Topic Modeling over Short Texts by Incorporating Word Embeddings

Jipeng Qiang[1,2,3](✉), Ping Chen[3], Tong Wang[3], and Xindong Wu[2,4]

[1] Yangzhou University, Yangzhou 225009, China
qjp2100@163.com
[2] Hefei University of Technology, Hefei 230009, China
[3] University of Massachusetts Boston, Boston, MA 02155, USA
[4] University of Louisiana at Lafayette, Lafayette, LA 70504, USA

Abstract. Inferring topics from the overwhelming amount of short texts becomes a critical but challenging task for many content analysis tasks. Existing methods such as probabilistic latent semantic analysis (PLSA) and latent Dirichlet allocation (LDA) cannot solve this problem very well since only very limited word co-occurrence information is available in short texts. This paper studies how to incorporate the external word correlation knowledge into short texts to improve the coherence of topic modeling. Based on recent results in word embeddings that learn semantically representations for words from a large corpus, we introduce a novel method, Embedding-based Topic Model (ETM), to learn latent topics from short texts. ETM not only solves the problem of very limited word co-occurrence information by aggregating short texts into long pseudo-texts, but also utilizes a Markov Random Field regularized model that gives correlated words a better chance to be put into the same topic. The experiments on real-world datasets validate the effectiveness of our model comparing with the state-of-the-art models.

Keywords: Topic modeling · Short text · Word embeddings

1 Introduction

Topic modeling has been proven to be useful for automatic topic discovery from a huge volume of texts. Topic model views texts as a mixture of probabilistic topics, where a topic is represented by a probability distribution over words. Many topic models such as Latent Dirichlet Allocation (LDA) have demonstrated great success on long texts (news article and academic paper) [2,5]. In recent years, knowledge-based topic models have been proposed, which ask human users to provide some prior domain knowledge to guide the model to produce better topics instead of purely relying on how often words co-occur in different contexts. For example, two recently proposed models, i.e., a quadratic regularized topic model based on semi-collapsed Gibbs sampler [10] and a Markov Random Field regularized Latent Dirichlet Allocation model based on Variational Inference [18], share the idea of incorporating the correlation between words.

J. Kim et al. (Eds.): PAKDD 2017, Part II, LNAI 10235, pp. 363–374, 2017.
DOI: 10.1007/978-3-319-57529-2_29

With the rapid development of the World Wide Web, short text has been an important information source not only in traditional web site, e.g., web page title and image caption, but in emerging social media, e.g., tweet, status message, and question in Q&A websites. Compared with long texts, topic discovery from short texts has the following three challenges: only very limited word co-occurrence information is available, the frequency of words plays a less discriminative role, and the limited contexts make it more difficult to identify the senses of ambiguous words [15]. Therefore, long text topic models cannot work very well on short texts [4,20]. Finally, how to extract topics from short texts remains a challenging research problem [16]. Three major heuristic strategies have been adopted to deal with how to discover the latent topics from short texts. One follows the simple assumption that each text is sampled from only one latent topic which is totally unsuited to long texts, but it can be suitable for short texts compared to the complex assumption that each text is modeled over a set of topics [19,21]. Therefore, many models for short texts were proposed based on this simple assumption [4,20]. Zhao et al. [21] proposed a Twitter-LDA model by assuming that one tweet is generated from one topic. But, the problem of very limited word co-occurrence information in short texts has not been solved yet. The second strategy takes advantage of various heuristic ties among short texts to aggregate them into long pseudo-texts before topic inference that can help improve word co-occurrence information [8,17]. For example, some models aggregated all the tweets of a user as a pseudo-text [17]. As these tweets with the same hashtag may come from a topic, Mehrotra et al. [8] aggregated all tweets into a pseudo-text based on hashtags. However, these schemes are heuristic and highly dependent on the data, which is not fit for short texts such as news titles, advertisements or image captions. The last scheme directly aggregates short texts into long pseudo-texts through clustering methods [15], in which the clustering method will face this same problem of very limited word co-occurrence information.

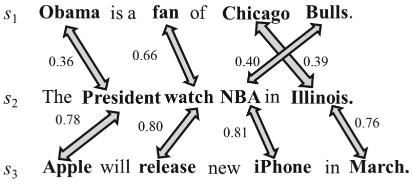
s_1 Obama is a fan of Chicago Bulls.
s_2 The President watch NBA in Illinois.
s_3 Apple will release new iPhone in March.
0.36
0.66
0.40
0.39
0.78
0.80
0.81
0.76

Fig. 1. An illustration of the relationship among short texts.

Figure 1 shows an example to explain the shortcomings of existing short text topic models. There are three short texts, and non-stop words are marked in bold. The shortest distances between two words from different short texts are labeled using the arrows, in which the distance is computed by word embeddings [13]. We can see s_1 and s_2 probably include two topics. 'Obama' and 'President' are likely to come from the same topic, and 'NBA' and 'Bulls' are from another topic. The simple assumption that each text is sampled from only one latent topic is unsuited to these texts. And if we directly aggregate the three short texts into two long pseudo-texts, it is very hard to decide how to aggregate these texts since they do not share the same words. But, it is very clear that s_1 is more similar to s_2 than s_3.

To overcome these inherent weaknesses and keep the advantages of three strategies, we propose a novel method, Embedding-based Topic Model (ETM), to discover latent topics from short texts. Our method leverages recent results by word embeddings that obtain vector representations for words [9]. ETM has the following three steps. ETM firstly builds distributed word embeddings from a large corpus, and then aggregates short texts into long pseudo-texts by incorporating the semantic knowledge from word embeddings, thus alleviates the problem of very limited word co-occurrence information in short texts. Finally, ETM discovers latent topics from pseudo-texts based on the complex assumption that each text of a collection is modeled over a set of topics. ETM adopts a Markov Random Field regularized model based on collapsed Gibbs sampling which utilizes word embeddings to improve the coherence of topic modeling. Within a long pseudo-text, if two words are labeled as similar according to word embedding, a binary potential function is defined to encourage them to share the same latent topic. Experiments demonstrate that ETM can discover more prominent and coherent topics than the baselines.

2 Algorithm

Our model includes three steps. First, we build distributed word embeddings for the vocabulary of the collection. Different from Word2Vec [9] that only utilizes local context windows, Pennington et al. later introduced a new global log-bilinear regression model, Glob2Vec [13], which combines global word-word co-occurrence counts and local context windows. Therefore, we adopt Glob2Vec to learn word vector representation. Second, we aggregate short texts into long pseudo-texts by incorporating the semantic knowledge from word embeddings. A new metric, Word Mover's Distance (WMD) [7], to compute the distance between two short texts. Third, we adopt a Markov Random Field regularized model based on collapsed Gibbs Sampling to improve the coherence of topic modeling. The framework of ETM is shown in Fig. 2.

2.1 Aggregate Short Texts into Long Pseudo-texts

After obtaining word embeddings of each word, we use the typical cosine distance measure for the distance between words, i.e., for word vector v_x and word vector v_y, we define the distance $d(v_x, v_y) = 1 - \frac{v_x}{\|v_x\|_2} \times \frac{v_y}{\|v_y\|_2}$. Consider a collection of short texts, $S = \{s_1, s_2, \ldots, s_i, \ldots, s_n\}$, for a vocabulary of V words, where s_i represents the i^{th} text. We assume each text is represented as a normalized bag-of-words (nBOW) vector, $\mathbf{r}_i \in \mathbb{R}^V$ is the vector of s_i, a V-dimension vector, $r_{i,j} = \frac{c_{i,j}}{\sum_{v=1}^{V} c_{i,v}}$ where $c_{i,j}$ denotes the occurrence times of the j^{th} word of the vocabulary in text s_i. We can see that a nBOW vector is very sparse as only a

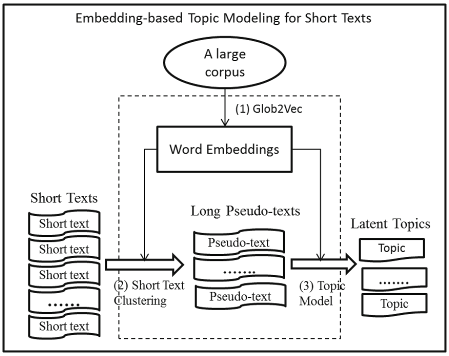

Fig. 2. Embedding-based topic model for short texts

few words appear in each text. For example, given three short texts in Fig. 1, if we adopt these metrics (e.g., Euclidean distance, Cosine Similarity) to measure distance between two texts, it is hard to find their difference. Therefore, we introduce WMD to compute the distance between texts. WMD computes the minimum cumulative cost that words from one text need to travel to match exactly the words of the other text as the distance of texts, in which the distance bewteen words is computed by word embeddings.

