A}$, $\mathbf{V}$ and $\mathbf{Z}$
    4. calculate negative loglikelihood
    /* word encoder step: */
    5. encode $\mathbf{X}$
    6. update $\mathbf{\Phi}$ and $\mathbf{c}$
    7. calculate $\mathbf{H}$
    8. update $\hat{\rho}$ with $\mathbf{\Phi}$ and $\mathbf{c}$
    9. update $\hat{\varphi}$ with $\mathbf{H}$
    10. calculate loss function
    11. update $\mathbf{\Phi}$ with backpropagation
    12. update $\mathbf{V}$ with $\mathbf{\Phi}$
end while
```

5 Experiments

In this section, we evaluate the effectiveness of our proposed framework TMSA from both the topic discovery task and the word embedding task.

5.1 Datasets We utilize two datasets for the evaluations. One is the 20 Newsgroups[1] and the other one is the Reuters-21578 corpus[2]. The two corpora are referred to as the *20News* and *Reuters* in the following. 20News has 11,311 documents for training and 7,529 for testing. It has 20 different categories. For Reuters, the largest 10 categories are selected for the experiment with 5,770 documents for

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

[2] http://www.nltk.org/book/ch02.html

training and 2,255 for testing. During the data processing step, stop words are removed and all words are converted to lowercase. For the construction of the word co-occurrence matrix for word embedding, we set the context window size as 10. For the topic modeling, the predefined topic difference ϵ and the topic Dirichlet prior α is set to 0.01 and 0.1, respectively. The total epoch number is set to 100. For the parameters of sparse autoencoder, we set both the sparsity level ξ and topic guidance weight τ as 0.1. The sparsity parameter ρ and the topic sparsity parameter η are both set as 0.05. The regularization parameter λ is 0.01. The number of topics are 20 and 10 for 20News and Reuters, respectively. The embedding dimension is set to 50.

5.2 Evaluation on Document Classification Here, we first evaluate how TMSA can benefit downstream applications. We focus on the document classification task and compare with the following topic modeling baselines:

- LDA [4]: the vanilla Latent Dirichlet Allocation;
- LFTM [20]: the Latent Feature Topic Modeling;
- TopicVec [15]: the generative topic embedding method;
- CLM [27]: the Collaborative Language Model.

In addition to the above baselines, we also compare with the state-of-the-art methods that use the learned word representations for document classification. They are:

- PV-DBOW and PV-DM [8]: the Doc2Vec model;
- MeanWV [15]: the mean word embedding of the TopicVec model;
- TWE [16]: the Topical Word Embedding method;
- Gaussian LDA [6]: the Gaussian LDA model;
- TV+MeanWV [15]: the concatenation of TopicVec and MeanWV.

In TWE, Gaussian LDA and TV+MeanWV, both topic representations and word embeddings of a document are concatenated as features for classification. In TMSA, we aggregate the word embeddings and use the mean as document features since the topical information has already been incorporated into the learned word embeddings. In the experiment, the macro-average precision, recall and F1 measures are used as the evaluation metrics. For LDA, LFTM, CLM, PV-DBOW, PV-DM, TWE and Gaussian-LDA, we use the same results reported in [27]. For TopicVec, MeanWV and TV+MeanWV, we report the same results from [15].

The performance on 20News and Reuters are shown in Table 1 and Table 2, respectively. The best results are highlighted in boldface. It can be observed that TMSA outperforms the compared methods significantly on both datasets. Compared to the second best method CLM, TMSA achieves 2.5% and 4.3% higher on Fscore for 20News and Reuters, respectively. As mentioned in [23], STE is proposed to learn topic-specific word embeddings to avoid the issue of polysemy. It is reported in [23] that STE achieves 82.5% of precision, 82.3% of recall and 82.5% of Fscore on 20News. There is no available result of STE on Reuters. We can see that TMSA still outperforms STE on 20News. In summary, the proposed TMSA combines the topic modeling and word embedding components via a mutual learning mechanism and achieves the best performance on both datasets.

Table 1: Document classification on the 20News dataset. The best results are highlighted in bold.

	Precision	Recall	Fscore
LDA	72.7%	72.2%	71.9%
LFTM	71.6%	71.4%	70.9%
TopicVec	71.3%	71.3%	71.2%
CLM	82.5%	81.8%	81.6%
PV-DBOW	51.0%	49.1%	45.9%
PV-DM	42.8%	38.6%	36.1%
MeanWV	70.4%	70.3%	70.1%
TWE	52.5%	46.6%	43.7%
Gaussian-LDA	30.9%	26.5%	22.7%
TV+MeanWV	71.8%	71.5%	71.6%
TMSA	**85.7%**	**83.7%**	**84.1%**

Table 2: Document classification on Reuters dataset.

	Precision	Recall	Fscore
LDA	88.8%	87.0%	87.9%
LFTM	89.3%	59.1%	66.1%
TopicVec	92.5%	92.1%	92.2%
CLM	94.4%	91.6%	92.9%
PV-DBOW	75.5%	50.5%	54.9%
PV-DM	68.1%	43.4%	50.7%
MeanWV	92.0%	89.6%	90.5%
TWE	79.4%	51.2%	62.6%
Gaussian-LDA	46.2%	31.5%	35.3%
TV+MeanWV	92.2%	91.6%	91.6%
TMSA	**97.3**%	**97.2**%	**97.2**%

5.3 Evaluation on Word Similarity Next, we evaluate the quality of word embedding learned from 20News, to illustrate the effectiveness of the proposed TMSA framework. Since 20News is a small corpus compared with the largest online encyclopedia Wikipedia, it is challenging to collect a large amount of local context information. By encoding the global topical information into the sparse autoencoder with local contexts as a kind of complementary information, the proposed TMSA can improve the word representation learning process significantly even for small corpora.

In this section, we compare with several word embedding baselines, including Skip-Gram and CBOW in [18], GloVe [21], SPPMI and SPPMI+SVD in [11], PV-DBOW and PV-DM in [8], TWE [16] and CLM [27]. We use word embeddings learned from all these methods to evaluate the word pair similarities on several datasets. These datasets include WordSim353 (WS353) [24], WordSim Relatedness (WS Rel) [1], Turk [22], simLex-999 [7] and Rare [17]. We test the performance of word embeddings by measuring the Spearman's correlation of the cosine similarities of word embeddings and the human-assigned similarities. The code for the word similarity evaluation is publicly available[3]. We run it to measure the performance of the proposed TMSA model on the task of word similarity. For all the baseline methods, we report the results from [27].

Table 3: Comparison of word similarity results.

	WS353	WS Rel	Turk	SimLex-999	Rare
SPPMI	0.461	0.444	0.551	0.131	0.245
SPPMI +SVD	0.451	0.435	0.489	0.166	0.349
GloVe	0.300	0.279	0.268	0.049	0.230
Skim-Gram	0.492	0.479	0.512	0.155	0.407
CBOW	0.488	0.451	0.529	0.151	0.407
PV-DBOW	0.477	0.442	0.488	0.139	0.285
PV-DM	0.297	0.304	0.339	0.013	0.157
TWE	0.317	0.231	0.260	0.084	0.184
CLM	0.526	0.486	0.525	0.189	0.411
TMSA	**0.551**	**0.531**	**0.586**	**0.261**	**0.591**

Table 3 shows the results of word similarities. Higher values indicate that the learned embeddings are closer to the human judgments on the word similarity task. We observe that TMSA outperforms all baseline methods on all datasets. Although CLM also performs well on these datasets, it cannot beat TMSA as it does not encode the topical information explicitly into the word representation learning process.

5.4 Qualitative Analysis In this section, we present two case studies to show the quality of generated topics and word embeddings as well as the correlations between them.

