Improving Distributed Word Representation and Topic Model by Word-Topic Mixture Model

Xianghua Fu fuxh@szu.edu.cn
Ting Wang wangtingwtc@163.com
Jing Li muziqingqing@yahoo.com
Chong Yu yuchong1001@163.com
Wangwang Liu tochange@163.com
College of Computer Science and Software Engineering, Shenzhen University, Guangdong China

Editors: Robert J. Durrant and Kee-Eung Kim

Abstract

We propose a Word-Topic Mixture(WTM) model to improve word representation and topic model simultaneously. Firstly, it introduces the initial external word embeddings into the Topical Word Embeddings(TWE) model based on Latent Dirichlet Allocation(LDA) model to learn word embeddings and topic vectors. Then the results learned from TWE are integrated in the LDA by defining the probability distribution of topic vectors-word embeddings according to the idea of latent feature model with LDA (LFLDA), meanwhile minimizing the KL divergence of the new topic-word distribution function and the original one. The experimental results prove that the WTM model performs better on word representation and topic detection compared with some state-of-the-art models.

Keywords: topic model, distributed word representation, word embedding, Word-Topic Mixture model

1. Introduction

Probabilistic topic model is one of the most common topic detection methods. Topic modeling algorithms, such as LDA(Latent Dirichlet Allocation) Blei et al. (2003) and related methods Blei (2011), infer probability distributions from frequency statistic, which can only reflect the co-occurrence relationships of words. The semantic information is less considered.

Recently, distributed word representations with NNLM(neural network language model) Bengio et al. (2003) have shown wonderful performances for NLP(natural language processing) and ML(machine learning) tasks Collobert and Weston (2008); Huang et al. (2012). But the existing word embedding framework, such as word2vec Mikolov et al. (2013a) and glove toolbox Pennington et al. (2014), only exploits a slide window context to predict the target word, which is insufficient to capture semantic, especially dealing with small corpus.

Many researches explored constructing topic models by latent feature representations of words Salakhutdinov and Hinton (2009); Cao et al. (2015). They achieve great improvement. But the quality of word embedding has a significant influence on topic-word mapping. Most word embedding methods get word embedding from external corpora, which is inaccurate for word expression, and words that are not included in external word embedding are ignored.

In this paper, we propose a word-topic mixture(WTM) model based on LDA Blei et al. (2003) for improving word representation and topic model simultaneously. On one hand, we use the idea of TWE model Liu et al. (2015). We first learn the word-topic assignment from LDA, and introduce external corpus to capture inital word embeddings. Then we train word embeddings and topic vectors. On the other hand, according to the idea of latent leature model with LDA(LFLDA) Nguyen et al. (2015), we integrate the probability distribution of topic vectors-word embeddings from TWE with topics-words probability distribution from LDA. We redefine new objective function using KL divergence to learn word embeddings and train topic model, and then we mine the latent topic.

2. Related works

Most topic detections are expanded based on probabilistic topic model and the basic structure is LDA Blei et al. (2003). LDA(Latent Dirichlet Allocation) is a three-Bayesian probabilistic topic model, including document, topic and word three levels. The probability of topic is calculated by word frequency statistics, which causes not ideal topic recognition rate without enough semantic information. Later, many scholars have proposed the using of external corpus, such as Wikipedia, Baidu Wikipedia, etc. to the semantic extension.

Distributed word representation is recently proposed as word vector representation and also called word embedding Turian et al. (2010). Word embedding represents a word as a dense, low-dimensional and real-valued vector. Each dimension contains a certain amount of semantic information. This is a very simple and efficient vector representation. Neural network language model Bengio et al. (2003) is firstly proposed by Bengio in 2003. It uses a four-layer structure of statistical language model to automatically learn word embedding representation that contains certain word meaning. Word2vec Mikolov et al. (2013a) developed by Tomas Mikolov team of Google in 2013 is one of the most widely used word embedding tools currently. It improves training efficiency and learns high-quality word representation by simplifying the internal structure of the NNLM. It removes hidden layer which is complicated and time-consuming. The projection and output layer are also optimized in the training process, which makes Word2vec trains more flexibly and efficiently.

Each word is represented as a single vector in most word embedding methods. But a word have multiple senses in practical. Some scholars proposed multi-prototype word embedding models Huang et al. (2012). Yang Liu et al. introduced word embeddings to multiple prototype probabilistic topic model in 2015. They proposed word embedding model TWE(Topical Word Embedding) Liu et al. (2015) based on LDA. The final word-topic assignment gotten from LDA is an auxiliary input. The word embedding-topic vector is learned by taking the corresponding topic information of the target word and context information into account together based on the original Skip-Gram of Word2vec.

Dat Quoc Nguyen et al. proposed latent feature model with LDA Nguyen et al. (2015) in 2015, referred to as LFLDA. The feature vector representations of words are trained on a large corpora. The model considers word embeddings and word frequency statistics to detect the topic. Word embeddings can capture both semantic and syntactic information of words and gather semantically similar words. The words frequency statistics are based on the co-occurrences. Generally speaking, if the times of words co-occurrences are more in the same corpus, the probabilities of that they appear in the same topic are bigger. So it can better solve the defect that traditional method calculates the probability distribution relying solely on word frequency without considering the internal semantic relationships. However, the word embedding is fixed because it uses pre-trained word vectors. The word embedding entirely depends on external corpus. Therefore, it cannot guarantee that words in the training text are highly semantically consistent with the word embedding. Besides, Das et al. (2015) proposed Gaussian LDA model for topic modeling by treating the document as a collection of word embeddings and topics itself as multivariate Gaussian distributions in the embedding space. The model can infer different topics relative to standard LDA and outperforms existing models at dealing with OOV words in held-out documents. Chenliang Li et al. proposed GPU-DMM Li et al. (2016) which extends the Dirichlet Multinomial Mixture(DMM) Yin and Wang (2014) model by incorporating the learned word relatedness provided by auxiliary word embeddings through the generalized Polya urn(GPU) Mahmoud (2008) model in topic inferences of short texts for solving sparsity problem.

