World Wide Web (2020) 23:3001–3023
https://doi.org/10.1007/s11280-020-00814-x

A Pseudo-document-based Topical N-grams model for short texts

Hao Lin[1] · Yuan Zuo[1] · Guannan Liu[1] · Hong Li[1] · Junjie Wu[1,2,3] · Zhiang Wu[4]

Received: 6 July 2019 / Revised: 19 January 2020 / Accepted: 30 March 2020 /
Published online: 23 July 2020

Abstract
In recent years, short text topic modeling has drawn considerable attentions from interdisciplinary researchers. Various customized topic models have been proposed to tackle the semantic sparseness nature of short texts. Most (if not all) of them follow the *bag-of-words* assumption, which, however, is not adequate since word order and phrases are often critical to capturing the meaning of texts. On the other hand, while some existing topic models are sensitive to word order, they do not perform well on short texts due to the severe data sparseness. To address these issues, we propose the Pseudo-document-based Topical N-Grams model (PTNG), which alleviates the data sparsity problem of short texts while is sensitive to word order. Extensive experiments on three real-world data sets with state-of-the-art baselines demonstrate the high quality of topics learned by PTNG according to UCI coherence scores and more discriminative semantic representation of short texts according to classification results.

Keywords Short text · Topic model · Word order · Topical N-Grams

1 Introduction

Short text is being the prevalent format of information on the Internet, due to the explosive growth of online social media like Twitter and Facebook. Almost 500 million tweets daily on Twitter, for example, can be produced by around 250 million active users. This massive short texts carry sophisticated information which can hardly be found in conventional sources of information [34]. The accurate knowledge discovery of short texts has therefore been recognized as a challenging yet promising research problem.

The archetypal topic model, i.e., Latent Dirichlet Allocation (LDA) [1], performs relatively poor when directly applied to short texts for the lack of word co-occurrence information [24] compared to normal size documents. Therefore, many research efforts have

✉ Yuan Zuo
zuoyuan@buaa.edu.cn

Extended author information available on the last page of the article.

been devoted to tackle incompetence of LDA in modeling short texts. Several customized topic models [5, 10, 21, 31, 32, 35, 36] have been proposed to alleviate the data sparsity issue of short texts. One potential limitation of the above models is that they all follow the *bag-of-words* assumption, which brings in computational efficiency but might severely hurt the accuracy of topic modeling for the ignorance of word order. We list two detailed reasons as follows:

- Sentences have the same *bag-of-words* representation could have quite different meanings. For instance, "the department chair couches offer" and "the chair department offers couche" are about quite different topics while have the same unigram statistics [25]. Different from normal size documents, many short texts contain single sentence, which makes the ordering of words become much more essential in short texts than in regular texts.
- A collocation (phrase)'s meaning typically is not derivable from its constituted words. For example, the meaning of *power supply* as an electronic device comes from neither the unigram *power* nor the unigram *supply*, but emerges from the bigram as a whole. Since LDA is prone to decompose collocations into different topics [11], then *power* might be generated from an electronic topic, while *supply* be generated from an economy topic.

As discussed above, under the *bag-of-words* assumption, two short texts might have the same representation while their topics are different, and unigrams of given n-gram might be assigned with different topics while they actually emerge from a collocation as a whole. