Short and Sparse Text Topic Modeling via Self-Aggregation

Xiaojun Quan[1], Chunyu Kit[2], Yong Ge[3], Sinno Jialin Pan[4]
[1]Institute for Infocomm Research, A*STAR, Singapore
[2]City University of Hong Kong, Hong Kong SAR, China
[3]UNC Charlotte, NC, USA
[4]Nanyang Technological University, Singapore
[1]quanx@i2r.a-star.edu.sg, [2]ctckit@cityu.edu.hk, [3]yong.ge@uncc.edu, [4]sinnopan@ntu.edu.sg

Abstract

The overwhelming amount of short text data on social media and elsewhere has posed great challenges to topic modeling due to the sparsity problem. Most existing attempts to alleviate this problem resort to heuristic strategies to aggregate short texts into pseudo-documents before the application of standard topic modeling. Although such strategies cannot be well generalized to more general genres of short texts, the success has shed light on how to develop a generalized solution. In this paper, we present a novel model towards this goal by integrating topic modeling with short text aggregation during topic inference. The aggregation is founded on general topical affinity of texts rather than particular heuristics, making the model readily applicable to various short texts. Experimental results on real-world datasets validate the effectiveness of this new model, suggesting that it can distill more meaningful topics from short texts.

1 Introduction

Short texts have been an important form of information carrier in modern society, widely observed in a wide range of web services from online advertising, instant messaging and email to recent vogues like social media. These texts are typically characterized by short length, informality and noise. Analyzing unannotated short texts is an effective means to acquire valuable insights from big text archives and to decipher their vast amounts of information. However, because of their huge size, such data cannot be handled by normal human power and hence demands effective and efficient tools. Topic modeling has proven to be instrumental in automatic discovery of thematic information from large archives of documents. It views documents as mixtures of probabilistic topics, where a topic is a probability distribution over words [Blei, 2012]. These topic components uncover certain latent structures in document collections and can be inferred by standard statistical inference. Typical topic modeling techniques such as latent Dirichlet allocation (LDA) have demonstrated great successes on long documents, but unfortunately, they have not been able to work very well on short texts [Hong and Davison, 2010; Zhao *et al.*, 2011]. This is mainly due to the fact that only very limited word co-occurrence information is available in such short and sparse texts as tweets compared with long documents [Wang and McCallum, 2006].

Two major heuristic strategies have been adopted to deal with the sparsity problem. One follows the relaxed assumption that each short text snippet is sampled from only one latent topic, an assumption adopted by early topic models such as mixture of unigrams [Nigam *et al.*, 2000]. Although this assumption does not fit long documents quite well [Blei *et al.*, 2003], it is suited for certain short texts and can help to alleviate the sparsity problem to certain extent [Zhao *et al.*, 2011]. The other strategy, widely used on social medial to cope with short texts, takes advantage of various heuristic ties between short text snippets to aggregate them into long pseudo-documents before a standard topic model is applied. Taking Twitter as an example, there is a plentiful set of such context information as hashtag, authorship, time, and location associated with tweets that can be used for the aggregation [Hong and Davison, 2010; Weng *et al.*, 2010; Mehrotra *et al.*, 2013]. However, this strategy cannot be easily generalized to deal with more general forms of short texts, e.g., questions and search queries, which are widely observable in many domains but hardly contain any useful tie.

Another conceivable way out is to enrich short texts using highly relevant long documents [Jin *et al.*, 2011; Guo *et al.*, 2013]. One may also exploit various automatic query expansion techniques broadly used in information retrieval to expand short texts [Xu and Croft, 1996; Voorhees, 1994]. Yet to the best of our knowledge, there has not been any effort towards this. A recent attempt by Yan et al. [2013] to provide a generalized topic model for short texts has come up with a new model by ignoring document identities and directly modeling word co-occurrences in short texts. However, this model brings in little additional word co-occurrence information and cannot alleviate the sparsity problem essentially.

This research is motivated by the success of heuristically aggregating short text snippets on social media for better topic modeling, and aims to provide a generalized solution for topic modeling of short texts of various forms. The rationale behind the aggregation lies in that more useful word co-occurrences can be created through effective aggregation of short texts with similar topics, leading to a solution that can potentially tackle the sparsity problem. Going beyond the general machinery of standard topic models, we further assume that each piece of short text snippet is sampled from a long pseudo-document unobserved in current text collection. The key to extract meaningful and interpretable topics is to find the right "documentship" for each text snippet. In our model, this is to be achieved by means of organic integration of topic modeling and text self-aggregation during topic inference, with the aggregation built upon general topical affinity of texts and applicable to various short texts.

In addition, we propose two new evaluation criteria in view of the deficiencies of existing evaluation metrics for short text topic modeling. We evaluate our model on two datasets of sentence-level short texts from different domains and compare it with other topic models. Experimental results confirm the capability of the new model in extracting meaningful topics from short and sparse texts.

2 Related Work

Most existing work on topic modeling of short texts uses data from social media, where messages are generally very short and associated with plentiful context information, enabling a straightforward solution without revising the basic machinery of standard topic models. More specifically, such context information as usership and hashtag can be employed to aggregate short messages into long pseudo-documents before standard topic modeling is applied.

One may naturally question whether the aggregation based on such context information is helpful enough to give rise to useful pseudo-documents for better topic modeling. Nevertheless, studies have shown that sometimes the aggregation is quite necessary and beneficial [Hong and Davison, 2010; Weng *et al.*, 2010]. For example, in the work of finding influential users on Twitter, Weng et al. [2010] aggregated tweets from the same user into a new pseudo-document and then performed standard topic modeling on them. Since their focus was on the topics that are of interest to each user rather than the topics in each tweet, the aggregation is sound. In another similar study by Mehrotra et al. [2013], hashtag was shown to be the best context information among others for tweets aggregation to yield meaningful topics. This is most likely because hashtags on Twitter are used to identify tweets with the same topics, and accordingly the generated pseudo-documents tend to be more coherent than those otherwise. Instead of using the multiple types of contexts separately as do the above models, Tang et al. [2013] proposed a multi-contextual topic model that generates both context-specific topics and consensus topics across contexts.

However, the above approaches cannot be readily applied to more general forms of short texts which provide hardly any such context information for use. To the best of our knowledge, the first effort towards a generalized short text topic model was made by Yan et al. [2013]. To maximize the usage of existent word co-occurrence information in short texts, their model directly captures word co-occurrence patterns during topic modeling. However, this model tends not to differentiate between long documents and short texts and creates little new word co-occurrence information for alleviating the sparsity problem. In contrast, the current work is built upon previous success on social media through tweets aggregation, and consequently, the new topic model is able to perform automatic self-aggregation of short texts during topic modeling. In addition, the aggregation is based on general topical affinity of texts, allowing a more generalized solution.

3 Topic Modeling via Self-Aggregation

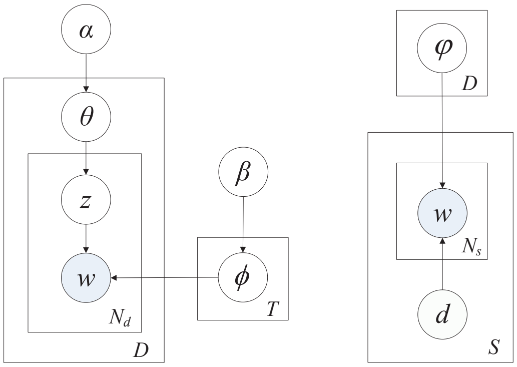
As mentioned before, the success of topic modeling on social media through heuristic aggregation of tweets has shed light on how to develop a generalized topic modeling solution for short texts. Motivated by this, a natural strategy would be to resort to automatic clustering algorithms for aggregating short texts into long pseudo-documents, before a standard topic model is applied. However, such solutions have at least two drawbacks. First, data sparsity is still an unavoidable issue for most clustering algorithms and needs to be taken care of beforehand [Jin *et al.*, 2011]. Second, the process of clustering and topic modeling would be separated, resulting in clusters of short texts that are likely to have only superficial affinity within each cluster but no latent topical relatedness.

In this section, we present a self-aggregation based topic model (SATM) for short and sparse texts by natural integration of clustering and topic modeling. In particular, we assume that short texts are the consequences of randomly crumbling long documents, which are generated using a standard topic model. In this sense, each piece of short text snippet can be considered to be sampled from an unobserved long pseudo-document. Finding the correspondence between the observed text snippet and the hidden pseudo-document is thus very critical for successful topic modeling. In our model, the correspondence is characterized in a way to follow the assumption of the mixture of unigrams model [Nigam *et al.*, 2000] that a document is sampled from only one latent topic rather than many as do many more advanced topic models. Although this simple assumption does not fit long documents very well [Blei *et al.*, 2003], surprisingly, it is suited to short texts of certain scenario [Zhao *et al.*, 2011]. More importantly, it is consonant with our assumption that each short text snippet is sampled from only one long pseudo-document.

In the following paragraphs we will first state the basic assumption employed by SATM on how to describe the generative process for a collection of short text snippets. Then, we introduce the inference of the involved hidden variables by means of Gibbs sampling.

3.1 Model

The generative process of SATM for short texts can be basically described in two phases. As shown in Figure 1(a), the first phase follows the assumption of standard topic models (e.g., LDA) to generate a set of D regular-sized documents, where each document d is composed of a word sequence, $\mathbf{w}_d$, of size N_d. In the second phase, each document will be used to generate a few short text snippets, as described in Figure 1(b), which corresponds to the assumption that each text snippet is sampled from a long document following the multinomial distribution. That means for each short text snippet s containing a sequence, $\mathbf{v}_s$, of N_s words, there is exactly one long document being responsible for its generation. The above generative procedure can be described as follows:



(a) Phase 1 (b) Phase 2

Figure 1: Graphical representation of the new topic model.

1. For each latent topic z:
   (a) Sample a multinomial distribution over words $\phi_z \sim \mathrm{Dirichlet}(\beta)$.
2. Sample a topic proportion $\theta_d \sim \mathrm{Dirichlet}(\alpha)$.
3. For each word $w \in \mathbf{w}_d$:
   (a) Sample a topic $z_w \sim \mathrm{Multinomial}(\theta_d)$.
   (b) Sample a word $w \sim \mathrm{Multinomial}(\phi_{z_w})$.
4. For each word $w \in \mathbf{v}_s$:
   (a) Sample a word from φ_d, the probability distribution over words for document d.

where α and β are hyperparameters, and ϕ_{z_w} refers to the multinomial distribution over words for topic z_w. Steps 1-3 of the procedure correspond to the first phase of generating long documents, and Step 4 corresponds to the second phase of generating short texts from these long documents.

In addition to the same hidden variables of z, ϕ, and θ as in the traditional topic models, SATM also involves d and φ as new hidden variables. Consequently, the key problem in the topic modeling of SATM is to estimate the posterior distribution of the hidden variables θ_d, d_s, and $\mathbf{z}_s$ for a given piece of short text snippet s, which amounts to

$$p(\theta_d, d_s, \mathbf{z}_s|\mathbf{v}_s, \alpha, \beta) = \frac{p(\theta_d, d_s, \mathbf{z}_s, \mathbf{v}_s|\alpha, \beta)}{p(\mathbf{v}_s|\alpha, \beta)}, \qquad (1)$$

where $\mathbf{z}_s$ denotes the sequence of topic identities assigned to the words of s, and d_s is the hidden document for generating s. It is generally intractable to compute this distribution because the normalization factor, $p(\mathbf{v}_s|\alpha, \beta)$, cannot be computed exactly. Fortunately, there have been a number of approximate inference techniques such as variational inference [Blei *et al.*, 2003] and Gibbs Sampling [Griffiths and Steyvers, 2004] that can be used to solve this problem. In what follows we will describe how to use Gibbs sampling to solve the above problem and to infer the hidden variables including a set of latent topics from short texts.

3.2 Gibbs Sampling

The Gibbs sampling process in SATM can be described in two essential steps. Before going into the details, we first use $\mathcal{W}$, $\mathcal{S}$ and $\mathcal{D}$ to represent a word vocabulary, a collection of short texts, and a collection of pseudo-documents, respectively. The role of the first step is to build correspondences between short texts and hidden pseudo-documents by maintaining a $\mathcal{S} \times \mathcal{D}$ relation matrix, which indicates how short texts are likely to be aggregated. To be more concrete, this step updates the $\mathcal{S} \times \mathcal{D}$ relation matrix based on the $\mathcal{W} \times \mathcal{D}$ relation derived from a previous iteration. In the second step, the new $\mathcal{S} \times \mathcal{D}$ matrix will be taken to the inference of topic assignments for words under the Markov Chain Monte Carlo framework. This step is intended to perform standard topic modeling based on the above "aggregated" short texts.

More specifically, the ij-th entry of the $\mathcal{S} \times \mathcal{D}$ matrix is denoted by the probability of the occurrence of a pseudo-document $d^{(j)}$ in $\mathcal{D}$ conditioned on a piece of short text snippet $s^{(i)}$ in $\mathcal{S}$, and is to be estimated following the mixture of unigrams model [Nigam *et al.*, 2000]:

$$p(d^{(j)}|s^{(i)}) = \frac{p(d^{(j)}) \prod_{k=1}^{W} p(w^{(k)}|d^{(j)})^{r_{ik}}}{\sum_{m=1}^{D} p(d^{(m)}) \prod_{n=1}^{W} p(w^{(n)}|d^{(m)})^{r_{in}}}, \qquad (2)$$

where $p(d^{(j)})$ refers to the probability of pseudo-document $d^{(j)}$, and $p(w^{(k)}|d^{(j)})$ is the probability of word $w^{(k)}$ conditioned on $d^{(j)}$, to form a component of $\varphi_{d^{(j)}}$. Both of the two probabilities are estimated based on the assignments in the previous iteration. W is the size of the word vocabulary and r_{ik} is the frequency of word $w^{(k)}$ in $s^{(i)}$. For the sake of simplicity, we use η_{ij} to represent $p(d^{(j)}|s^{(i)})$.

The second step of Gibbs sampling needs to deal with two sets of unobserved variables, d and z. In our work, a similar strategy as the author-topic model [Rosen-Zvi *et al.*, 2004] is employed to draw a pair of d_i and z_i jointly for the ith word token w_i, meaning that the assignments of pseudo-documents for words will also be carried out at the same time. Accordingly, each (d_i, z_i) pair is drawn as a block like:

$$p(d_i = j, z_i = k|w_i = m, s_i, \mathbf{d}_{-i}, \mathbf{z}_{-i}) \propto \eta_{s_i j} \cdot \frac{U_{kj} + \alpha}{U_{\cdot j} + T\alpha} \cdot \frac{V_{mk} + \beta}{V_{\cdot k} + W\beta}, \qquad (3)$$

where $w_i = m$ denotes that the current work token corresponds to the mth word in the word vocabulary, s_i is the short text snippet containing w_i, and $d_i = j$ and $z_i = k$ represent the assignments of word w_i to pseudo-document j and topic k, respectively. Besides, $\mathbf{d}_{-i}$ and $\mathbf{z}_{-i}$ refer to all pseudo-document and topic assignments not including the current token, matrices U and V denote topic-pseudo-document and word-topic assignments, respectively, and $U_{\cdot j}$ and $V_{\cdot k}$ are the sums of the jth and kth columns of U and V over rows. Finally, T is the number of latent topics.

The general idea behind the above two steps is to append the aggregations of short texts as some special "states" to the Markov chain of the conventional Gibbs sampling for better topic modeling of short texts, for which convergence can be guaranteed. Note that in order to reduce the computational cost of the above sampling process, a threshold can be set for η (e.g., 0.001 for this work) so that unimportant correspondences between short texts and pseudo-documents are filtered out and the process can be accelerated greatly.

4 Experiment

In this section we report on the evaluation of our new topic model on two corpora of short texts from different domains. We also explore new evaluation metrics in the light of the limitations of existing metrics for short text topic modeling.

4.1 Data

NIPS. The first corpus consists of 1,740 NIPS conference papers over the period of 2000 to 2012, with most of the papers falling in the general area of learning algorithms. The title, abstract and main body of each paper are used, while the acknowledgement, references and bibliography are discarded. By removing words with document frequency below 5, we obtain a word vocabulary of size 10,297. Since these scientific documents are formalized in language and concentrated in content, the topics extracted from them by a standard topic model such as LDA can be considered desirable or even ideal at some level. For the purpose of evaluation, we treat these documents as "unobserved" pseudo-documents and use them to generate a set of short texts by crumbling each document into sentences, resulting in a set of 200,879 pieces of short text snippets. In a sense, a short text topic model can be considered good if it is able to extract as meaningful and interpretable topics from these short texts as those from the original long documents.

Yahoo! Answers. This corpus, crawled from Yahoo! Answers[1], consists of 88,120 questions from 11 categories, with each category containing from 2,243 to 23,352 questions. These questions are generally very short and each contains only several meaningful words after removal of stop words. Words appearing in less than 3 questions are also discarded. This gives rise to a dictionary of 5,972 words. The resulting dataset will be used for short text classification in a way to evaluate the proposed topic model on a specific task.

4.2 New Evaluation Metrics

Finding an effective evaluation metric for topic models is not as straightforward as it might look like. 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 [Chang *et al.*, 2009]. Recently, some new metrics have been proposed by measuring the coherence of topics in documents [Newman *et al.*, 2010; Mimno *et al.*, 2011]. For example, one can use pointwise mutual information to measure the coherence of a topic z_i [Newman *et al.*, 2010] as

$$Coh_i = \frac{2}{K(K-1)} \sum_{j<k\leq K} \log \frac{p(w_j, w_k)}{p(w_j)p(w_k)}, \tag{4}$$

where K is the number of most probable words in each topic, $p(w_j, w_k)$ is the probability of words w_j and w_k co-occurring in a document, and $p(w_j)$ and $p(w_k)$ are the probabilities of occurrence of words w_j and w_k in current document collection, respectively. Typically, an average coherence score over all topics can be used to measure the overall quality.

Although such metrics tend to be reasonable for long document scenarios, they can be problematic for short texts. Given the limited word co-occurrences in short texts, even if we obtain desirable topics in certain way, they may not really be given very high coherence scores by the above metrics. To demonstrate this, we view the topics extracted from the NIPS long documents with LDA as gold-standard, and examine if they can receive high coherence scores on the NIPS short texts using the metric described in Equation (4). Meanwhile, we also extract the same numbers (50 to 300) of topics from the short texts directly and calculate their coherence scores on the short texts for comparison. For this demonstration, the parameter of K is set to 20. If the existing metric is effective for our problem, the gold-standard topics should always receive higher coherence scores than those topics directly extracted from short texts. However, as the result shown in Table 1, the effectiveness cannot be supported.

T	50	100	150	200	250	300
Coh_l	0.76	1.03	1.24	1.49	1.71	2.07
Coh_s	1.05	1.67	1.87	1.78	1.63	1.48

Table 1: Coherence scores of gold-standard topics (Coh_l) and comparative topics (Coh_s) estimated on NIPS short texts.

To provide alternative metrics, we propose that the objective of topic modeling for short texts is to distill topics that are in maximum alignment with those extracted from the corresponding long documents if available. This objective can be measured in at least two ways. One is still to measure topic coherence, but instead of on the short texts, it is performed on the long documents. The other is to use the *purity* metric from clustering evaluation. Here a purity score can be obtained by selecting a set of K most probable words from each topic respectively and comparing the sets of words with those from topics extracted on long documents:

$$Purity = \frac{1}{TK} \sum_i \max_j |\mathcal{T}_{z_i} \cap \mathcal{T}_{g_j}|, \tag{5}$$

where z_i and g_j refer to a specific topic extracted from short texts and long documents, respectively, and $\mathcal{T}_{z_i}$ and $\mathcal{T}_{g_j}$ are the sets of K most probable words from topic z_i and g_j.

Parameters of the topic models to be studied below are set as follows. First, the number of iterations for Gibbs sampling is set to 3000, which is generally sufficient enough for convergence. Then, following previous work [Griffiths and Steyvers, 2004; Weng *et al.*, 2010], the parameters of α and β are set to $50/T$ and 0.1, respectively.

4.3 Number of Pseudo-Documents

One would naturally expect that the ideal number of pseudo-documents is the same as or very close to the actual number if known. However, this can be unrealistic in practice as the fragments of a document are not simply assembled in the same way as the aggregation of short texts. We study this on the NIPS short texts and examine how the number of pseudo-documents affects topic quality using the two new evaluation metrics. Meanwhile, the numbers of latent topics, T, from

[1] https://answers.yahoo.com/

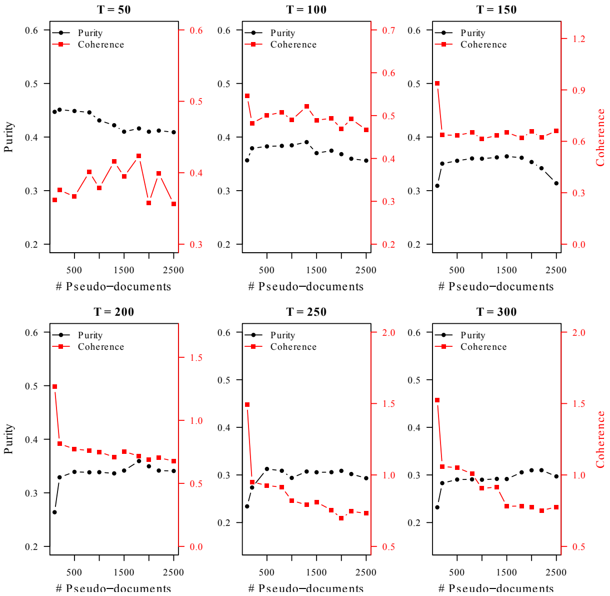


Figure 2: Impact of the number of pseudo-documents.

