Computational Linguistics and Chinese Language Processing
Vol. 20, No. 2, December 2015, pp. 45-64

Word Co-occurrence Augmented Topic Model in Short Text

Guan-Bin Chen[*] and Hung-Yu Kao[*]

Abstract

The large amount of text on the Internet cause people hard to understand the meaning in a short limit time. Topic models (e.g. LDA and PLSA) has been proposed to summarize the long text into several topic terms. In the recent years, the short text media such as tweet is very popular. However, directly applies the transitional topic model on the short text corpus usually gating non-coherent topics. Because there is no enough words to discover the word co-occurrence pattern in a short document. The Bi-term topic model (BTM) has been proposed to improve this problem. However, BTM just consider simple bi-term frequency which cause the generated topics are dominated by common words. In this paper, we solve the problem of the frequent bi-term in BTM. Thus, we proposed an improvement of word co-occurrence method to enhance the topic models. We apply the word co-occurrence information to the BTM. The experimental result that show our PMI-β-BTM gets well result in the both of regular short news title text and the noisy tweet text. Moreover, there are two advantages in our method. We do not need any external data and our proposed methods are based on the original topic model that we did not modify the model itself, thus our methods can easily apply to some other existing BTM based models.

Keywords: Short Text, Topic Model, Document Clustering, Document Classification

1. Introduction

With the advancement of information and communication technology, the information we obtained is very abundant and multivariate. Especially, in the recent 15 years, many type of the Internet media grow up so that people can get large amount of the information in a short time. These internet media include Wikipedia, blogs and the recently popular social medial

[*] Department of Computer Science and Information Engineering, National Cheng Kung University
E-mail: gbchen@ikmlab.csie.ncku.edu.tw; hykao@mail.ncku.edu.tw

such as Twitter, Facebook et.al. Generally, the articles/documents in the Wikipedia, and blogs are usually the long text and have the complete content. While the short text social media, such as Twitter, become very popular in the recent years. The reason is that these short text social media provide a very convenient way to share the people feeling and thinking.

Generally, these Internet media deliver the people thinking by using the text. However, the large amount of text on the Internet cause people hard to understand the meaning in a short limit time. To solve the problem, many document summarization technologies have been proposed. Among them, topic models summarize the context in large amount of documents into several topic terms. By reading these topic terms, people will understand the content in a short time. Topic model can be performed by the vector space model or the probability model. In the recent years, the probability models such as Probabilistic Latent Semantic Analysis (pLSA) (Hofmann, 1999) and Latent Dirichlet Allocation (LDA) (Blei *et al.*, 2003) are very popular because the probability models base on the document generation process. The inspirations of the document generation process come from the human written articles. When a person writes an article, he or she will inspire some thinking in mind, then extend these thinking into some related words. Finally, they write down these words to complete an article. Probability topic models simulate the behavior of above document generating process. In the view of the vectorization of the probability topic models, when we have a text corpus, we have known the documents and its words distribution by statistic the word vector. Then, the probability topic models split the document-word matrix into the document-topic and topic-word matrices. The distribution of the document-topic matrix describes that the degree of each document belongs each topic while the topic-word matrix describes the degree of each word belongs each topic. The "topic" in these two matrices is the latent factor as the human thinking.

In essence, the topic models capture the word co-occurrence information and these highly co-occurrence words are put together to compose a topic (Divya *et al.*, 2013; Mimno *et al.*, 2011). So, the key to find out high quality topics is that the corpus must contain a large amount of word co-occurrence information and the topic model has the ability to correctly capture the amount of the word co-occurrence. However, the traditional topic models work well in the long text corpus but work poorly in short text corpus. The reason is that the original intention of LDA is designed to model the long text corpus. Exactly, LDA capture the word co-occurrence in document-level (Divya *et al.*, 2013; Yan *et al.*, 2013), but there are no enough words to well judge the word co-occurrence in document-level in a short text document. Figure 1 is an example which shows the difference of the topic model in between the long text and short text corpus. In the long text corpus, each document provides a lot of word co-occurrence information, so that LDA can well capture these information to discover the high quality topics. While in the short text document, there are no enough words in a single document to discover the word co-occurrence information.

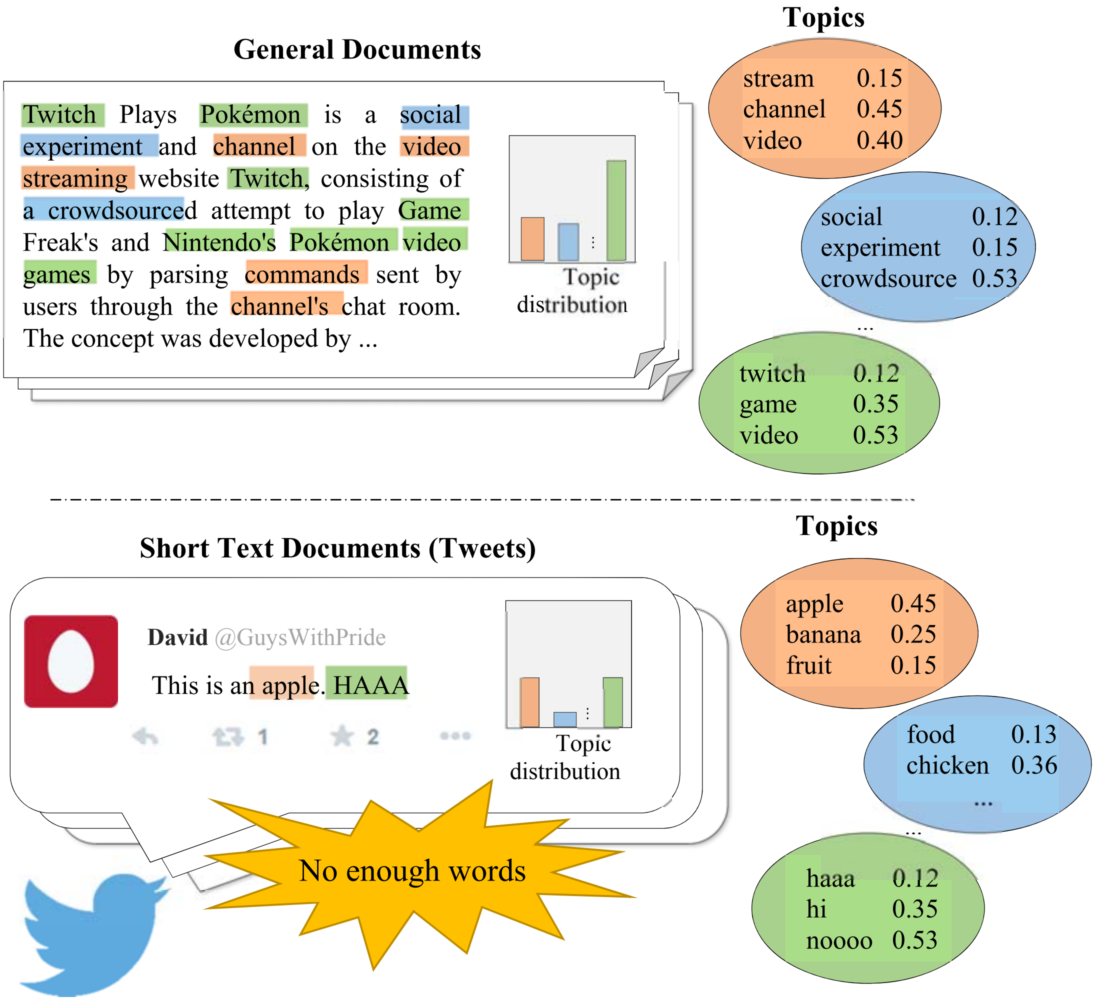
General Documents
Twitch Plays Pokémon is a social experiment and channel on the video streaming website Twitch, consisting of a crowdsourced attempt to play Game Freak's and Nintendo's Pokémon video games by parsing commands sent by users through the channel's chat room. The concept was developed by ...
Topic distribution
Topics
stream 0.15
channel 0.45
video 0.40
social 0.12
experiment 0.15
crowdsource 0.53
...
twitch 0.12
game 0.35
video 0.53
Short Text Documents (Tweets)
David @GuysWithPride
This is an apple. HAAA
1
2
Topic distribution
No enough words
Topics
apple 0.45
banana 0.25
fruit 0.15
food 0.13
chicken 0.36
...
...
haaa 0.12
hi 0.35
noooo 0.53

Figure 1. An example of LDA in the long text and short text corpus

To overcome above problems in short text, many researchers consider a simpler topic model, mixture of unigrams model. Mixture of unigrams model samples topics in global corpus level (Nigam *et al*., 2000; Zhao *et al*., 2011). More specifically, the word co-occurrence in document-level means that the amount of the word co-occurrence relation comes from a single document. On the contrary, the word co-occurrence in corpus-level means that the amount of the word co-occurrence relation comes from a full corpus which contains many documents. Mixture of unigrams overcomes the lack of words in the short text documents. Further, Xiaohui Yan *et al*. proposed the Bi-term Topic Model (BTM) (Yan *et al*., 2013; Cheng *et al*., 2014) which directly model the word co-occurrence and use the corpus-level bi-term to overcome the lack of the text information problem. A bi-term is an unordered word pair co-occuring in a short text document. The major advantage of BTM is that 1) BTM model the word co-occurrence by using the explicit bi-term, and 2) BTM aggregate these word co-occurrence patterns in the corpus for topic discovering (Yan *et al*., 2013; Cheng *et al*., 2014). BTM abandons the document-level directly. A topic in BTM contains several bi-term and a bi-term crosses many documents. BTM emphasizes that the co-occurrence information comes from all bi-terms in whole corpus. However, BTM will make the common words be performed excessively because the frequency of bi-term comes from the whole corpus instead of a short document.

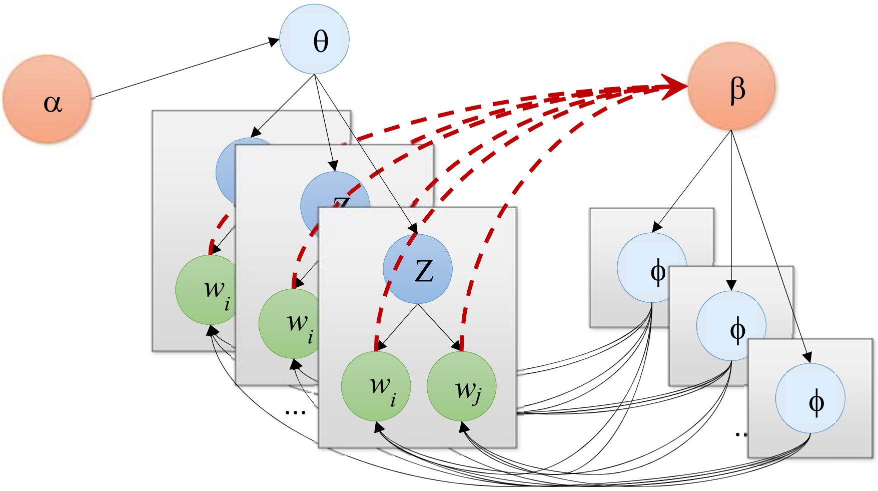

Figure 2. The graphical representation of the PMI-β-BTM

In this paper, we solve the frequent bi-term problem in BTM. We propose an approach base on BTM. For the problem in BTM, a simple and intuitive solution is to use pointwise mutual information (PMI) (Church & Hanks, 1990) to decrease the statistical amount of the frequent words in whole corpus. With respect to the frequency of bi-term, the PMI can normalize the score by each single word frequency in the bi-term. Otherwise, the priors in the topic models usually set symmetric. This symmetric priors mean that there is not any preference of words in any specific topic (Wallach *et al.*, 2009). An intuitive idea is that why not adopt some word co-occurrence information in priors to restrict the generated topics. Base on above two ideas, we propose a novel prior adjustment method, PMI-β priors, which first use the PMI to mine the word co-occurrence from the whole corpus. Then, we transform such PMI scores to the priors of BTM. Figure 2 shows the graphical representation of the PMI-β-BTM.

In summary, the proposed approach enhance the amount of the word co-occurrence and also based on the original topic model. Basing on the original topic model means we did not modify the model itself, thus our methods can easily apply to some other existing BTM based models, to overcome the short text problem without any modification. To test the performance of our two methods completely, we prepare two different types of short text corpus for the experiments. One is the tweet text and another is the news title. The context of news title dataset is regular and formal while the text in tweet usually contain many noise. Experimental results show our PMI-β priors method is better than the BTM in both tweet and news title datasets.

The remaining of this paper shows below. In Section 2, we show the survey of some traditional topic models and the previous works of topic model to overcome the short text. Section 3 shows our proposed PMI-β priors and the re-organized document methods. The experiment results show in Section 4. Finally, we conclude this research in Section 5.

2. Related Work

2.1 The Survey of the Traditional Topic Models for Normal Text

Topic Model is a method to find out the hidden semantic topics from the observed documents in the text corpus. Topic Models have been researched several years. Generally, topic model can be performed by the vector space model or the probability model. The early one of the vector space topic model, Latent Semantic Analysis (LSA) (Landauer *et al.*, 1998), uses the singular value decomposition (SVD) to find out the latent topic. However, LSA does not model the polysemy well and the cost of SVD is very high (Hofmann, 1999; Blei *et al.*, 2003). Afterward, Thomas Hofmann proposed the one-document-multi-topics model, probabilistic Latent Semantic Analysis (pLSA) (Hofmann, 1999). pLSA bases on the document generation process which like the human writing. However, the numerous parameters of pLSA cause the overfitting problem and pLSA does not define the generation of the unknown documents. In 2003, Blei *et al.* proposed a well-known Latent Dirichlet Allocation (LDA) (Blei *et al.*, 2003), LDA use the prior probability in Bayes theory to extents pLSA and simplify the parameters estimate process in pLSA. Also, the non-zero priors let LDA have the ability to infer the new documents.

However, there are some drawbacks in LDA. First, LDA works under the bag-of-word model hypothesis. In the bag-of-word model, each word of the document is no order and independent of others (Wallach, 2006). The hypothesis compared with the human writing behavior is unreasonable (Divya *et al.*, 2013). Second, LDA emphasizes the relations between topics are week, but actually, the topics may have hierarchical structure. Third, LDA requires the large number of articles and well-structured long articles to get the high quality topics. Apply LDA on the short text or uncompleted sentences corpus usually get poor results. The fourth drawback is that in spite of the LDA has the concept of the prior probabilities but LDA priors generally set the symmetric values in each prior vector, like <0.1> or <0.01>. The symmetric prior means no bias of each words in the specific topic (Wallach *et al*., 2009). In this situation, the priors only provide the smooth technology to avoid the zero probability and the model only use the statistical information from the data to discover the hidden topics.

To overcome above four drawbacks, many researchers propose new modify models. Such as N-gram Topic Model (Wang *et al*., 2007) and HMM-LDA (Griffiths *et al*., 2004) provide the context modeling. Wei Li *et al*. proposed the Pachinko Allocation Model (PAM) (Li & McCallum, 2006) which adds the super topic concept and make the topic have the hierarchical structure. Otherwise, Zhiyuan Chen *et al*. apply the must-link and cannot-link information to guide the document generation process which words must or not to be put into a topic (Chen & Liu, 2014).

2.2 Topic Models for Short Text

With the rise of social media in recent years, topic models have been utilized for social media analysis. For example, some researches apply topic models in social media for event tracking (Lin *et al*., 2010), content characterizing (Zhao *et al*., 2011; Ramage *et al*., 2010), and content recommendation (Chen *et al*., 2010; Phelan *et al*., 2009). However, to share people thinking conveniently, the context is usually short. These short text contexts make topic models hard to discover the amount of word co-occurrence. For the short text corpus, there are three directions to overcome the insufficient of the word co-occurrence problem. One is using the external resources to guide the model generation, another is aggregating several short texts into a long text, and the other is improving the model to satisfy the short text properties. For the first direction, Phan *et al*. (Phan *et al*., 2008) proposed a framework that adopt the large external resources (such as Wiki and blog) to deal with the data sparsity problem. R.Z. Michal *et al*. proposed an author topic model (Rosen-Zvi *et al*., 2004) which adopt the user information and make the model suitable for specific users. Jin *et al*. proposed the Dual-LDA model (Jin *et al*., 2011), it use not only the short text corpus but also the related long text corpus to generate topics, respectively. The generation process use the long text to help the short text modeling. If the quality of the external long text or knowledge base is high, the generated topic quality will be improve. However, we cannot always obtain the related long text to guide short text and the related long text is very domain specific. So, using external resources is not suitable for the general short text dataset. In addition to adopt the long text, Hong et al. aggregate the tweets which shared the same words and get better results than the original tweet text (Hong & Davison, 2010).

For the model improvement, Wayne *et al.* use the mixture of unigrams model to model the tweets topics from whole corpus text (Zhao *et al*., 2011). Their experimental results verify that the mixture of unigram model can discover more coherent topics than LDA in the short text corpus. Further, Xiaohui Yan *et al.* proposed the Bi-term Topic Model (BTM) (Yan *et al.*, 2013; Cheng *et al.*, 2014) which directly model the word co-occurrence and use the corpus level bi-term to overcome the lack of the text information problem. A bi-term is a word pair containing a co-occur relation in this two words. The advantage is that BTM can model the general text without any domain specific external data. Comparing with the mixture of unigram, BTM is a special case of the mixture of unigram. They both model the corpus level topic but BTM generates two words (bi-term) every time the generation process. However, BTM discovers the word co-occurrence just by considering the bi-term frequency. The bi-term frequency will be failed to judge the word co-occurrence when the bi-term frequency is high but one of the frequency of two words in a bi-term is high and another is low.

3. The Word Co-occurrence Augmented Methods

Topic models learn topics base on the amount of the word co-occurrence in the documents. The word co-occurrence is a degree which describes how often the two words appear together. BTM, discovers topics from bi-terms in the whole corpus to overcome the lack of local word co-occurrence information. However, BTM will make the common words be performed excessively because BTM identifies the word co-occurrence information by the bi-term frequency in corpus-level. Thus, we propose a PMI-β priors methods on BTM. Our PMI-β priors method can adjust the co-occurrence score to prevent the common words problem. Next, we will describe the detail of our method of PMI-β priors.

We first describe the detail of BTM. First, we introduce the notation of "bi-term". Bi-term is the word pair co-occurring in the short text. Any two distinct words in a document construct a bi-term. For example, a document with three terms will generate three bi-term (Yan *et al.*, 2013):

$$(t_1,t_2,t_3) \Rightarrow \{(t_1,t_2),\ (t_2,t_3),\ (t_1,t_3)\}. \tag{1}$$

Note that each bi-term is unordered. For a real case example, we have a document and the context is "I visit apple store". Because "I" is a stop-word, we remove it. The remaining three terms "visit", "apple" and "store" will generate three bi-terms "visit apple", "apple store", and "visit store". We generate all possible bi-terms for each document and put all bi-terms in the bi-term set B.

Second, we describe the parameter estimation of the BTM. The aim of the parameter estimation of BTM is to estimate the topic assignment z, the corpus-topic posteriori distribution θ and the topic-word posteriori distribution ϕ. But the Gibbs sampling can integrate θ and ϕ due to use the conjugate priors. Thus, the only one parameter z should be estimate. Clearly, we should assign a suitable topic for each bi-term. The Gibbs sampling equation shows below:

$$P(z = k \mid \mathbf{z}_{\neg b}, \mathbf{B}, \boldsymbol{\alpha}, \boldsymbol{\beta}) \propto \theta \cdot \varphi \,, \tag{2}$$

where z is the topic assignment, k means the kth topic, **B** is the bi-term set, α is the corpus-topic prior distribution and β is the topic-word prior distribution. The θ and ϕ in Eq. (2) show following:

$$\theta = \frac{(n_{k,\neg b} + \alpha_k)}{\sum\limits_{k=1}^{K} (n_{k,\neg b} + \alpha_k)} \,, \tag{3}$$

$$\varphi = \frac{(n^{w_1}_{k,\neg b} + \beta^{w_1}_{k})}{\sum\limits_{t=1}^{V} (n^{w_t}_{k,\neg b} + \beta^{w_t}_{k})} \times \frac{(n^{w_2}_{k,\neg b} + \beta^{w_2}_{k})}{\sum\limits_{t=1}^{V} (n^{w_t}_{k,\neg b} + \beta^{w_t}_{k})} \,, \tag{4}$$

where V is the number of unique words in the corpus, $n_{k,-b}$ is the statistical count for the document-topic distribution, and $n^{w_t}_{k,\neg b}$ is the statistical count for the document-topic distribution. When the frequency of bi-term is high the two terms in this bi-term tend to be put into the same topic. Otherwise, to overcome the lack of words in a single document BTM abandons the document-level directly. A topic in BTM contains several bi-term and a bi-term crosses many documents. BTM emphasizes that the co-occurrence information comes from all bi-terms in whole corpus.

However, just consider the frequency of bi-term in corpus-level will generate the topics which contain too many common words. To solve this problem, we consider the Pointwise Mutual Information (PMI) (Church & Hanks, 1990). Since the PMI score not only considers the co-occurrence frequency of the two words, but also normalizes by the single word frequency. Thus, we want to apply PMI score in the original BTM. A suitable way to apply PMI scores is modifying the priors in the BTM. The reason is that the priors modifying will not increase the complexity in the generation model and very intuitive. Clearly, there are two kinds of priors in BTM which are β-prior and β-priors. The β-prior is a corpus-topic bias without the data. While the β-priors are topic-word biases without the data. Applying the PMI score to the β-priors is the only one choice because we can adjust the degree of the word co-occurrence by modifying the distributions in the β-priors. For example, we assume that a topic contains three words "pen", "apple" and "banana". In the symmetric priors, we set <0.1, 0.1, 0.1> which means no bias of these three words, while we can apply <0.1, 0.5, 0.5> to enhance the word co-occurrence of "apple" and "banana". Thus the topic will prefer to put the "apple" and "banana" together in the topic sampling step.

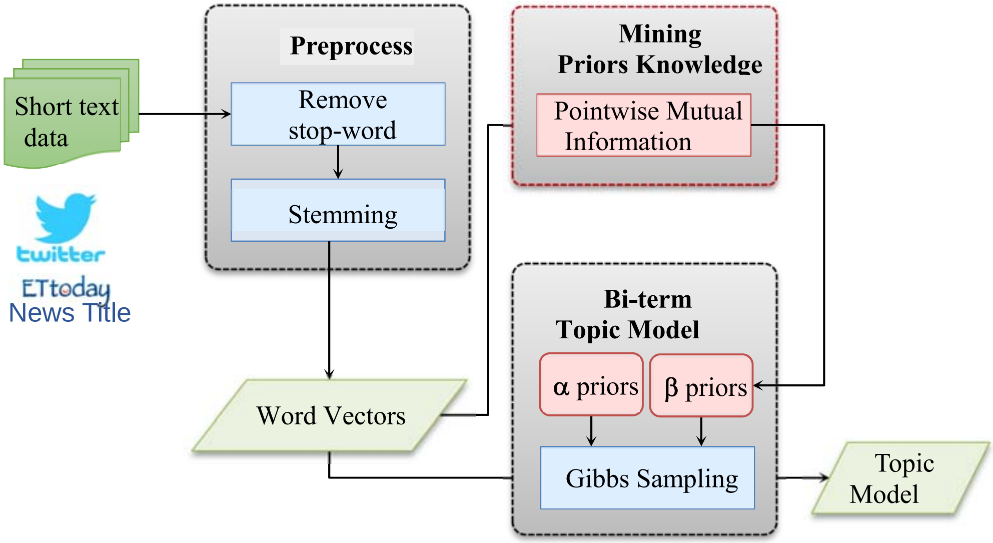

Figure 3. The PMI-β priors approach

Figure 3 shows our PMI-β-priors approach. After pre-procession, we first calculate the PMI score of each bi-term <w_x, w_y> as

$$\mathrm{PMI}(w_x, w_y) = \log \frac{p(w_x, w_y)}{p(w_x)p(w_y)}, \tag{5}$$

Because the priors can view as an additional statistics count of the target probability, the value ordinarily should be greater than or equal to zero. Thus, we adjust the value of NPMI to [0, 2] by adding one as:

$$\mathrm{NPMI}(w_x, w_y) = \frac{\mathrm{PMI}(w_x, w_y)}{-\log p(w_x, w_y)} + 1. \tag{6}$$

After getting the NPMI scores, we transform these scores to meet the β-priors. Let β_{SYM} is the original symmetric β-priors and the PMI β-priors, denote β_{PMI}, define as

$$\beta_{\mathrm{PMI}}^{w_x, w_y} = \beta_{\mathrm{SYM}} + 0.1 \times \mathrm{NPMI}(w_x, w_y). \tag{7}$$

There is a constant value 0.1 in Eq. (7). This constant value 0.1 prevent the target probability being dominated by the priors. The partial of the word co-occurrence information should still be captured by the original model and the priors provide the additional information to enhance the word co-occurrence in the model. The following shows how we apply PMI-β-priors into the BTM. We apply the β_{PMI} of w1 and w2 in Eq. (6) and the new equation of shows below:

$$\varphi = \frac{(n_{k,\neg b}^{w_1} + \beta_{\mathrm{PMI}}^{w_1,w_2})}{\sum_{t=1}^{V} (n_{k,\neg b}^{w_t} + \beta_{k}^{w_t})} \times \frac{(n_{k,\neg b}^{w_2} + \beta_{\mathrm{PMI}}^{w_1,w_2})}{\sum_{t=1}^{V} (n_{k,\neg b}^{w_t} + \beta_{k}^{w_t})} . \tag{8}$$

Finally, we sample topic assignments by Gibbs sampling (Liu, 1994) approach.

4. Experiments

How to justly evaluate the quality of the topic model is still a problem. The reason is that the topic model is an unsupervised method. There are no prominent words or labels can directly assign to each topic. Thus, many researchers apply topic model in other applications, such as clustering, classification and information retrieval (Blei *et al.*, 2003; Yan *et al.*, 2013). In classification task, instead of using the original word vectors to identify the document categories, it use the reduced vectors which generating from the topic model. The topic model plays as a dimensional reduction role and the classification result shows how well the model to represent the original features. Topic model can also look as the document clustering approach by just considering a document assign to which topic(s). In this paper, we evaluate topic models by clustering and classification tasks. Otherwise, to make our experiment more robust, we adopt two different types of short text dataset - Twitter2011 and ETtoday Chinese news title. The properties of these two corpus are different. The text of ETtoday Chinese news title is very regular, while the text of Twitter2011 usually contains emotional words, simplified texts and some unformed words. For example, “haha” is the emotional word, and “agreeeee” is the unformed word.

Table 1 shows the statistics of short text datasets. The number of average words per document is not more than ten words. The number of documents in each class are shown in Figure 4. The property of both two dataset is skew. The skew dataset may cause the results that the fewer documents are dominated by the larger one. In summary, the challenges of these two datasets are not only the short text problem but also the unbalance category. The top-3 classes in the Twitter2011 dataset are “#jan25”, “#superbowl” and “#sotu”. And the top-3 classes in the ETtoday News Title dataset are “entertainment”, “physical” and “political”.

Table 1. The Statistics of Two Short Text Datasets

Property	Twitter2011	ETtoday News title
The number of documents	49,461	17,814
The number of domains	50	25
The number of distinct words	30,421	31,217
Avg. words per document	5.92	9.25

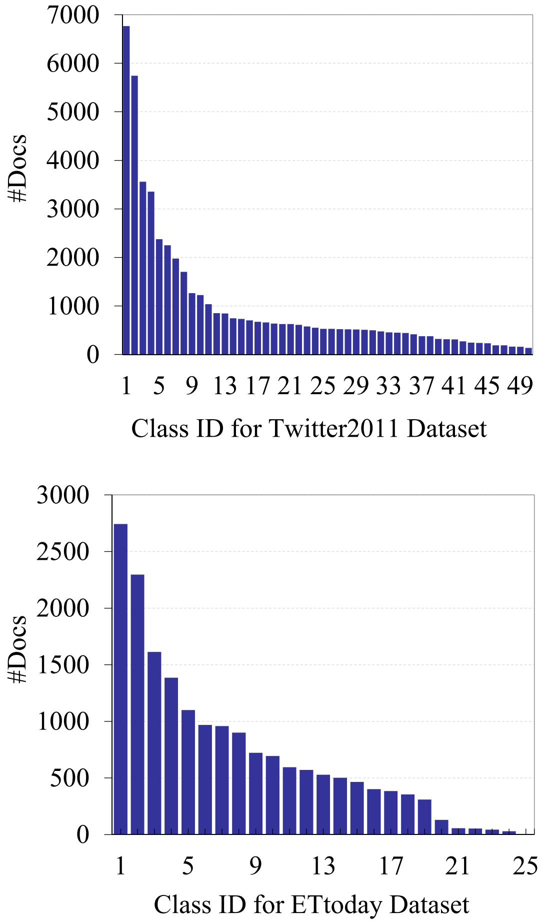

Figure 4. The number of documents in each class

4.1 Experimental Setup

All of the experiments were done on the Intel i7 3.4 GHz CPU and 16G memory PC. All of the pre-process and topic models were written by JAVA code. The parameters α priors and the base β priors of topic models are all set <0.1>. The number of iterations in Gibbs sampling is set 1,000. To make our results more reliable, we run each experiments 10 times and average these scores.

For the clustering experiment, we first get the document-topic posteriori probability distribution ϕ and we use the highest probability topic P(z|d) as the cluster assignment for each document in ϕ. For the classification experiment, we divide our dataset into five parts in which four parts for training and one for testing. After training the topic model, we fix the topic-word distribution ϕ and then we re-infer document-topic posteriori probability distribution θ of all original short text documents. Instead of using the original word vectors to do the classification task, we take this re-inferred posteriori probability distribution θ as the reduced feature matrix. Finally we use this reduced feature matrix to classify the documents by LIBLINEAR[1].

We compare our methods with the previous topic models: 1) LDA, 2) Mixture of unigrams, and 3) BTM. In addition to the above three topic models, we also compare with our PCA-β priors methods. We use the principal component analysis (PCA) to discover the whole corpus principal component. Then, we transform the principal component to the topic-word prior distribution.

4.2 Evaluation Criteria

In this part, we list three criteria for the clustering experiment and one for classification. In the clustering experiment, let Ω = {ω_1, ω_2, ... , ω_K} is the output cluster labels, and C = {c_1, c_2, ... , c_p} is the gold standard labels of the documents. We first describe the three criteria for the clustering.

- **Purity**

Purity is a simple and transparent measure which perform the accuracy of all cluster assignments as the following equation:

$$\mathrm{Purity}(\Omega,\mathrm{C}) = \frac{\sum_{k} \max_{j} \left\| \varpi_k \cap c_j \right\|}{N}, \tag{9}$$

where N is the total number of documents. Note that the high purity is easy to achieve when the number of clusters is large. In particular, purity is 1 if each document gets its own cluster.

- **Normalized Mutual Information (NMI)**

NMI score is based on the information theory. Let I(Ω, C) denotes the mutual information between the output cluster Ω and the gold standard cluster C. The mutual information of NMI is normalized by each entropy denoted H(Ω) and H(C). This normalization can avoid the influence of the number of clusters. The equation of NMI shows following:

$$\mathrm{NMI}(\Omega,\mathrm{C}) = \frac{\mathrm{I}(\Omega,\mathrm{C})}{[\mathrm{H}(\Omega)+\mathrm{H}(C)]/2}, \tag{10}$$

where I(Ω, C), H(Ω) and H(Ω) denote:

[1] http://www.csie.ntu.edu.tw/~cjlin/liblinear/

$$\mathrm{I}(\Omega,\mathrm{C})=\sum_{k}\sum_{j}\mathrm{P}(\varpi_k \cap c_j)\log\frac{\mathrm{P}(\varpi_k \cap c_j)}{\mathrm{P}(\varpi_k)\,\mathrm{P}(c_j)}, \tag{11}$$

$$\mathrm{H}(\Omega)=-\sum_{k}\mathrm{P}(\varpi_k)\log\mathrm{P}(\varpi_k)\,. \tag{12}$$

- **Rand Index**

Rand Index (RI) (Rand, 1971) consider the clustering result as a pair-wise decision. More clearly, RI penalizes both true positive and true negative decisions during clustering. If two documents are both in the same class and the same cluster, or both in different classes and different clusters, this decision is correct. For other cases, the decision is false. The equation of RI shows following:

$$RI=\frac{TP+TN}{TP+FP+FN+TN}, \tag{13}$$

where *TP*, *FP*, *FN*, and *TN* are the true positive count, false positive count, false negative count and true negative count respectively. For the classification experiment, we adopt the accuracy as the measure. The definition of the accuracy is the same as the *RI* score in Eq. (13), but just change the cluster label to the classification label.

4.3 Experimental Results for the Twitter2011 Dataset

The Twitter2011 dataset was published in TREC 2011 microblog track[2]. It contains approximately 16 million tweets sampled between January 23rd and February 8th, 2011. It is worth mentioning that there are some semantics tags, called hashtag, in some tweets. The hashtags had been given when the author wrote a tweet. Because these hashtags can identify the semantics of tweets, we use the hashtags as our ground truth for both clustering and classification experiments. However, there are about 10 percentages of all tweets contain hashtags and some hashtags are very rare. Also, there are contains multilingual tweets. To reduce the effect of noise in this dataset, we just extract the English tweets with top-50 frequent hashtags. After tweet extraction, we totally get the 49,461 tweets. Then, we remove the hashtags and stop-words from the context. Finally, we stem all the words in all tweets by the English stemming in the Snowball library.

Table 2 shows the clustering results on the Twitter2011 dataset, when we set the number of topic to 50. As expected, BTM is better than Mixture of unigram and LDA got the worst result when we adopt the symmetric priors <0.1>. When apply the PMI-β priors, we get the

[2] http://trec.nist.gov/data/tweets/

better result than BTM with symmetric priors. Otherwise, our baseline method, PCA-β, is better than the original LDA because the PCA-β prior can make up the lack of the global word co-occurrence information in the original LDA.