We propose a Word-Topic Mixture(WTM) model which combines the ideas of TWE and LFLDA. It introduces the external extended corpus to learn initial word embeddings and also use the word-topic assignment learned from LDA to train word embeddings and topic vectors based on TWE. Rather than relying solely on external word embeddings, as in LFLDA, we combine the topic vectors-word embeddings learned from TWE with the topics-words learned from LDA. We redefine a new objective function by minimizing the KL divergence of the new topic vectors-word embeddings probability distribution and the topics-words probability distribution to train word embedding and topic model simultaneously.

3. Models

3.1. LDA model

Latent Dirichlet Allocation(LDA) Blei et al. (2003) is a three-layer probabilistic graphical model. It represents a document as a mixture of topics and topics as a distribution over words. The structure of LDA is in Fig. 1 and LDA assumes the following generative process.

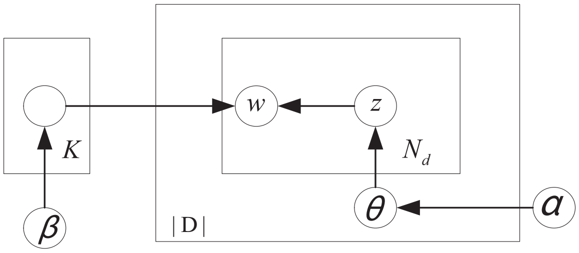

Figure 1: Architecture of LDA model.

(1) for each topic z

i) sample a topic-word distribution $\varphi_z \sim Dir(\beta)$

(2) for each document d in corpus D

i) sample a document-topic distribution $\theta_d \sim Dir(\alpha)$

ii) for each word w_i in the document d, $i \in \{1, 2, 3, ..., N_d\}$

a) choose a topic $z_{d_i} \sim Cat(\theta_d)$

b) choose a word $w_i \sim Cat(\varphi_{z_{d_i}})$

Dir and Cat stand for a Dirichlet distribution and a Categorical distribution. α and β are the Dirichlet hyper-parameters. K is the number of topics. N_d is the number of words in document d . z_{d_i} is the topic of the i^{th} word w_i in the document d .

A Gibbs Sampling algorithm is used to estimate LDA by Griffiths and Steyvers (2004). The Gibbs Sampling algorithm uses the conditional distribution $P(z_{d_i}|\mathbf{z}_{\neg d_i})$ to sample the topic z_{d_i} for the i^{th} word w_i in the document d by integrating out θ and φ. $\mathbf{z}_{\neg d_i}$ is the topic assignments of all words in the D, except the word w_i. There are:

$$P(z_{d_i} = t|\mathbf{z}_{\neg d_i}) \propto (N^t_{d\neg_i} + \alpha) \cdot \frac{N^{t,w_i}_{\neg d_i} + \beta}{N^t_{\neg d_i} + V \cdot \beta} \tag{1}$$

$N^t_{\neg d_i}$ is the number of words assigned to topic t, ignoring the i^{th} word in document d. $N^{t,w_i}_{\neg d_i}$ is the number of times that the word w_i is generated by topic t, except the i^{th} word in document d. $N^t_{d\neg_i}$ stands for the number of words in the topic t in document d without its i^{th} word. V is the size of the vocabulary.

3.2. TWE model

Topical Word Embedding(TWE) Liu et al. (2015) is a flexible model for learning topical word embeddings based on Skip-Gram Mikolov et al. (2013b), meanwhile integrating LDA. In the TWE paper, there are three TWE models: TWE-1, TWE-2 and TWE-3, but the experimental results show that TWE-1 has the best performance due to the separate and simultaneous learning of word and topic embedding. So we only describe TWE-1 in Fig. 2.

In the TWE model, each word w_i has a labeled topic z_i inferred from LDA, forming a word-topic pair $< w_i, z_i >$. It extends Skip-Gram Mikolov et al. (2013b) to implement as shown in Fig. 2, where the window size is c, given a contextual word-topic stream $\{w_{i-c} : z_{i-c}, w_{i-c+1} : z_{i-c+1}, ..., w_{i+c-1} : z_{i+c-1}, w_{i+c} : z_{i+c}\}$. Each topic can be regarded as a pseudo word that contains topic information. It learns topic vectors and word embeddings separately and then build the topical word embedding of $< w_i, z_i >$ according to the embeddings of w_i and z_i. The learning objective function can be defined as follows.

$$L_{TWE} = \frac{1}{V} \sum_{i=1}^{V} \sum_{-c \le j < c, i \ne 0} (\log p(w_{i+j}|w_i) + \log p(w_{i+j}|z_i)) \tag{2}$$

V is the vocabulary size. $p(w_k\,|w_i\,)$ is a softmax function as follows, where $\mathbf{w}_i$ and $\mathbf{w}_k$ are the vector representations of target word w_i and context word w_k, and W is the vocabulary.

$$p(w_k\,|w_i\,) = \frac{\exp(\mathbf{w}_k \cdot \mathbf{w}_i)}{\sum\limits_{w_i \in W} \exp(\mathbf{w}_k \cdot \mathbf{w}_i)} \tag{3}$$

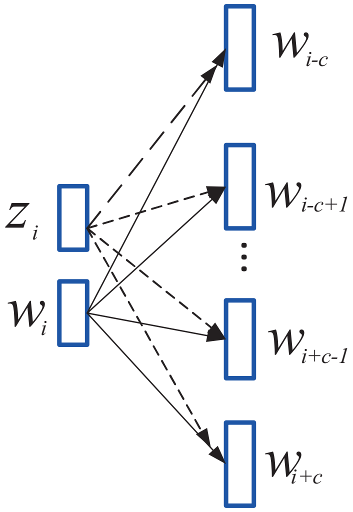

Figure 2: Graphical representation of the TWE model.