5.4.1 Qualitative Assessment of Topic Modeling This subsection provides examples of how the proposed framework improves topic coherence. Table 4 compares the top words produced by TopicVec and TMSA for four topics. TopicVec [15] is one of the state-of-the-art method for topic discovery. In Table 4, for Topic 1 both TopicVec and TMSA produce words which share clear and similar themes (religion for Topic 1). But for Topic 2, Topic 3 and Topic 4, TMSA finds more meaningful words than TopicVec. In TMSA, Topic 2 is about email communications, Topic 3 is language related and Topic 4 is more related to industries. In contrast, TopicVec discovers fewer meaningful words related to these three topics. The words in TopicVec are not that coherent. This shows that TMSA has more powerful capacity of generating topics with interpretable themes.

5.4.2 Qualitative Assessment of Word Embedding Here, we qualitatively assess word embeddings from two perspectives. First, we test the performance of word embeddings on the task of word analogy. Word analogy aims at measuring whether word embedding can cluster word/phrase pairs of similar relations together. Given four words "a", "b", "c" and "d", the usual format for such analogy is "a is to b" as "c is to d", where "d" is hidden and needs to be inferred from the vocabulary. "d" can be inferred by optimizing 3CosAdd [10] as $\operatorname{argmin}_{d \in V}(\cos(d, c - a + b))$. In this subsection, we use the Google dataset [18] to test the quality of the word embeddings learned from TMSA on 20News. The Google dataset contains syntactic analogies such as "good is to better as rich is to richer" and semantic analogies such as "king is to queen as man is to woman".

Table 5 shows the top five analogies for the word analogy task discovered from 20News by ranking the optimized 3CosAdd value in a descending order. The last column is the optimized 3CosAdd value for each word analogy question. It can be observed that TMSA cannot only discover the syntactic analogies such as "flying is to flew as playing is to played", but also find the semantic analogies such as "husband is to wife as father is to mother".

In addition to examples of word analogy, we also present a figure of a two-dimensional PCA projection of word embedding clusters as in Figure 2. Words which have higher scores than a threshold are selected to represent a cluster of related word embeddings. Five clusters with distinct themes can be observed, roughly as, religion, manufacturing, astronomy, computer-related and electronic. Further, the locations of those five themes in the embedding space are not random either. Computer-related and electronic are closer and located on the above while manufacturing, religion and astronomy are closer and located on the below. Those word embedding clusters are evidently affected or guided by the topic words generated from TMSA$_{topic}$. Similar words can be observed from topics generated in TMSA$_{topic}$ in Table 4. Topic 1 and Topic 4 correspond to religion and manufacturing respectively. In addition, topics about space sciences, astronomy and computers can be observed in the output of TMSA$_{topic}$ too. It shows that the mutual learning is working effectively in TMSA.

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

Table 4: Comparisons of topics generated between TopicVec and TMSA, with the most relevant words for four topics.

Topic	Method	Word					
Topic 1	TopicVec	God	Jesus	Bible	Christ	Christian	Church
	TMSA	God	Jesus	Christian	Religion	Truth	Faith
Topic 2	TopicVec	Email	Trash	Address	Sell	Send	Geek
	TMSA	Email	Shipping	Address	Reply	Send	Mail
Topic 3	TopicVec	Dictionary	Lemieux	Language	gainey	Nyr	Det
	TMSA	Thesaurus	Grammar	Encyclopedia	Dictionaries	Idioms	Synonyms
Topic 4	TopicVec	Sectors	Clair	Garden	Eau	Ashland	Unmarked
	TMSA	Procurement	Manufactures	Agenices	Sector	Escrow	Management

Table 5: Examples from 20News for word analogy. The top 5 word pairs are shown.