Table 2. The Clustering Results on Twitter2011 dataset

Model	β priors	Purity	NMI	RI
LDA	<0.100>	0.4174	0.3217	0.9127
	PCA-β	0.4348	0.3325	0.9266
Mix	<0.100>	0.4217	0.3358	0.8687
	PCA-β	0.3748	0.3305	0.7550
BTM	<0.100>	0.4318	0.3429	0.9092
	PCA-β	0.4367	0.4000	0.8665
	PMI-β	0.4427	0.3927	0.9284

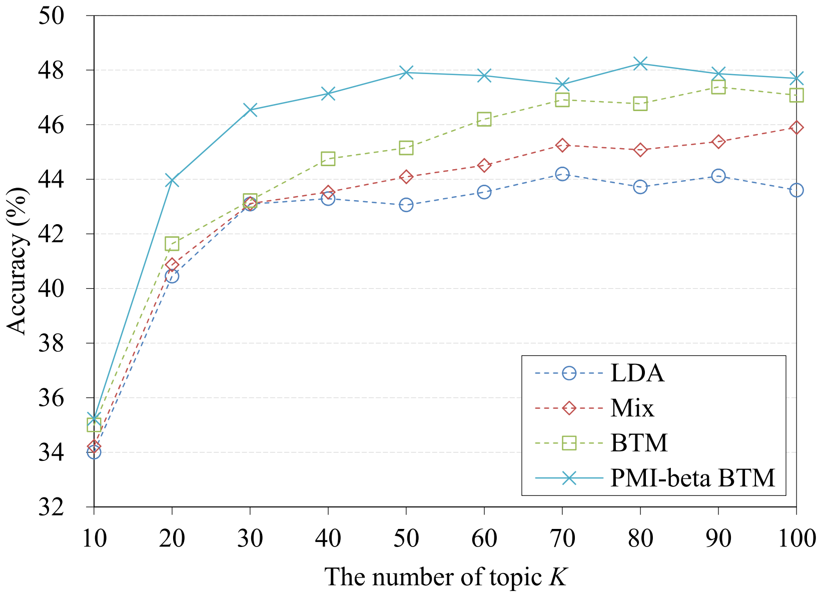

Figure 5. The Classification Results on Twitter2011 dataset

Figure 5 shows the classification results on the Twitter2011 dataset by using LIBLINEAR classifier. When apply the PMI-β priors, we get the better result than BTM with symmetric priors. Table 3 presents the top-10 topic words of the "job" topic in the Twitter2011 dataset for LDA, mixture of unigram, BTM and PMI-β-BTM respectively, when the number of topic is 70. The top-10 words are the 10 highest probability words of the topics. The bold words in this table are the words which highly correlated with the topic by the human judgment. The topic words in the LDA and mixture of unigram models are almost non-correlated or low-correlated with the topic “job”, such as “jay” and “emote”. In BTM and PMI-β-BTM, the model capture the more high-correlated words, such as “engineer” and “management”.

Table 3. The top-10 topic words of the “job” topic in Twitter2011 dataset

	Top-10 Topic words
LDA	**job**, house, jay, steal, material, burglary, **construct**, park, pick, ur
Mix	**job**, robbery, material, **construct**, steal, warehouse, emote, feel, woman, does
BTM	**job**, **management**, **engineer**, media, social, open, **sale**, analyst, **develop**, **senior**
PMI-β-BTM	**job**, real, open, estate, **management**, **market**, **company**, **sale**, **develop**, **engineer**

4.4 Experimental Results for ETtoday News Title Dataset

The ETtoday News Title dataset is collected from the overview list of the ETtoday News website[3] between January 1st and January 31, 2015. There are totally 25 predefined news labels in the dataset. These labels include some classical news category such as “society news”, “international news” and “political news”, and some special news category such as “animal and pets”, “3C” and “games”. In both the clustering and the classification experiments, we use these labels as the ground-truth. Because the Chinese text does not contain the break word, we must adopt the additional word breaker in the pre-process step. We adopt the jieba[4], the Python Chinese word segmentation module, to segment all news title into several words.

Figure 6 shows the classification results on the ETtoday News Title dataset. The three original topic model LDA, mixture of unigram, and BTM perform the same order as the results of the Tweet2011 dataset. The PMI-β BTM is outperform all other methods. Our PMI-β-BTM is also suitable to model the regular short text. The top-10 topic words of the “baseball” topic of ETtoday news title dataset lists in the Table 4. Because these words are almost Chinese, we also attach the simple explanation in English. There are many non-related words in the LDA and mixture of unigram, such as “年終” (Year-end bonuses) and “不” (no). Especially, we compare the topic words in BTM with in PMI-β-BTM, the topic words in BTM contain some frequent but low-correlated words with the topic, such as “年” (means year) and “萬” (means ten thousand). While in the PMI-β-BTM, this noisy words do not appear. The reason is that the original BTM just consider the simple bi-term frequency and this bi-term frequency make some frequent words be extracted together with other words from the

[3] http://www.ettoday.net/news/news-list.htm

[4] https://github.com/fxsjy/jieba

document. Our PMI-β priors can decrease the probability of the common words by the word normalization effect in the PMI.

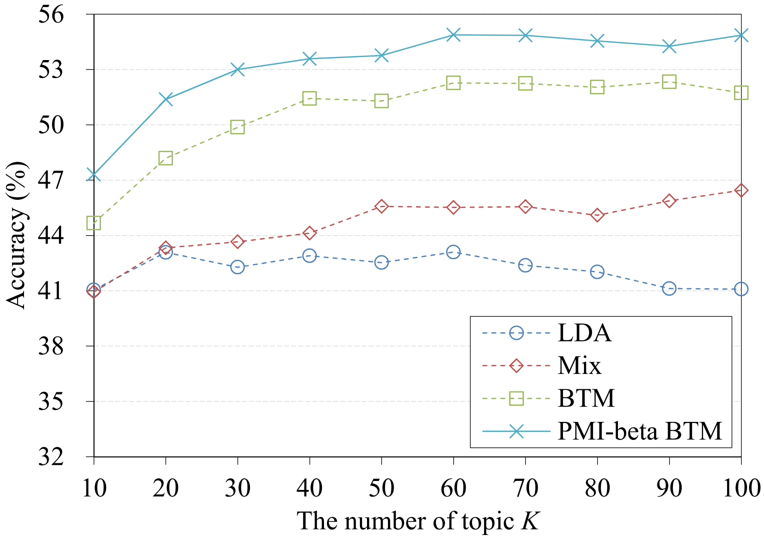

Figure 6. The Classification Results on ETtoday dataset

Table 4. The top-10 topic words of the "baseball" topic in ETtoday News Title dataset