TWE learns word embeddings and topic vectors following the same optimization method as that of Skip-Gram in Mikolov et al. (2013b). NS(Negative Sampling) or HS(Hierarchical Softmax) Mikolov et al. (2013b) is approximate to the objective function, updating targets by stochastic gradient descent(SGD) and back-propagation algorithm.

3.3. WTM model

In this section, we propose Word-Topic Mixture(WTM) model for training distributed word representation and topic distribution simultaneously. First, we introduce the external corpus to learn the initial word embedding v'_w and use word-topic assignment learned from LDA in order to train word embeddings v_w and topic vectors τ_z based on TWE. Second, we combine the topic vectors-word embeddings learned from TWE with the topic-word distribution learned from LDA to train word embedding and topic model simultaneously according to the idea of LFLDA Nguyen et al. (2015). The architecture of WTM is in Fig. 3. WTM assumes the following generative process for a document d in the corpus D.

(1) sample a document-topic distribution $\theta_d \sim Dir(\alpha)$

(2) for each word w_i in document d, $i \in \{1, 2, 3, ..., N_d\}$

  i) choose a topic $z_{d_i} \sim Cat(\theta_d)$
  ii) choose the assignment method of topic-word $S_{d_i} \sim Ber(\lambda)$
  iii) generate a word $w_i \sim (1 - S_{d_i}) \cdot Cat\left(\varphi_{z_{d_i}, w_i}\right) + S_{d_i} \cdot Cat\left(\gamma_{\tau_{z_{d_i}}, v_{w_i}}\right)$

S_{d_i} is a binary indicator variable. It is sampled from a Bernoulli distribution to decide whether w_i is generated by the probability distribution of topic-word under vector space representations or the original LDA Blei et al. (2003). λ is the probability of a word generated by the topic-word distribution under vector space representations. $Cat(\theta_d)$ is categorical distribution drawn from a symmetric dirichlet distribution with a prior parameter

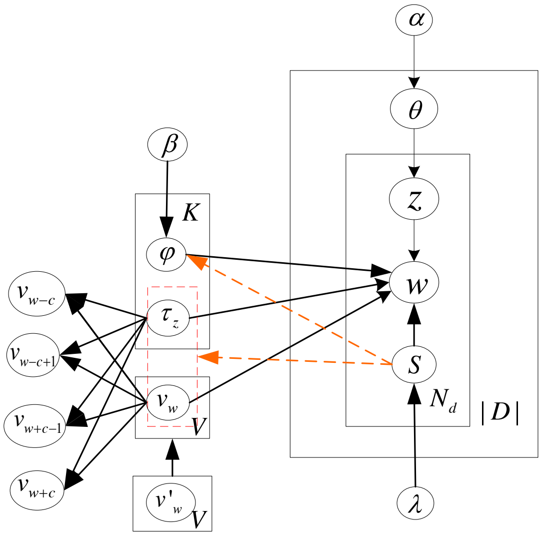

Figure 3: Graphical representation of the WTM model.

α. $Cat\left(\varphi_{z_{d_i},w_i}\right)$ is the topic-word categorical distribution of LDA as described in section 3.1. $Cat\left(\gamma_{\tau_{z_{d_i}},v_{w_i}}\right)$ is the topic-word categorical distribution of vector representation.

In WTM, the probability distribution of topic-word under vector space representations γ_{τ_z,v_w} referred to the LFLDA model Nguyen et al. (2015) and that of the standard LDA $\varphi_{z,w}$ which is the same as section 3.1 are defined as follows. W is the word vocabulary. w is a word in W. N_z^w is the number of times that the word w assigned to topic z, ignoring w in current document. N_z stands for the number of words generated from topic z, except the word w. V is the size of the vocabulary. β is the hyper-parameter.

$$\gamma_{\tau_z,v_w} = \frac{\exp(v_w \cdot \tau_z)}{\sum\limits_{w' \in W} \exp(v_{w'} \cdot \tau_z)} \tag{4}$$

$$\varphi_{z,w} = \frac{N_z^w + \beta}{N_z + V \cdot \beta} \tag{5}$$

In general, the topic-word distribution obtained from the vector space representations is different from that of LDA in expression, but they should be consistent with a criterion: two expressions of documents should be as close as possible to each other on semantic. Therefore, we define a new objective function L_{WTM} with L2 regularization term as follows.

$$L_{WTM} = KL\left(\gamma_{\tau_z,v_w}, \varphi_{z,w}\right) + \mu \left\|v_w\right\|_2^2 \tag{6}$$

In the equation, $KL\left(\gamma_{\tau_z,v_w}, \varphi_{z,w}\right)$ stands for the symmetric KL divergence between distributions γ_{τ_z,v_w} and $\varphi_{z,w}$, μ is the regularization factor.

$$KL\left(\gamma_{\tau_z,v_w}, \varphi_{z,w}\right) = \frac{\sum\limits_{w \in W} \left(\gamma_{\tau_z,v_w} \cdot \log \frac{\gamma_{\tau_z,v_w}}{\varphi_{z,w}} + \varphi_{z,w} \cdot \log \frac{\varphi_{z,w}}{\gamma_{\tau_z,v_w}}\right)}{2} \tag{7}$$

We get the model parameters by minimizing L_{WTM} and v_w can be updated by :

$$\frac{\partial_{L_{WTM}}}{\partial_{v_w}} = \frac{1}{2}\left(1 + \log\frac{\gamma_{\tau_z,v_w}}{\varphi_{z,w}} - \frac{\varphi_{z,w}}{\gamma_{\tau_z,v_w}}\right) \cdot \frac{\exp(v_w \cdot \tau_z) \cdot \tau_z \cdot \left(\sum\limits_{w' \in W} \exp(v_{w'} \cdot \tau_z) - \exp(v_w \cdot \tau_z)\right)}{\left(\sum\limits_{w' \in W} \exp(v_{w'} \cdot \tau_z)\right)^2} + 2\mu v_w \tag{8}$$

Here we use the Gibbs sampling algorithmn Robert and Casella (2004) for posterior inference to compute conditional probabilities. $N^t_{\neg d_i}$ is the number of words assigned to topic t by LDA, except the i^{th} word in document d, and $K^t_{\neg d_i}$ is the same number of words by topic vector-word embedding distribution. The pseudo code is described in Alg. 1.

Algorithm 1 Learning Algorithm for the WTM model.

```
Input: the preprocessed dataset $D$ , topic number $K$ , hyper-parameters $\alpha$ , $\beta$ , the iteration number $N$ ,
    word embeddings $v_w$ , distributed topic representations $\tau_z$ , initial word-topic assignments $\langle w, z \rangle$
Output: word embeddings $v_w$ , word-topic assignments $\langle w, z \rangle$ , document-topic probability matrix, topic-
    word probability matrix
for iteration number =1, 2...$N$ do
    for document $d$ =1, 2...$|D|$ do
        for word index $i$ =1, 2...$N_d$ do
            for topic $t$ =1, 2...$K$ do
                compute $P\left(z_{d_i} = t|\mathbf{z}_{\neg d_i}\right) \propto \left(N^t_{d\neg_i} + K^t_{d\neg_i} + \alpha\right) \cdot \left((1-\lambda) \cdot Cat\left(\varphi_{z_{d_i}, w_i}\right) + \lambda \cdot Cat\left(\gamma_{\tau_{z_{d_i}}, v_{w_i}}\right)\right)$
            end for
            sample topic $z_{d_i}$ for word $w_i$
            update word embeddings $v_{w_i}$ and word-topic assignments $\langle w_i, z_{d_i} \rangle$
        end for
    end for
end for
```

4. Experiments

To investigate the performance of our WTM model on word embeddings learning and topic detection, we compare it with a series of other correlation models, including the Skip-Gram and TWE, LDA and LFLDA. The former is set for word embeddings evaluation. We analyze results with document classification and word similarity tasks. The latter is set for topic model. We evaluate them on topic coherence and document clustering tasks.

4.1. Datasets

4.1.1. External Datasets

We use the Chinese Wikipedia and English Wikipedia corpus, the largest online knowledge base, as external information to improve the initial word embeddings and topic representations for TWE Liu et al. (2015). The Chinese Wikipedia consists of 777,961 documents, about 998.5M tokens. We pre-train the word vectors by using the Google Word2Vec toolkit Mikolov et al. (2013a). Vector dimension is chosen according to the size of training dataset.

4.1.2. Experimental Datasets

We conduct experiments on two different types of dataset. One is the standard benchmark 20-Newsgroups dataset and Reuter-21578 dataset. The other is social media datasets that crawling from Tianya By-talk of Tianya Forum and Sina Blog, which are famous Chinese online Bulletin Board System community.

The 20-Newsgroups dataset consists of 18,828 newsgroup documents evenly grouped into 20 different categories. Each document has only one topic. Reuter-21578 dataset contains 52 different categories, about 9,100 documents. Each report has several attributions, such as title, document id, content and so on. We only use the content as training data.

Tianya-12261 dataset is constructed in 2015 Fu et al. (2015). There are about 72,585 posts. Each post includes text, title, replies and so on. The author discards posts whose replies number is less than 10, meanwhile merging the original posts and their replies into an article for training. After preprocessing, the Tianya-12261 dataset consists of 12,261 different documents. The Sina Blog dataset has 12,979 news reports and comments for 8 topics, which occur in 2008. We extract content by Xpath.

Our experimental Datasets contain Chinese and English. The data processing includes word segment and filtering out high and low frequency words, stop words, improper characters, and single words. Empirically, if word occurs less than 5 times in a corpus, we regard it as a low frequency word. If the frequency of a word occurrence appears more than 20% of the total tokens, then it is treated as a high-frequency word. For the Chinese corpus, we choose the ICTCLAS2016 as the word segment system which can support the user dictionary and find new words. The English datasets don't need the word segment processing. We also convert all characters to lower case and remove non-alphabetic characters. After preprocessing, the distributions of experimental datasets are shown in Tab. 1 and Tab. 2.

Table 1: The distribution of all the experimental datasets.

Datasets	Categories	Train	Test	Total
20-Newsgroups	20	12,385	5,308	1,7693
Reuter-21578	52	8,000	1,100	9,100
Tianya-12261	-	1,0957	1,304	1,2261
Sina Blog	8	1,1680	1,299	1,2979

Table 2: The document distribution of each topic in Sina Blog dataset.

Event	甲流 /cold	火车实名制 /train	酒后驾驶 /drink	农民工返乡 /farmer	三鹿奶粉 /milk	西南大旱 /drought	房价 /house	石油 /oil
Posts number	2,283	932	2,289	563	1,425	1,499	2,153	2,074

4.2. Parameter Setting

The Java package for the LDA and DMM(Dirichlet Multinomial Mixture)(jLDADMM) topic models is used to get the initial topic-word assignment. Hyper-parameters of the topic model α and β are 0.1 and 0.01, which is a common setting in Griffiths and Steyvers (2004). The iteration number is 1500. We evaluate the topics assigned to words in the last sample. To train the word embeddings and distributed topic vectors, the parameters are as follows: initial learning rate=0.025, window size=10(five front and five future words), HS (Hierarchical Softmax)=1, sample is le^{-3}, 12-threads, binary=0. The dimension of vectors is different to different corpus and tasks. In the WTM model, we learn word embeddings and topic model by running 500 further iterations.

4.3. Evaluation Methods and Analysis of Results

4.3.1. Word Embeddings Evaluation

We examine the quality of word embeddings induced by WTM on Multi-class document classification and word similarity tasks under two types of corpus. We use the famous 20-Newsgroup dataset as the English dataset and Sina Blog dataset as the Chinese dataset.

In our experiment, we compare the results of document classification on WTM with the following baseline models, LDA, Skip-Gram, PV-DM(paragraph vector-distributed memory), PV-DBOW(paragraph vector-distributed bag-of-words) Le and Mikolov (2014) and TWE, and the results of other methods on 20-Newsgroup dataset are shown in Liu et al. (2015). The dimensions word embeddings and topic embeddings are 400. Topic number of 20-Newsgroup dataset is 80, which is the same setting as in Liu et al. (2015). The Sina Blog is a dataset which is labeled by topics. So the topic number is category number. We set it to 8. Performances of each model are listed in Tab. 3.

Table 3: Results of document classification on different models.

Datasets	Models	accuracy
20-Newsgroup	LDA	72.20%
	Skip-Gram	75.40%
	PV-DM	72.40%
	PV-DBOW	75.40%
	TWE	78.55%
	WTM	**80.94%**
Sina Blog	LDA	74.72%
	Skip-Gram	72.01%
	PV-DM	71.62%
	PV-DBOW	73.26%
	TWE	71.70%
	WTM	**76.41%**

We can see that WTM outperforms all baselines significantly on the two datasets. For the 20 Newsgroup dataset whose topic number is 80, our method achieves an accuracy rate of 80.94%, which yields an improvement approximately 2.5% over the best result of other models(here is TWE Liu et al. (2015)). In the Sina Blog dataset whose topic number is 8, we also get the best performance on the WTM model at an accuracy of 76.41% in all models. The classification results of the two datasets on the WTM is better than the other models. It shows that on the feature representation of social media data, WTM can indeed learn a word with richer semantic information. In addition, as for the very difference results of TWE on two datasets, it is because that the accuracy is related to the number of topics and corpus. The structure of 20-Newsgroup is organized and has 80 topics, but Sina blog grabbed from social media corpus is casual and has 8 topics.

Except document classification, we also evaluate word embeddings on word similarity tasks. As we known, word embeddings trained by the Neural Probabilistic Language Model Bengio et al. (2003) can gather words with similar semantic to close embedding space. So we can use cosine similarity measure to compute the similarity words for each target word. Tab. 4 lists four topic words and their top-10 most similarity words on the Sina Blog dataset.

Table 4: Examples of the 10 most similarity words for four topic words on Sina Blog dataset.

Target word	Models	Similarity Words
流感/flu	**Skip-Gram**	甲型/A, 异同/distinct, H1N1/H1N1, 感染/infected, 病毒/viruses, 秋冬季/Autumn and winter, 禽流感/influenza, 慕盛学/Mu Shengxue, HN/HN, 刘政/Liu Zhen
	TWE	甲型/A, 疫情/epidemic, 暴发/break out, 没事/all right, 治疗学/therapeutics, 阴影/shadow, 陈国芳/Chen Guofang, 你一言我一语/everybody chimes in, 成百/hundreds, H1N1/H1N1
	WTM	甲型/A, 感染/infect, 病毒/viruses, 疫情/epidemic, H1N1/H1N1, 患者/patient, 流行/prevalence, 预防/prevention, 病例/case, 感冒/cold
地沟油/ hogwash oil	**Skip-Gram**	餐桌/table, 砒霜/arsenic, 湘菜/Hunan CAI, 炼油厂/refinery, 炼制/refine, 回流/backflow, 三百万/three million, 地沟/trench, 餐饮业/catering, 百倍/centuplicate
	TWE	地沟/gutter, 餐馆/restaurant, 餐桌/table, 砒霜/arsenic, 危害/harm, 泔水/swill, 餐饮业/catering, 回收/recycle, 饭店/hotel, 废油/waste oil
	WTM	餐桌/table, 食品/ food, 餐馆/restaurant, 地沟/gutter, 黑心/greedy, 饭店/hotel, 砒霜/arsenic, 泔水/swill, 回流/reflux, 垃圾/garbage
农民工/migrant workers	**Skip-Gram**	民工/migrant worker, 返乡/return, 输出地/exporter, 外省/provinces, 务工/worker 务工人员/workers, 陆川县/Lu Chuan County, 就业/employment, 产业工人/industrialist, 就学/go to school
	TWE	返乡/return, 打工者/worker, 农工/laborer, 民工/migrant worker, 陆川/Lu Chuan, 打工/work, 外出/go out, 务农/farming, 结队/troop, 陈凡顺/Chen fanshun
	WTM	民工/migrant worker, 返乡/return, 就业/employment, 打工/work, 农村/ rural, 失业/unemployment, 农民/farmer, 创业/entrepreneurship, 劳动/ labor, 民工荒/labor shortage
三鹿/ SANLU	**Skip-Gram**	奶粉/milk, 襄汾/Xiang Feng, 三聚氰胺/melamine, 狗宝/Gou bao, 美赞臣/Mead Johnson, 婴幼儿/baby, 肾结石/renal calculus, 毒奶粉/tainted-milk, 戕害/harm, 冰海/ice
	TWE	奶粉/milk, 毒奶粉/tainted-milk, 三聚氰胺/melamine, 冰海凌峰/Like Ling Feng, 襄汾/Xiangfen, 毒酒/poison, 事件/event, 词典/dictionary, 毒大米/poisoned rice, 林海峰/Haifeng Lin
	WTM	奶粉/milk, 事件/event, 山西/Shanxi, 毒奶粉/tainted-milk, 里面, /inner, 老板/boss, 孩子/children, 宝宝/baby, 几乎/almost, 良心 /conscience

Tab. 4 shows top 10 nearest words from the Skip-Gram, TWE and WTM models for 4 different topic words. As we can see that all of these models have the ability of gathering words that have similar semantics into the same topic space, which also proved that word embeddings with NNLM did have a good performance on semantics learning. The Skip-Gram model is more inclined to gather together the words with similar meanings. As for topic word ” 地沟油/hogwash oil”, Skip-Gram also learns some useless words such as ” 三百万/three million”, ” 百倍/centuplicate”. While the TWE model also learned some words that are related to the event, such as ” 疫情/epidemic”, ” 暴发/break out” under topic word ” 流感/flu”, ” 餐馆/restaurant”, ” 危害/harm”, ” 回收/recycle” under topic word ” 地沟油/hogwash oil”, and ” 打工/work”, ” 外出/go out”, ” 结队/troop” under topic word ” 农民工/migrant workers”. However, there are many useless words mixed, such as ” 没事/all right”, ” 你一言我一语/everybody chimes in” under topic word ” 流感/flu”, ” 冰海凌峰/Like Ling Feng”, ” 词典/dictionary” under topic word ” 三鹿/ SANLU”. Besides, the results of TWE and WTM are similar for topic word ” 地沟油/hogwash oil”. It may be because we give only 10 most similar words. We can find that WTM is better than TWE from the results of other three target words, so it won't influence the overall experiment results. In general, compared with other two methods, most of words gotten from WTM are semantically related to the target word, and in terms of the word similarity list, these ten words are most related to the topic events.

4.3.2. Topic Model Evaluation

To evaluate performance of our WTM model on topic-word mappings, we set experiments with 20-Newsgroups, Reuter-21578 and Sina Blog datasets and compare them to the standard LDA and LFLDA model in topic coherence and document clustering tasks.

We report the results of 20-Newsgroups on two metrics: Purity and NMI (normalized mutual information) Manning et al. (2008) with topic number K =6, 20, 40, 80, embedding size 300, and adjustment factor λ =0.6(best value), which are the same as in LFLDA Nguyen et al. (2015). Fig. 4 presents Purity and NMI scores under three models. It shows that WTM outperforms other models. The purity and NMI are the highest scores when topic number is 40 and 20. WTM is also consistent with other two methods.

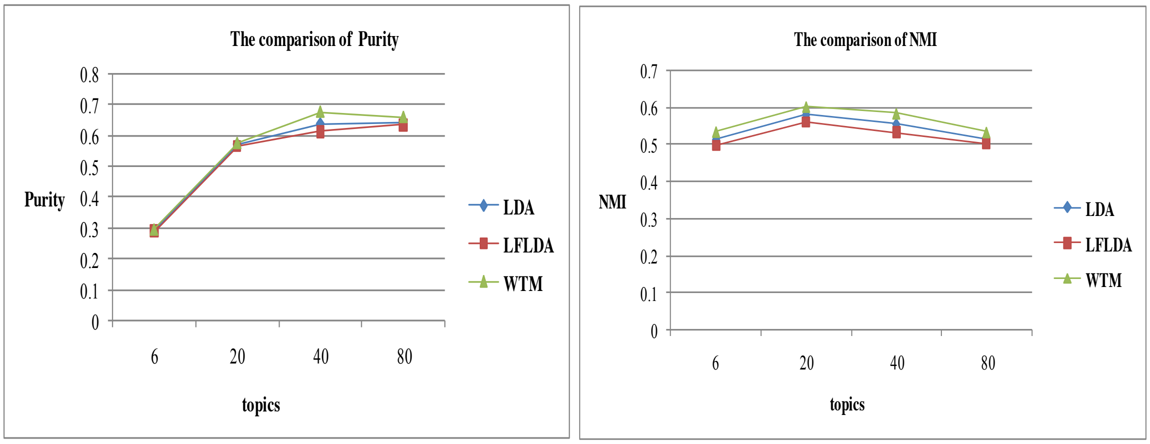

Figure 4: The Purity and NMI scores for 20-Newsgroups.

Tab. 5 shows clustering results produced by the three models on all datasets. Word vector size is 300. λ is 0.6. Topic number is set to the category number. That is 20 for 20-Newsgroups dataset, 52 for Reuter-21578 dataset and 8 for Sina Blog dataset.

Table 5: Purity, NMI and Perplexity results on all datasets.

Datasets	Models	Purity	NMI	Perplexity
Reuter-21578	LDA	0.79	0.49	392.91
	LFLDA	0.80	0.51	416.69
	WTM	**0.82**	**0.53**	**379.96**
20-Newsgroups	LDA	0.52	0.56	3499.07
	LFLDA	0.57	0.56	3487.15
	WTM	**0.58**	**0.60**	**2189.65**
Sina Blog	LDA	0.66	0.52	3834.55
	LFLDA	0.67	0.54	4060.33
	WTM	**0.68**	**0.55**	**3698.31**

As expected, the WTM is better than other two models, LDA and LFLDA, on all datasets. In terms of Purity and NMI scores, WTM achieves about 2% improvement than LFLDA on the Reuter-21578 dataset, and more than 1% improvement on other two datasets. However, the WTM model gets about 3%-5% higher than LDA on all datasets. The performance on perplexity illustrates that improvement of topic-word mappings is not as obvious as that of the document-topic assignments in our WTM model.

Besides, we also analyze the results of topic coherence on the unlabeled social dataset, Tianya-12261. The topic number follows the best results in the Fu et al. (2015). It is 8. We list some examples of each topic and their 20 most probable topical words in Tab. 6.

Table 6: Examples of each topic and their 20 most probable topic words on Tianya-12261 dataset.

Model	Topic	20 most probable topical words
LDA	Topic1	孩子/children, 学生/student, 老师/teacher, 女人/woman, 教育/education, 学校/school, 李双江/Li Shuangjiang, 生活/life, 男人/man, 儿子/son, 朋友/friend, 大学/college, 李天一/Li tianyi, 事情/event, 时间/time, 社会/social, 工作/work, 父母/parents, 同学/classmates, 故事/story
	Topic2	中国/China, 美国/American, 日本/Japan, 国家/country, 朝鲜/North Korea, 世界/world, 人口/population, 经济/economics, 国际/international, 台湾/Taiwan, 战争/war, 问题/question, 人类/human, 历史/history, 政府/government, 技术/technology, 民族/nation, 转基因/transgenic, 政策/policy, 发展/development
	Topic 3	电话/call, 重庆/Chongqing, 网友/net friend, 时间/time, 微博/micro-blog, 手机/cell phone, 网络/network, 视频/video, 赵红霞/Zhao hongxia, 朋友/friend, 记者/journalist, 问题/question, 天涯/Tianya, 信息/information, http:, 新闻/news, 朱瑞峰/Zhu ruifeng, 公司 ph/company, 照片/photo, 网站/website
	Topic 4	法院/court, 法律/lawyer, 书记/secretary, 领导/leader, 证据/evidence, 工作/job, 行为/behavior, 局长/director, 公司/company, 事实/fact, 人员/staff, 案件/event, 规定/regulations, 干部/cadre, 公安/police, 市委/committee, 公安局/police, 机关/office, 法官/judge, 信访/visit
TWE	Topic 1	公安局/police, 人民法院/court, 涉嫌/Suspected, 市委/Municipal Party committee, 检察院/procuratorate, 违纪/discipline, 廉政/uncorrupted, 一审/firsttrial, 被告/defendant, 证人/witness, 殴打/beat, 纪委/commission, 案件/event, 公安/police, 审理/judge, 警方/police, 有期徒刑/imprisonment, 新华/Xinhua, 政法委/committee, 公署/government office

Model	Topic	20 most probable topical words
	Topic 2	控股/shares, 有限公司/company, 房地产/housing, 物业/property, 股份/stock, 资产/assets, 股权/stock right, 证券/securities, 置业/home, 商场/market, 地产/house, 基金/fund, 投资/investment, 物流/logistics, 高新技术/high technology, 投资者/investor, 餐饮/restaurant, 股票/stock, 购物/shopping, 期货/futures
	Topic 3	常委/committee, 常委会/committee, 市委/Municipal Party committee, 省委/Provincial Party committee, 书记/secretary, 党组/arty, 人大/National People's Congress, 组织部/organization department, 办公厅/office, 省人大/provincial People's Congress, 政协/CPPCC, 代市长/Municipal Party committee, 全国人大/National People's Congress, 全国人大常委会/National People's Congress Standing Committee, 党委/Party committee, 厅长/Director, 宣传部/Propaganda Department, 全国政协/Chinese People's Political Consultative Conference, 省政协/Provincial Political Consultative Conference, 政法委/Politics and Law Committee
	Topic 4	月刊/monthly, 侦探/detect, 漫画/cartoon, 专栏/columnist, 柯南/Conan, 童话/tale, 散文/Prose, 小说/novel, 出版社/publication 同名/homonymy, 周刊/weekly, 金庸/Jin Yong, 剧本/script, 喜剧/comedy, 古装/ancient, 监制/Producer, 编剧/Screenwriter, 爱情/love, 中文版/Chinese version, 科幻/science fiction
WTM	Topic 1	李双江/Li shuangjiang, 重庆/Chongqing, 警察/policeman, 派出所/Police Station, 记者/journalist, 警方/police, 事故/event, 赵红霞/Zhao hongxia, 人员/people, 视频/video, 李天一/Li tianyi, 受害人/victim, 儿子/son, 事件/event, 电梯/elevator, 朱瑞峰/Zhu ruifeng, 民警/police, 交通/traffic, 网友/friend, 交警/traffic police
	Topic 2	中国/China, 美国/American, 国家/country, 日本/Jacana, 朝鲜/North Korea, 世界/world, 历史/history, 人民/people, 毛泽东/Mao Zedong, 人类/human, 政治/Politics, 战争/war, 问题/question, 革命/Politics, 民族/nation, 台湾/Tai Wan, 国际/national, 主席/chairman, 经济/economics, 思想/thought
	Topic 3	社会/social, 中国/China, 国家/national, 官员/official, 腐败/corrupt, 问题/question, 领导/leader, 人民/people, 政府/government, 制度/system, 权力/power, 经济/economics, 工作/work, 人口/population, 改革/reform, 政策/policy, 发展/development, 群众/masses, 政治/Politics, 利益/interest
	Topic 4	法律/lawyer, 法院/court, 案件/case, 证据/evidence, 行为/action, 派出所/police station, 公安/police, 事实/fact, 公安局/Public Security Bureau, 人员/staff, 警察/police, 犯罪/crime, 机关/office, 受害人/victim, 书记/secretary, 领导/leader, 民警/police, 司法/justice, 情况/situation, 警方/polices

In Tab. 6, we can see the topics of the LDA model consist of similar words and it is difficult to label. In the LDA, for instance, topic 1 contains words " 孩子/children", " 学生/student", " 老师/teacher", " 女人/woman", and topic 4 includes words " 书记/secretary", " 领导/leader", " 局长/director", " 干部/cadre". As for the results of TWE, it can gather certain vocational related words together, but they can't reflect actual events. For example, topic 1 of TWE relates to "criminal case(刑事案件)", topic 2 is about "investment(投资)", topic 3 is generally a mixture of "government department(政府部门)", and topic 4 is about "literature(文学)". It is still difficult to understand what people are concerning about in the TWE. While the WTM model trains word embeddings, topic representations and topic model simultaneously, so it have a better performances to detect topics. Combined with the current events, we can learn that topic 1 is about "Li Tianyi rape event (李天一强奸事件)", topic 2 is about "National political and economic problems(国家政治经济问题)", topic 3 is the discussion on "Corruption of government officials(政府官员腐败)", and topic 4 is the attitude or opinion of people about a civil procedure.

5. Conclusions and Future works

5.1. Conclusions

In this paper, we proposed the Word-Topic Mixture(WTM) model trains word embeddings and topic model simultaneously based on LDA and word embeddings. WTM combines the ideas of TWE and LFLDA. It first uses LDA to capture the word-topic assignment and introduces external corpus as the words semantic supplement into TWE to learn topic vectors and word embeddings. Then the vector representation is applied to LDA model. WTM defines the probability distribution of topic vectors-word embeddings and integrates it with topics-words distribution to detect topics. Finally, our model trains word embeddings and topic model together by defining a new objective function which minimize the KL divergence of the new topic-word distribution function and original one on the LDA. The experimental results prove that WTM performs better on word representation and topic detectionas compared with some state-of-the-art models. It also shows that there is a certain practical significance to train word embedding and topic model simultaneously.

5.2. Future works

The existing topic detection methods can only process the small data and it is difficult to deal with the practical application with the development of big data. Future research may be two directions of the data parallelization and model parallelization. We can consider improvements to existing methods by using Hadoop, Spark and other cloud-based frameworks. On the characteristic expression of data, there is also a better opportunity to the work of topic detection and analysis as the popularity and application of deep learning models.

On the other hand, the proposed method takes semantics associations in the text into account, but it ignores the influence of time on topic distribution. The following study will consider using existing inference algorithms such as the collapse of Gibbs sampling(CGS) Robert and Casella (2004), variational Bayesian(VB) Beal (2003) to detect dynamic topic. We also consider studying more efficient inference algorithm to accommodate the training requirements of large-scale corpus. In addition, the collaborative training among topic detection, emotion analysis and the deep learning of words feature representation is the inevitable trend in the future large-scale social media data analysis and processing.

Acknowledgments

This research is supported by the National Nature Science Foundation of China under Grants 61472258, 61402294, National Key Technology Research and Development Program of the Ministry of Science and Technology of China (2014BAH28F05), Technology Planning Project from Guangdong Province (2014B010118005), Science and Technology Foundation of Shenzhen City under Grants JCYJ20140509172609162 and JCYJ20130329102032059.

References

Matthew James. Beal. Variational algorithms for approximate bayesian inference. *University College London*, 2003.

Yoshua Bengio, Holger Schwenk, Jean S¨¦bastien Sen¨¦cal, Fr¨¦deric Morin, and Jean Luc Gauvain. A neural probabilistic language model. *Journal of Machine Learning Research*, 3(6):1137–1155, 2003.

David M. Blei. Probabilistic topic models. *Communications of the Acm*, 55(4):55–65, 2011.

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

Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. A Novel Neural Topic Model and Its Supervised Extension. *Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2210–2216, 2015.

Ronan Collobert and Jason Weston. A unified architecture for natural language processing: deep neural networks with multitask learning. In *International Conference*, pages 160–167, 2008.

Rajarshi Das, Manzil Zaheer, and Chris Dyer. Gaussian LDA for Topic Models with Word Embeddings. In *Meeting of the Association for Computational Linguistics and the International Joint Conference on Natural Language Processing*, 2015.

Xianghua Fu, Kun Yang, Joshua Zhexue Huang, and Laizhong Cui. Dynamic nonparametric joint sentiment topic mixture model. *Knowledge-Based Systems*, 82:102–114, 2015.

T. L. Griffiths and M Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101 Suppl 1(1):5228–5235, 2004.

Eric H. Huang, Richard Socher, Christopher D. Manning, and Andrew Y. Ng. Improving word representations via global context and multiple word prototypes. In *Meeting of the Association for Computational Linguistics: Long Papers*, pages 873–882, 2012.

Quoc V. Le and Tomas Mikolov. Distributed Representations of Sentences and Documents. *Computer Science*, 4:1188–1196, 2014.

Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. Topic Modeling for Short Texts with Auxiliary Word Embeddings. In *International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2016.

Yang Liu, Zhiyuan Liu, Tat Seng Chua, and Maosong Sun. Topical word embeddings. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, pages 2418–2424, 2015.

Hosam Mahmoud. Polya Urn Models. *Crc Texts in Statistical Science*, 43(2):xii+290, 2008.

Christopher D. Manning, Prabhakar Raghavan, Sch, and Hinrich tze. *Introduction to Information Retrieval*, volume 43. Cambridge University Press, 2008.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient Estimation of Word Representations in Vector Space. *Computer Science,Computation and Language*, 2013a. URL http://arxiv.org/abs/1301.3781.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean. Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, 26:3111–3119, 2013b.

Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. Improving Topic Models with Latent Feature Word Representations. *Transactions of the Association for Computational Linguistics*, 3:299–313, 2015.

Jeffrey Pennington, Richard Socher, and Christopher Manning. Glove: Global Vectors for Word Representation. In *Conference on Empirical Methods in Natural Language Processing*, 2014.

Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer Texts in Statistics. Springer New York, 2004. ISBN 978-1-4419-1939-7. doi: 10.1007/978-1-4757-4145-2. URL http://link.springer.com/book/10.1007/978-1-4757-4145-2.

Ruslan Salakhutdinov and Geoffrey E. Hinton. Replicated Softmax: an Undirected Topic Model. In *Advances in Neural Information Processing Systems 22: Conference on Neural Information Processing Systems 2009. Proceedings of A Meeting Held 7-10 December 2009, Vancouver, British Columbia, Canada*, pages 1607–1614, 2009.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 384–394, 2010.

Jianhua Yin and Jianyong Wang. A Dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 233–242, 2014.