	(a, b)	(c, d)	3CosAdd
1	(Stockholm, Sweden)	(Helsinki, Finland)	0.978
2	(scream, screaming)	(listen, listening)	0.972
3	(jumping, jumped)	(playing, played)	0.970
4	(flying, flew)	(playing, played)	0.965
5	(husband, wife)	(father, mother)	0.964

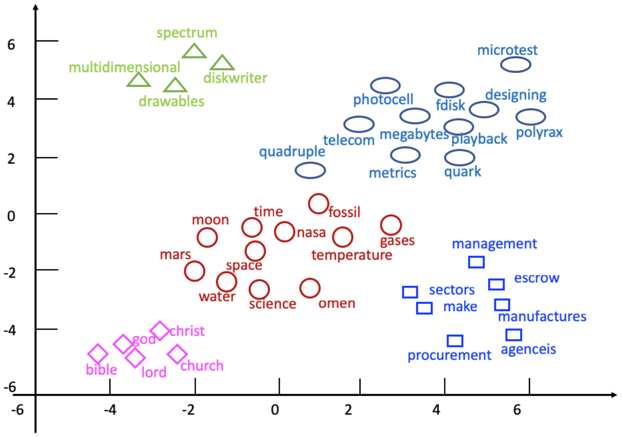

Figure 2: 2D PCA projection of word embeddings. Five different word clusters are shown.

6 Conclusions

This work proposes a mutual learning model TMSA for global topic discovery and local word embedding. In TMSA, the topic discovery component TMSA$_{topic}$ learns topics for input word co-occurrence. The learned word topics are then passed to TMSA$_{word}$ to add topic sparsity to enhance the construction of count-based word embedding. In return, word embeddings are passed back to TMSA$_{topic}$ to improve topic discovery. The experimental results show that both topics and word embeddings demonstrate better performances.

In future, more theoretical studies will be made to explore the optimized integration between autoencoder, topic modeling and word embedding. For example, besides the parametric model based on LDA, we may consider the non-parametric model, such as hierarchical Dirichlet process [25]. Secondly, topics of documents and embeddings can be jointly learned to help boost the document classification performance. Another direction is to explore the integration of knowledge graph into topic modeling. Through the joint learning process, more interesting discoveries will be made on the associations between topic generations, word representation learning and knowledge graph embedding.

References

[1] Eneko Agirre, Enrique Alfonseca, Keith B. Hall, Jana Kravalova, Marius Pasca, and Aitor Soroa. A study on similarity and relatedness using distributional and wordnet-based approaches. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, May 31 - June 5, 2009, Boulder, Colorado, USA*, pages 19–27, 2009.

[2] Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. A neural probabilistic language model. *Journal of Machine Learning Research*, 3:1137–1155, 2003.

[3] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. In *Advances in Neural Information Processing Systems 19, Proceedings of the Twentieth Annual Conference on Neural Information Processing Systems, Vancouver, British Columbia, Canada, December 4-7, 2006*, pages 153–160, 2006.

[4] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[5] Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel P. Kuksa. Natural language processing (almost) from scratch. *Journal of Machine Learning Research*, 12:2493–2537, 2011.

[6] Rajarshi Das, Manzil Zaheer, and Chris Dyer. 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 of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31,*

2015, Beijing, China, Volume 1: Long Papers, pages 795–804, 2015.

[7] Felix Hill, Roi Reichart, and Anna Korhonen. Simlex-999: Evaluating semantic models with (genuine) similarity estimation. *Computational Linguistics*, 41(4):665–695, 2015.

[8] Quoc V. Le and Tomas Mikolov. Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014*, pages 1188–1196, 2014.

[9] Rémi Lebret and Ronan Collobert. Rehabilitation of count-based models for word vector representations. In *Computational Linguistics and Intelligent Text Processing - 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part I*, pages 417–429, 2015.

[10] Omer Levy and Yoav Goldberg. Linguistic regularities in sparse and explicit word representations. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning, CoNLL 2014, Baltimore, Maryland, USA, June 26-27, 2014*, pages 171–180, 2014.

[11] Omer Levy and Yoav Goldberg. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 2177–2185, 2014.