50 to 300 are investigated. The parameter, K, is fixed at 20 throughout this paper unless mentioned otherwise. As the result shown in Figure 2, it is not difficult to find that the best number of pseudo-documents is likely to be dependent on the number of latent topics. Another observation is that the purity metric tends to be more stable than the topic coherence metric. For this reason, we will only report the purity scores in the rest evaluations.

4.4 Effect of Clustering Algorithms

Despite that our model involves self-aggregation of short texts during topic modeling, other approaches may achieve similar effect. For example, short texts can be aggregated into longer pseudo-documents using any clustering algorithms, and then standard topic modeling can be applied. We study this by using two clustering algorithms. First, the K-means clustering algorithm is performed to aggregate short texts into pseudo-documents and LDA is applied for topics extraction. To do this, each short text snippet is represented using a *tf-idf* vector and the cosine similarity is used to measure the similarity of two short snippets based on their *tf-idf* vectors. Second, since our model resorts to the mixture of unigrams model for short text aggregation, we use it as another clustering algorithm. The pseudo-document of a short text snippet is determined as the one which has the highest probability with it. Since the two clustering algorithms might both lead to different clustering results given different starting points, we run each of them 10 times and report the average performance. As shown in Figure 3, the numbers of pseudo-documents of 500, 1000 and 1500 are studied. From the figures, we can observe that while K-means clustering performs the worst and the mixture of unigrams model performs very competitively for short text topic modeling, our model constantly performs the best, showing the effectiveness of the integration of topic modeling with self-aggregation of short texts. Another observation is that as the number of topics increases, the quality of topics keeps falling, which is likely due to that as more topics are extracted, each of them becomes less coherent.

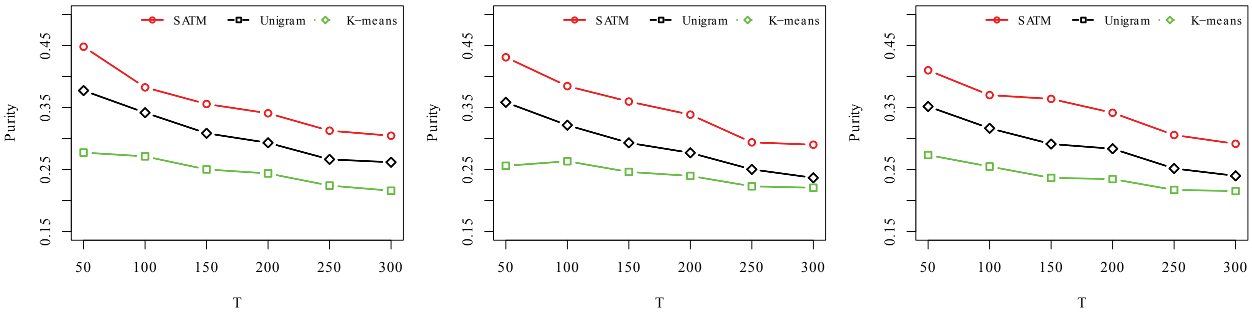
4.5 Comparison with Other Topic Models

Three baseline topic models are implemented on the NIPS short texts for comparison. The first one is the standard LDA, which works well on long documents but not on short and sparse texts [Hong and Davison, 2010; Zhao *et al.*, 2011]. The second model for comparison is the mixture of unigrams model, initially proposed for semi-supervised text categorization [Nigam *et al.*, 2000]. Its parameters are learned using the expectation-maximization (EM) algorithm. The last baseline model is Biterm, which appears to be the first attempt towards topic modeling of general short texts [Yan *et al.*, 2013]. The number of pseudo-documents of SATM is fixed at 1000. The result in terms of the purity metric is depicted in Figure 4, in which different values of K from 10 to 30 are studied. From the figures we can notice that LDA and Biterm are not able to extract as desirable topics from short texts as the other two models. The performance of the mixture of unigrams model is relatively competitive, which is consistent with the finding in [Zhao *et al.*, 2011]. As expected, the new model SATM achieves the best performance in this experiment.