Let $\mathbf{r}_i$ and $\mathbf{r}_j$ be the nBOW representation of s_i and s_j. Each word of $\mathbf{r}_i$ can be allowed to travel to the word of $\mathbf{r}_j$. Let $T \in \mathbb{R}^{m \times m}$ be a flow matrix, where $T_{u,v}$ represents how much of the weight of word u of $\mathbf{r}_i$ travels to word v of $\mathbf{r}_j$. To transform all weights of $\mathbf{r}_i$ into $\mathbf{r}_j$, we guarantee that the entire outgoing flow from vertex u equals to $r_{i,u}$, namely $\sum_v T_{u,v} = r_{i,u}$. Correspondingly, the amount of incoming flow to vertex v must equal to $r_{j,v}$, namely, $\sum_u T_{u,v} = r_{j,v}$. At last, we can define the distance of two texts as the minimum cumulative cost required to flow from all words of one text to the other text, namely, $\sum_{u,v} T_{u,v} d(u,v)$. The best average time complexity of solving the WMD problem is O(m^3log$m$), where m is the number of unique words in the text. To speed up the optimization problem, we relax the WMD optimization problem and remove one of the two constraints. Consequently, the optimization becomes,

$$\min_{T \geq 0} \sum_{u,v}^{m} T_{u,v} d(u,v) \quad s.t. \sum_{v}^{m} T_{u,v} = r_{i,u} \forall u \in \{1, 2, ..., m\} \tag{1}$$

The optimal solution is the probability of each word in one text is moved to the most similar word in the other text. The time complexity of WMD can be reduced to O($m$log$m$). Once the distance between texts have been computed, we aggregate short texts into long pseudo-texts based on K-Means algorithm [1].

2.2 Topic Inference by Incorporating Word Embeddings

Model Description: We adopt the MRF model to learn the latent topics which can incorporate word distances into topic modeling for encouraging words labeled similarly to share the same topic assignment [18]. Here, we continue to use word embeddings to compute the distance between words. We can see from Fig. 3, MRF model extends the standard LDA model [2] by imposing a Markov Random Field on the latent topic layer.

Suppose the corpus contains K topics and long pseudo-texts with L texts over V unique words in the vocabulary. Following the standard LDA, Φ is represented by a $K \times V$ matrix where the kth row ϕ_k represents the distribution of words in topic k, Θ is represented by a $L \times K$ where the lth row θ_l represents the topic distribution for the lth long pseudo-texts, α and β are hyperparameters, z_{li} denotes the topic identities assigned to the i_{th} word in the lth long pseudo-text.

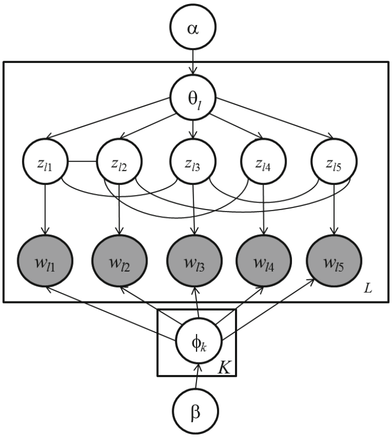

Fig. 3. Markov random field regularized model

The key idea is that if the distance between two words in one pseudo-text is smaller than a threshold, they are more likely to belong to the same topic. For example, in Fig. 1, 'President' and 'Obama' ('Bulls' and 'NBA') are likely to belong to the same topic. Based on this idea, MRF model defines a Markov Random Field over the latent topic. Given a long pseudo-text l consisting of n_l words $\{w_{li}\}_{i=1}^{n_l}$. If the distance between any word pair (w_{li}, w_{lj}) in l is smaller than a threshold, MRF model creates an undirected edge between their topic assignments (z_{li}, z_{lj}). Finally, MRF creates an undirected graph G_l for the lth pseudo-text, where nodes are latent topic assignments $\{z_{li}\}_{i=1}^{n_l}$ and edges connect the topic assignments of correlated words. For example, in Fig. 3, G_l is consisted of five nodes $(z_{l1}, z_{l2}, z_{l3}, z_{l4}, z_{l5})$ and five edges $\{(z_{l1}, z_{l2},), (z_{l1}, z_{l3},), (z_{l2}, z_{l4},), (z_{l2}, z_{l5},), (z_{l3}, z_{l5})\}$.

The same to LDA, MRF model uses the unary potential for z_{li} as $p(z_{li} \mid \theta_l)$. The difference is MRF model defines binary potential over each edge (z_{li}, z_{lj}) of G_l as $exp\{\mathcal{I}(z_{li} = z_{lj})\}$, which produces a large value if the two topic assignments are the same and generates a small value if the two topic assignments are different, where $\mathcal{I}(\cdot)$ is the indicator function. Hence, similar words in one pseudo-text have a high probability to be put into the same topic. The joint probability of all topic assignments $\mathbf{z}_l = \{z_{li}\}_{i=1}^{n_l}$ in MRF model can be calculated as

$$p(\mathbf{z}_l \mid \theta_l, \lambda) = \prod_{i=1}^{n_l} p(z_{li} \mid \theta_l) exp\{\lambda \frac{\sum_{(li,lj)\in\mathcal{P}_l} \mathcal{I}(z_{li} = z_{lj})}{\mid \mathcal{P}_l \mid}\} \tag{2}$$

where $\mathcal{P}_l$ represents all edges of G_l and $\mid \mathcal{P}_l \mid$ is the number of all edges. Here, λ is a user-specified parameter that controls the tradeoff between unary potential and binary potential. If $\lambda = 0$, MRF model is reduced to LDA. Different from LDA that topic label z_{li} is determined by topic distribution θ_l, z_{li} in MRF depends on both θ_l and the topic assignments of similar words in the lth pseudo-text.

Formally, the generative process of MRF model is described as follows.

(1) Draw $\Theta \sim$ Dirichlet(α)
(2) For each topic $k \in [1, K]$
  (a) draw $\phi_k \sim$ Dirichlet(β)
(3) For each pseudo-text l in long pseudo-texts
  (a) draw topic assignments $\mathbf{z}_l$ for all words in pseudo-text l using Eq. (2)
  (b) draw $w_{li} \sim$ Multinomial($\phi_{z_{li}}$) for each word in lth pseudo-text

There have been a number of inference methods that have been used to estimate the parameters of topic models, from basic expectation maximization [6], to approximate inference methods like Variational Inference [2] and Gibbs sampling [5]. Variational Inference tends to approximate some of the parameters, such as Φ and Θ, not explicitly estimate them, may face the problem of local optimum. Therefore, we will use collapsed Gibbs sampling to estimate parameters under Dirichlet priors in this paper, not variational inference [18].

These parameters that need to be estimated include the topic assignments of $\mathbf{z}$, the multinomial distribution parameters Φ and Θ. Using the technique of collapsed Gibbs sampling, we only need to sample the topic assignments of $\mathbf{z}$ by integrating out ϕ and θ according to the following condition distribution:

$$p(z_{li} = k \mid \mathbf{z}_{l,-li}, \mathbf{w}_{l,-li}) = (n_{l,-li}^{k} + \alpha) \frac{n_{k,-li}^{w_{li}} + \beta}{n_{k,-li} + V\beta} exp(\lambda \frac{\sum_{j\in\mathcal{N}_{li}} (z_{lj} = k)}{\mid \mathcal{N}_{li} \mid}) \tag{3}$$

where z_{li} denotes the topic assignment for word w_{li} in the lth pseudo-text, $\mathbf{z}_{l,-li}$ denotes the topic assignments for all words except w_{li} in the lth pseudo-text, $n_{l,-li}^{k}$ is the number of times assigned to topic k excluding w_{li} in the lth pseudo-text, $n_{k,-li}^{w_{li}}$ is the number of times word w_{li} assigned to topic k excluding w_{li}, $n_{k,-li}$ is the number of occurrences of all words V that belongs to topic k excluding w_{li}, $\mathcal{N}_{li}$ denotes the words that are labeled to be similar to w_i in the lth pseudo-text, and $\mid \mathcal{N}_{li} \mid$ is the number of words in $\mathcal{N}_{li}$.

Parameter Estimation: There are three types of variables ($\mathbf{z}$, Φ and Θ) to be estimated for our model ETM. For the lth pseudo-text, the joint distribution of all known and hidden variables is given by the hyperparameters:

$$p(\mathbf{z}_l, \theta_l, \mathbf{w}_l, \Phi \mid \alpha, \beta, \lambda) = p(\Phi|\beta) \cdot \prod_{li=1}^{n_l} p(w_{li} \mid \phi_{z_{li}}) \cdot p(\mathbf{z}_l \mid \theta_l, \lambda) \cdot p(\theta_l \mid \alpha) \tag{4}$$

We can obtain the likelihood of the lth pseudo-text $\mathbf{w}_l$ of the joint event of all words by integrating out ϕ and θ and summing over z_{li}.

$$p(\mathbf{w}_l \mid \alpha, \beta, \lambda) = \int\int p(\theta_l \mid \alpha) \cdot p(\Phi|\beta) \cdot \prod_{li=1}^{n_l} p(w_{li} \mid \phi_{z_{li}}, \Phi, \lambda) \tag{5}$$

Finally, the likelihood of all pseudo-texts $\mathbf{W} = \{\mathbf{w}_l\}_{l=1}^{L}$ is determined by the product of the likelihood of the independent pseudo-texts:

$$p(\mathbf{W} \mid \alpha, \beta, \lambda) = \prod_{l=1}^{L} p(\mathbf{w}_l \mid \alpha, \beta, \lambda) \tag{6}$$

We try to formally derive the conditional distribution $p(z_{li} = k \mid \mathbf{z}_{l,-li}, \mathbf{w}_{l,-li})$ used in our ETM algorithm as follows.

$$p(z_{li} = k \mid \mathbf{z}_{l,-li}, \mathbf{w}_{l,-li}) = \frac{p(\mathbf{w}, \mathbf{z} \mid \alpha, \beta, \lambda)}{p(\mathbf{w}, \mathbf{z}_{l,-li} \mid \alpha, \beta, \lambda)} \propto \frac{p(\mathbf{w}, \mathbf{z} \mid \alpha, \beta, \lambda)}{p(\mathbf{w}_{l,-li}, \mathbf{z}_{l,-li} \mid \alpha, \beta, \lambda)} \tag{7}$$

From the graphical model of ETM, we can see

$$p(\mathbf{w}, \mathbf{z} \mid \alpha, \beta, \lambda) = p(\mathbf{w} \mid \mathbf{z}, \beta) p(\mathbf{z} \mid \alpha, \lambda) \tag{8}$$

The same to LDA, the target distribution $p(\mathbf{w} \mid \mathbf{z}, \beta)$ is obtained by integrating over ϕ,

$$p(\mathbf{w} \mid \mathbf{z}, \beta) = \prod_{z_{li}=1}^{K} \frac{\Delta(\mathbf{n}_{z_{li}} + \beta)}{\Delta(\beta)}, \mathbf{n}_{z_{li}} = \{n_{z_{li}}^{(w)}\}_{w=1}^{V} \tag{9}$$

where $n_{z_{li}}^{(w)}$ is the number of word w occurring in topic z_{li}. Here, we adopt the Δ function in Heinrich (2009), and we can have $\Delta(\beta) = \frac{\prod_{w=1}^{V} \Gamma(\beta)}{\Gamma(V\beta)}$ and $\Delta(\mathbf{n}_{z_{li}} + \beta) = \frac{\prod_{w \in \mathbf{w}} \Gamma(n_k^w + \beta)}{\Gamma(n_k + V\beta)}$, where Γ denotes the gamma function.

According to Eq. (3), we can get

$$p(\mathbf{z}_l \mid \theta_l, \lambda) = exp\{\lambda \frac{\sum_{(li,lj) \in \mathcal{P}_l} \sum_{k=1}^{K} (z_{li} z_{lj})}{\mid \mathcal{P}_l \mid}\} \prod_{k=1}^{K} \theta_k^{n_l^k} \tag{10}$$

Similarly, $p(\mathbf{z}_l \mid \alpha, \lambda)$ can be obtained by integrating out Θ as

$$\begin{aligned} p(\mathbf{z} \mid \alpha, \lambda) &= \int p(\mathbf{z} \mid \Theta, \lambda) p(\Theta \mid \alpha) \\ &= \prod_{l=1}^{L} exp\{\lambda \frac{\sum_{(li,lj) \in \mathcal{P}_l} \sum_{k=1}^{K} (z_{li} z_{lj})}{\mid \mathcal{P}_l \mid}\} \frac{\Delta(\mathbf{n}_l + \alpha)}{\Delta(\alpha)} \end{aligned} \tag{11}$$

where $p(\Theta \mid \alpha)$ is a Dirichlet distribution, and $\mathbf{n}_l = \{n_l^{(k)}\}_{k=1}^{K}$.

Finally, we put the joint distribution $p(\mathbf{w}, \mathbf{z} \mid \alpha, \beta, \lambda)$ into Eq. (11), the conditional distribution in Eq. (3) can be derived

$$\begin{aligned} p(z_{li} = k \mid \mathbf{z}_{l,-li}, \mathbf{w}_{l,-li}) &\propto \frac{p(\mathbf{w}, \mathbf{z} \mid \alpha, \beta, \lambda)}{p(\mathbf{w}_{l,-li}, \mathbf{z}_{l,-li} \mid \alpha, \beta, \lambda)} \\ &\propto \frac{\Delta(\mathbf{n}_l + \alpha)}{\Delta(\mathbf{n}_{l,-li} + \alpha)} \frac{\Delta(\mathbf{n}_{z_{li}} + \beta)}{\Delta(\mathbf{n}_{z_{l,-li}} + \beta)} exp(\lambda \frac{\sum_{j \in \mathcal{N}_{li}} (z_{lj} = k)}{\mid \mathcal{N}_{li} \mid}) \\ &\propto (n^k_{l,-li} + \alpha) \frac{n^{w_{li}}_{k,-li} + \beta}{n_{k,-li} + V\beta} exp(\lambda \frac{\sum_{j \in \mathcal{N}_{li}} (z_{lj} = k)}{\mid \mathcal{N}_{li} \mid}) \end{aligned} \tag{12}$$

3 Experiments

Datasets and Setup: We study the empirical performance of ETM on two short text datasets, Tweet2011 and GoogleNews[1]. Similar to existing papers [20], we utilize Google news as a dataset to evaluate the performance of topic models. We took a snapshot of the Google news on April 27, 2015, and crawled the titles of 6,974 news articles belonging to 134 categories. For each dataset, we conduct the same preprocessing with this paper [14]. We compare our model ETM with the following baselines. Three short text topic models: Unigrams [12], DMM [20], and BTM [4]. Two Long text topic models: LDA [5] and MRF-LDA [18]. For the baselines, we chooses the parameters according to their original papers. For LDA, Unigrams and BTM, both hyperparameters α and β are set to $50/K$ and 0.01. For DMM and ETM, both hyperparameters α and β are set to 0.1. For MRF-LDA, $\alpha = 0.5$ and $\lambda = 1$. For ETM, λ is set to 1. For our model and MRF-LDA, words pairs with distance lower than 0.4 are labeled as correlated.

A lot of metrics have been proposed for measuring the coherence of topics in texts [11]. Most conventional metrics try to estimate the likelihood of held-out testing data based on parameters inferred from training data. However, this likelihood is not necessarily a good indicator of the quality of extracted topics [3]. Similar to [18], we also evaluate our model in a qualitative and quantitative manner. And we validate topic models on short text clustering and short text classification. Due to the space limit, we omit some experiments. First, we discuss some exemplar topics learned by the six methods on the two datasets. Each topic is visualized by the top ten words. Then, we evaluate our model based on the coherence measure (CM) to assess how coherent the learned topics are. For each topic, we choose the top 10 candidate words and ask human annotators to judge whether they are relevant to the corresponding topic. To do this, annotators need to judge whether a topic is interpretable or not. If not, the 10 words of the topic are labeled as irrelevant; otherwise these words are identified by annotators as relevant words for this topic. Coherence measure (CM) is defined as the ratio between the number of relevant words and the total number of candidate words.

[1] http://news.google.com.

Qualitative Evaluation: On Tweet2011 dataset, there is no category information for each tweet. Manual labeling might be difficult due to the incomplete and informal content of tweets. Fortunately, some tweets are labeled by their authors with hashtags in the form of '#keyword' or '@keyword'. We manually choose 10 frequent hashtags as labels and collect documents with their hashtags. These hashtags are 'NBA', 'NASA', 'Art', 'Apple', 'Barackobama', 'Worldprayr', 'Starbucks', 'Job', 'Travel', 'Oscars', respectively. On GoogleNews dataset, the four topics are events on April 27, 2015, which are "Nepal earthquake", "Iran nuclear", "Indonesia Bali", and "Yemen airstrikes".

Table 1 shows some topics learned by the six models. Each topic is visualized by the top ten words. Words that are noisy and lack of representativeness are highlighted in bold. From Tabel 1, our model ETM can learn more coherent topics with fewer noisy and meaningless words than all baseline models. Long text topic modelings (LDA and MRF-LDA) that model each text as a mixture of topics does not fit for short texts, as short text suffers from the sparsity of word co-occurrence patterns. MRF-LDA incorporating word correlation knowledge cannot improve the coherence of topic modeling since binary potential of MRF cannot work when short text only consists of a few words. In addition, MRF-LDA based on variational inference may face the problem of local optimum. Consequently, the top 10 words of yemeb of LDA and Apple of MRF-LDA are not relevant to the corresponding topic.

The existing short text topic models suffer from two problems. On one hand, the frequency of words in short text plays a less discriminative role than long text, making it hard to infer which words are more correlated in each text. On the other hand, these models bring in little additional word co-occurrence information and cannot alleviate the sparsity problem. As a consequence, the topics extracted from these three short text topic models are not satisfying. For example, Unigrams cannot identify topic "iran", BTM cannot identify topic "yemen", and the learned topics of DMM consists of meaning-less words such as *going*, *today*, etc.

Our method ETM incorporates the word correlation knowledge provided by words embedding over the latent topic to cluster short texts to generate long pseudo-text. In this condition, the frequency of words in pseudo-text plays an important role to discover the topics based on this assumption each text is modeled as a mixture of topics. After aggregating short texts into long pseudo-texts, more similar words are in one text than the original text. Therefore, the Markove Random Field regularized model can paly an important in learning latent topics from pseudo-texts, which uses the word correlation knowledge over the latent topic to encourage correlated words to share the same topic label. Hence, although similar words may not have high co-occurrence in the corpus, they remain have a high probability to be put into the same topic. Consequently, from Table 1 we can see that the topics learned by our model are far better than those learned by the baselines. The learned topics have high coherence and contain fewer noisy and irrelevant words. Our model also can recognize the topic words that only have a few occurrences in the collection. For instance, the word *flight* from topic "NASA", *writer* from topic "Art", and *tablet* of topic "Apple" can only be recognized by ETM.