	Top-10 Topic words
LDA	中職 (baseball game in Taiwan), 月 (month), 萬 (ten thousand), 年 (year), 大 (big), 元 (dollars), 吳誌揚 (a politician), 臺北 (Taipei), 臺灣 (Taiwan), 年終 (Year-end bonuses)
Mix	中職, 日 (day), 臺灣, 大, 英雄 (hero), 聯盟 (league baseball), 世界 (world), 棒球 (baseball), 不 (no), 挑戰 (challenge)
BTM	中職，義大 (a baseball team), 兄弟 (a baseball team), **MLB**, 統一 (a baseball team), 年，桃猿 (a baseball team), 萬，獅 (a baseball team), 人 (human)
PMI-β-BTM	中職，**MLB**，兄弟，日職 (baseball game in Japan), 棒球，桃猿，先發 (Starting Pitcher), 總冠軍 (champion), 陳偉殷 (a Taiwanese professional baseball pitcher), 統一 (a baseball team)

5. Conclusions

In this paper, we propose a solution for topic model to enhance the amount of the word co-occurrence relation in the short text corpus. First, we find the BTM identifies the word co-occurrence by considering the bi-term frequency in the corpus-level. BTM will make the common words be performed excessively because the frequency of bi-term comes from the whole corpus instead of a short document. We propose a PMI-β priors method to overcome this problem. The experimental results show our PMI-β-BTM get the best results in the regular short news title text.

Moreover, there are two advantages in our methods. We do not need any external data and the proposed two improvement of the word co-occurrence methods are both based on the original topic model and easy to extend. Bases on the original topic model means we did not modify the model itself, thus our methods can easily apply to some other existing BTM based models to overcome the short text problem without any modification. In the future, we can extend some other steps in PMI-priors to deal the further improvement, such as removing the redundant documents by clustering.

References

Hofmann, T. (1999). Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, 289-296.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, *3*, 993-1022.

Divya, M., Thendral, K., & Chitrakala, S. (2013). A Survey on Topic Modeling. *International Journal of Recent Advances in Engineering & Technology (IJRAET)*, *1*, 57-61.

Mimno, D., Wallach, H. M., 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*, 262-272.

Yan, X., Guo, J., Lan, Y., & Cheng, X. (2013). A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, Rio de Janeiro, Brazil, 1445-1456.

Nigam, K., McCallum, A. K., Thrun, S., & Mitchell, T. (2000). Text classification from labeled and unlabeled documents using EM. *Machine learning*, *39*(2), 103-134.

Zhao, W. X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., *et al.* (2011). *Comparing twitter and traditional media using topic models*. In Advances in Information Retrieval. ed: Springer, 338-349.

Cheng, X., Yan, X., Lan, Y., & Guo, J. (2014). BTM: Topic Modeling over Short Texts. *Knowledge and Data Engineering, IEEE Transactions on*, *26*(12), 2928-2941.

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, *16*(1), 22-29.

Wallach, H. M., Minmo, D., & McCallum, A. (2009). Rethinking LDA: Why priors matter. In *Advances in Neural Information Processing Systems 22 (NIPS 2009)*.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, *25*(2&3), 259-284.

Wallach, H. M. (2006). Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, 977-984.

Wang, X., McCallum, A., & Wei, X. (2007). Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In *Seventh IEEE International Conference on Data Mining (ICDM 2007)*, 697-702.

Griffiths, T. L., Steyvers, M., Blei, D. M., & Tenenbaum, J. B. (2004). Integrating topics and syntax. In *Advances in neural information processing systems 17*, 537-544.

Li, W. & McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. In *Proceedings of the 23rd international conference on Machine learning*, 577-584.

Chen, Z. & Liu, B. (2014). Mining topics in documents: standing on the shoulders of big data. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, New York, New York, USA, 1116-1125.

Lin, C. X., Zhao, B., Mei, Q., & Han, J. (2010). PET: a statistical model for popular events tracking in social communities. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining*, 929-938.

Ramage, D., Dumais, S. T., & Liebling, D. J. (2010). Characterizing Microblogs with Topic Models. In *Fourth International AAAI Conference on Weblogs and Social Media*.

Chen, J., Nairn, R., Nelson, L., Bernstein, M., & Chi, E. (2010). Short and tweet: experiments on recommending content from information streams. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1185-1194.

Phelan, O., McCarthy, K., & Smyth, B. (2009). Using twitter to recommend real-time topical news. In *Proceedings of the third ACM conference on Recommender systems*, 385-388.

Phan, X.-H., Nguyen, L.-M., & Horiguchi, S. (2008). Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on World Wide Web*, 91-100.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th conference on Uncertainty in artificial intelligence*, 487-494.

Jin, O., Liu, N. N., Zhao, K., Yu, Y., & Yang, Q. (2011). Transferring topical knowledge from auxiliary long texts for short text clustering. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, 775-784.

Hong, L. & Davison, B. D. (2010). Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, 80-88.

Liu, J. S. (1994). The collapsed Gibbs sampler in Bayesian computations with applications to a gene regulation problem. *Journal of the American Statistical Association*, *89*(427), 958-966.

Rand, W. M. (1971). Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, *66*(336), 846-850.

Ramage, D., Hall, D., Nallapati, R., & Manning, C. D. (2009). Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, *1*, 248-256.