4.6 Short Text Classification

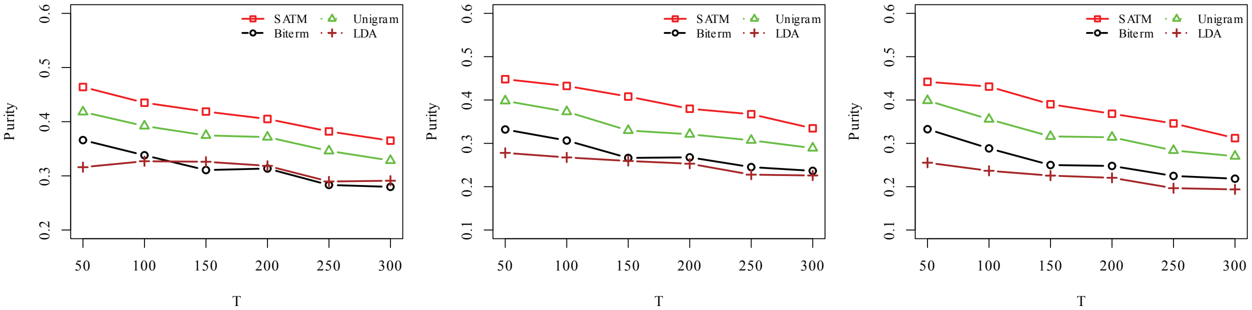
Another reasonable way to evaluate a topic model is to apply the learned topics to an external task. The quality of the topics can be assessed by their performance on the task [Blei *et al.*, 2003]. We conduct a short text classification task for such a purpose on the Yahoo! Answers corpus by randomly dividing it into two equal-sized subsets for training and testing. Similar to [Blei *et al.*, 2003], the parameters of a topic model are firstly estimated on the whole corpus without reference to their true class labels. In this way, a low-dimensional representation can be obtained for each question based on its distribution over topics and a support vector machines (SVM) classifier is then trained on the training set and evaluated on the testing set. For the implementation of SVM, the LIBSVM library [Chang and Lin, 2011] is employed, with its parameters chosen by five-fold cross-validation on the training set. Accuracy is used to measure the performance of this classification task. As shown in Figure 5, while the mixture of unigrams model and LDA have achieved comparable performance, our model constantly performs the best in this task.

One might have noted that the Biterm model is not studied in the above classification task. This is because Biterm cannot explicitly give low-dimensional representations of documents in its modeling process but has to resort to certain post inference strategies [Yan *et al.*, 2013], making it not directly comparable with other models. However, the post inference applies to other topic models as well, yet the classification performance in this case would relate not only to the specific topic model but also to the post inference used. Here we take LDA and SATM as an example and adopt the following post inference strategy to represent each question s: $p(z|s) = \sum_w p(z|w)p(w|s)$, where $p(z|w)$ is estimated by $\frac{p(z)p(w|z)}{\sum_z p(z)p(w|z)}$ and $p(w|s)$ is estimated using the relative frequency of w in s. The result is plotted in Figure 6, from



(a) # pseudo-documents = 500 (b) # pseudo-documents = 1000 (c) # pseudo-documents = 1500

Figure 3: Effect of different clustering algorithms on short text topic modeling.



(a) K = 10 (b) K = 20 (c) K = 30

Figure 4: Performance comparison with baseline topic models on NIPS short texts.

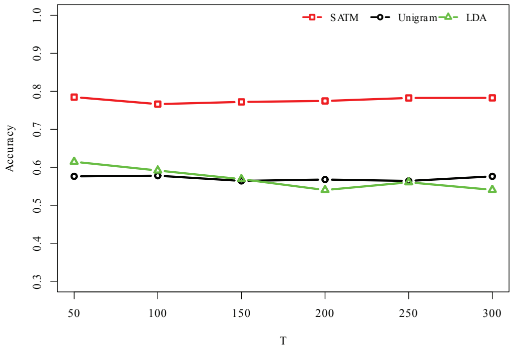


Figure 5: Classification result with direct representations.