Table 1. Topics learned from Tweet2011 and GoogleNews dataset

Data	Class	Method	Top 10 words
Tweet2011	NBA	LDA	Game lebron kobe player lakers team coach **going** james points
		MRF-LDA	Game lebron kobe player **museum** lakers play **tonight** james **better**
		Unigrams	Game lebron kobe player lakers team **going** james play allen
		DMM	Game lebron kobe player lakers team james points **going lead**
		BTM	Game kobe lebron lakers team player scored points **going** james
		ETM	Game lebron kobe player lakers team points james play allen
	NASA	LDA	Space shuttle launch nasa atlantis **live** video weather **watch check**
		MRF-LDA	Space shuttle **great** launch **good** nasa **store watch today** atlantis
		Unigrams	Space shuttle launch nasa atlantis **check live watch** weather crew
		DMM	Space shuttle launch nasa atlantis **live check** video weather **today**
		BTM	Space shuttle launch nasa atlantis **live** crew weather **watch** image
		ETM	Space shuttle launch nasa flight weather atlantis crew image ares
	Art	LDA	Artist museum **great check** photo **blog** artists gallery painting modern
		MRF-LDA	**Time** Artist **video twitter blog year record coming** work artists
		Unigrams	Artist museum **good** artists painting photo **blog check** gallery exhibition
		DMM	Artist museum **check** photo painting artists exhibition modern gallery **blog**
		BTM	Artist **great** museum **check miami** painting artists gallery **blog free**
		ETM	Artist museum writer painting gallery artists exhibition modern photo arts
	Apple	LDA	Apple iphone store **time** steve jobs snow **best good google**
		MRF-LDA	**Apple iphone check team live love follow star coach going**
		Unigrams	Apple iphone store **time good** ipod jobs video snow steve
		DMM	Apple iphone store **time** steve snow jobs **google great good**
		BTM	Apple iphone **good** steve video store **time** jobs ipod **going**
		ETM	Apple iphone store video ipod **twitter** tablet steve **blog google**
GoogleNews	Nepal	LDA	Nepal death israel quake rescue aid israeli israelis relief help
		MRF-LDA	Nepal death israel quake rescue aid everest israeli israelis relif
		Unigrams	Nepal quakes toll death quake everest tops aid rises israelis
		DMM	Nepal israelis quake toll israel rescue death aid everest israeli
		BTM	Nepal aids quake rescue israel toll death aid everest israeli
		ETM	Nepal israelis quake toll israel rescue death aid everest israeli
	Iran	LDA	Iran nuclear meet kerry zarif talks **good** israel **foreign deal**
		MRF-LDA	Iran meet kerry nuclear zarif **deal victim** talks powers arms
		Unigrams	**Nepal quakes israel rescue quake aid israeli relief help good**
		DMM	Iran nuclear meet kerry zarif israel weapon npt **deal foreign**
		BTM	**Yeman** Iran nuclear meet **saudi kerry** arms talks zarif **good**
		ETM	Iran nuclear meet kerry zarif israel weapon npt talks powers
	Bali	LDA	Bali Indonesia execution excecutions chan marries andrew duo death **deal**
		MRF-LDA	Bali Indonesia death **toll** execution executions chan andrew marries **nuclear**
		Unigrams	Bali Indonesia execution executions chan marries andrew death duo drug
		DMM	Bali Indonesia execution executions chan marries andrew death duo drug
		BTM	Bali chan executions andrew marries sukumaran **final** duo myuran **ahead**
		ETM	Bali Indonesia execution executions chan marries andrew death duo drug
	Yemen	LDA	**Yemen Nepal toll death saudi quake quakes Iran strikes tops**
		MRF-LDA	Yemen Saudi talks **drug iran** strikes yemeni war saudis **babies**
		Unigrams	Yemen **Iran nuclear meet** kerry saudi **zarif** talks arms **israel**
		DMM	Yemen Saudi **Iran** strikes yemeni saudis talks strike **tops** houthis
		BTM	**Nepal Bali chan drug aids arms chaims death duo pair**
		ETM	Yemen Saudi strikes yemeni saudis strikes war talks houthis arms

Table 2. CM (%) on Tweet2011 and GooleNews (Ai represents the ith annotator)

Method	Tweet2011					GoogleNews				
	A1	A2	A3	A4	Mean	A1	A2	A3	A4	Mean
LDA	54	42	45	67	52 ± 11.2	95	95	79	95	91 ± 8
MRF-LDA	44	46	46	57	48.2 ± 5.9	91	85	80	74	82.5 ± 7.2
Unigrams	66	45	56	59	56.5 ± 8.7	88	73	79	94	83.5 ± 9.3
DMM	70	49	50	60	57.2 ± 9.8	94	93	90	93	92.5 ± 1.7
BTM	62	45	50	77	58.5 ± 14.2	80	85	75	78	79.5 ± 4.2
ETM	**72**	**62**	**73**	**83**	**72.5 ± 8.5**	**96**	**96**	**94**	**96**	**95.5 ± 1.0**