[12] Omer Levy, Yoav Goldberg, and Ido Dagan. Improving distributional similarity with lessons learned from word embeddings. *TACL*, 3:211–225, 2015.

[13] Dingcheng Li, Peini Liu, Ming Huang, Yu Gu, Yue Zhang, Xiaodi Li, Daniel Dean, Xiaoxi Liu, Jingmin Xu, Hui Lei, and Yaoping Ruan. Mapping client messages to a unified data model with mixture feature embedding convolutional neural network. In *2017 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2017, Kansas City, MO, USA, November 13-16, 2017*, pages 386–391, 2017.

[14] Dingcheng Li, Jingyuan Zhang, and Ping Li. Representation learning for question classification via topic sparse autoencoder and entity embedding. *IEEE Big Data*, 2018.

[15] Shaohua Li, Tat-Seng Chua, Jun Zhu, and Chunyan Miao. Generative topic embedding: a continuous representation of documents. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*, 2016.

[16] Yang Liu, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. Topical word embeddings. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, pages 2418–2424, 2015.

[17] Thang Luong, Richard Socher, and Christopher D. Manning. Better word representations with recursive neural networks for morphology. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning, CoNLL 2013, Sofia, Bulgaria, August 8-9, 2013*, pages 104–113, 2013.

[18] Tomas Mikolov, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems 26: 27th Annual Conference on Neural Information Processing Systems 2013. Proceedings of a meeting held December 5-8, 2013, Lake Tahoe, Nevada, United States.*, pages 3111–3119, 2013.

[19] Andriy Mnih and Geoffrey Hinton. Three new graphical models for statistical language modelling. In *International conference on Machine learning*, pages 641–648. ACM, 2007.

[20] Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. Improving topic models with latent feature word representations. *TACL*, 3:299–313, 2015.

[21] Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[22] Kira Radinsky, Eugene Agichtein, Evgeniy Gabrilovich, and Shaul Markovitch. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th International Conference on World Wide Web, WWW 2011, Hyderabad, India, March 28 - April 1, 2011*, pages 337–346, 2011.

[23] Bei Shi, Wai Lam, Shoaib Jameel, Steven Schockaert, and Kwun Ping Lai. Jointly learning word embeddings and latent topics. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, Shinjuku, Tokyo, Japan, August 7-11, 2017*, pages 375–384, 2017.

[24] Fabrice Souvannavong, Bernard Mérialdo, and Benoit Huet. Improved video content indexing by multiple latent semantic analysis. In *Image and Video Retrieval: Third International Conference, CIVR 2004, Dublin, Ireland, July 21-23, 2004. Proceedings*, pages 483–490, 2004.

[25] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Sharing clusters among related groups: Hierarchical dirichlet processes. In *Advances in Neural Information Processing Systems 17 [Neural Information Processing Systems, NIPS 2004, December 13-18, 2004, Vancouver, British Columbia, Canada]*, pages 1385–1392, 2004.

[26] Pascal Vincent, Hugo Larochelle, Yoshua Bengio, and Pierre-Antoine Manzagol. Extracting and composing robust features with denoising autoencoders. In *Machine Learning, Proceedings of the Twenty-Fifth International Conference (ICML 2008), Helsinki, Finland, June 5-9, 2008*, pages 1096–1103, 2008.

[27] Guangxu Xun, Yaliang Li, Jing Gao, and Aidong Zhang. 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, Halifax, NS, Canada, August 13 - 17, 2017*, pages 535–543, 2017.

[28] Guangxu Xun, Yaliang Li, Wayne Xin Zhao, Jing Gao, and Aidong Zhang. A correlated topic model using word embeddings. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, pages 4207–4213, 2017.

[29] He Zhao, Lan Du, Wray L. Buntine, and Mingyuan Zhou. Inter and intra topic structure learning with word embeddings. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 5887–5896, 2018.