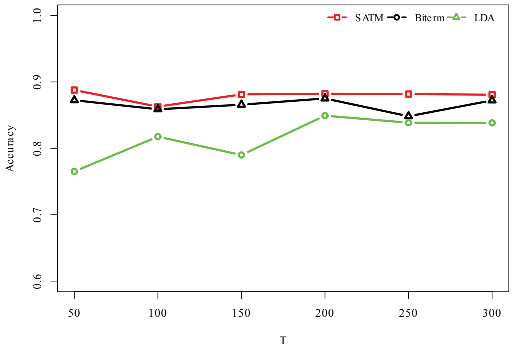


Figure 6: Classification result with indirect representations.

which we can find that the classification performance of LDA and SATM can both be largely improved by using the new representations. This suggests that in addition to topic models themselves, the way how short texts are represented also plays a vital role in the classification. Therefore, even though a model like Biterm can achieve very competitive performance in this classification task, it does not necessarily mean that it can extract more meaningful topics than other models.

5 Conclusions

In this paper, we have presented a generalized solution for topic modeling of very short and sparse texts. Compared with existing solutions which rely seriously on limited contextual information for aggregation of short texts to alleviate the sparsity problem, our model involves an automatic text aggregation during topic modeling, which is founded on more general topical affinity of short texts and can be applied to various forms of short texts. We empirically evaluated this new model on real-world short texts from different domains using two new evaluation metrics. Experimental results confirm the effectiveness of this model, indicating that it is able to extract more meaningful and interpretable topics from short texts than the conventional topic models. Among the existing topic models, the mixture of unigrams model has shown very competitive performance, suggesting that its assumption can be more suited for short texts than for long documents.

Acknowledgments

This research is partially supported by National Institutes of Health under Grant 1R21AA023975-01 and National Center for International Joint Research on E-Business Information Processing under Grant 2013B01035. Sinno J. Pan is supported by the NTU Singapore Nanyang Assistant Professorship (NAP) grant M4081532.020.

References

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

[Blei, 2012] David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

[Chang and Lin, 2011] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):27, 2011.

[Chang *et al.*, 2009] Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-graber, and David M Blei. Reading tea leaves: How humans interpret topic models. In *Advances in Neural Information Processing Systems*, pages 288–296, 2009.

[Griffiths and Steyvers, 2004] Thomas L Griffiths and Mark Steyvers. Finding scientific topics. *Proceedings of the National academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

[Guo *et al.*, 2013] Weiwei Guo, Hao Li, Heng Ji, and Mona T Diab. Linking tweets to news: A framework to enrich short text data in social media. In *Proceedings of the 51th Annual Meeting of the Association for Computational Linguistics*, pages 239–249, 2013.

[Hong and Davison, 2010] Liangjie Hong and Brian D Davison. Empirical study of topic modeling in twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pages 80–88. ACM, 2010.

[Jin *et al.*, 2011] Ou Jin, Nathan N Liu, Kai Zhao, Yong Yu, and Qiang Yang. Transferring topical knowledge from auxiliary long texts for short text clustering. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 775–784. ACM, 2011.

[Mehrotra *et al.*, 2013] Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. Improving lda topic models for microblogs via tweet pooling and automatic labeling. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development*, pages 889–892. ACM, 2013.

[Mimno *et al.*, 2011] David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. Optimizing semantic coherence in topic models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 262–272. Association for Computational Linguistics, 2011.

[Newman *et al.*, 2010] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108. Association for Computational Linguistics, 2010.

[Nigam *et al.*, 2000] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using em. *Machine learning*, 39(2-3):103–134, 2000.

[Rosen-Zvi *et al.*, 2004] Michal Rosen-Zvi, Thomas Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence*, pages 487–494. AUAI Press, 2004.

[Tang *et al.*, 2013] Jian Tang, Ming Zhang, and Qiaozhu Mei. One theme in all views: modeling consensus topics in multiple contexts. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 5–13. ACM, 2013.

[Voorhees, 1994] Ellen M Voorhees. Query expansion using lexical-semantic relations. In *SIGIR94*, pages 61–69. Springer, 1994.

[Wang and McCallum, 2006] Xuerui Wang and Andrew McCallum. Topics over time: a non-markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 424–433. ACM, 2006.

[Weng *et al.*, 2010] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the Third ACM International Conference on Web Search and Data Mining*, pages 261–270. ACM, 2010.

[Xu and Croft, 1996] Jinxi Xu and W Bruce Croft. Query expansion using local and global document analysis. In *Proceedings of the 19th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 4–11. ACM, 1996.

[Yan *et al.*, 2013] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. In *Proceedings of the 22nd International Conference on World Wide Web*, pages 1445–1456. International World Wide Web Conferences Steering Committee, 2013.

[Zhao *et al.*, 2011] Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. Comparing twitter and traditional media using topic models. In *Advances in Information Retrieval*, pages 338–349. Springer, 2011.