Quantitative Evaluation: Table 2 shows the coherence measure of topics inferred on Tweet2011 and GoogleNews datasets, respectively. We can see our model ETM significantly outperforms the baseline models. On Tweet2011 dataset, ETM achieves an average coherence measure of 72.5%, which is larger than long text topic models (LDA and MRF-LDA) with a large margin. Compared to short text topic models, ETM still has a big improvement. In GoogleNews dataset, our model is also much better than the baselines.

4 Conclusion

We propose a novel model, Embedding-based Topic Modeling (ETM), to discover the topics from short texts. ETM first aggregates short texts into long pseudo-texts by incorporating the semantic knowledge from word embeddings, then infers topics from long pseudo-texts using Markov Random Field regularized model, which encourages words labeled as similar to share the same topic assignment. Therefore, by incorporating the semantic knowledge ETM can alleviate the problem of very limited word co-occurrence information in short texts.

Acknowledgement. This research is partially supported by the National Key Research and Development Program of China (2016YFB1000900), and the National Natural Science Foundation of China (No. 61503116).

References

1. Aggarwal, C.C., Zhai, C.: A survey of text clustering algorithms. In: Mining Text Data, pp. 77–128 (2012)
2. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. JMLR **3**, 993–1022 (2003)
3. Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., Blei, D.M.: Reading tea leaves: how humans interpret topic models. In: NIPS, pp. 288–296 (2009)
4. Cheng, X., Yan, X., Lan, Y., Guo, J.: BTM: topic modeling over short texts. TKDE **26**(12), 2928–2941 (2014)

5. Griffiths, T., Steyvers, M.: Finding scientific topics. Proc. Nat. Acad. Sci. **101**, 5228–5235 (2004)
6. Hofmann, T.: Probabilistic latent semantic indexing. In: SIGIR, pp. 50–57 (1999)
7. Kusner, M.J., Sun, Y., Kolkin, N.I., Weinberger, K.Q.: From word embeddings to document distances. In: ICML, pp. 957–966 (2015)
8. Mehrotra, R., Sanner, S., Buntine, W., Xie, L.: Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In: SIGIR, pp. 889–892 (2013)
9. Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. In: NIPS, pp. 3111–3119 (2013)
10. Newman, D., Bonilla, E.V., Buntine, W.: Improving topic coherence with regularized topic models. In: NIPS, pp. 496–504 (2011)
11. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: NAACL, pp. 100–108 (2010)
12. Nigam, K., McCallum, A.K., Thrun, S., Mitchell, T.: Text classification from labeled and unlabeled documents using EM. Mach. Learn. **39**(2–3), 103–134 (2000)
13. Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: EMNLP, pp. 1532–1543 (2014)
14. Qiang, J., Chen, P., Ding, W., Wang, T., Fei, X., Wu, X.: Topic discovery from heterogeneous texts. In: ICTAI (2016)
15. Quan, X., Kit, C., Ge, Y., Pan, S.J.: Short and sparse text topic modeling via self-aggregation. In: ICAI, pp. 2270–2276 (2015)
16. Wang, X., Wang, Y., Zuo, W., Cai, G.: Exploring social context for topic identification in short and noisy texts. In: AAAI (2015)
17. Weng, J., Lim, E.-P., Jiang, J., He, Q.: Twitterrank: finding topic-sensitive influential twitterers. In: WSDM, pp. 261–270 (2010)
18. Xie, P., Yang, D., Xing, E.P.: Incorporating word correlation knowledge into topic modeling. In: NACACL (2015)
19. Yan, X., Guo, J., Lan, Y., Xu, J., Cheng, X.: A probabilistic model for bursty topic discovery in microblogs. In: AAAI, pp. 353–359 (2015)
20. Yin, J., Wang, J.: A dirichlet multinomial mixture model-based approach for short text clustering. In: SIGKDD, pp. 233–242 (2014)
21. Zhao, W.X., Jiang, J., Weng, J., He, J., Lim, E.-P., Yan, H., Li, X.: Comparing Twitter and traditional media using topic models. In: Clough, P., Foley, C., Gurrin, C., Jones, G.J.F., Kraaij, W., Lee, H., Mudoch, V. (eds.) ECIR 2011. LNCS, vol. 6611, pp. 338–349. Springer, Heidelberg (2011). doi:10.1007/978-3-642-20161-5_34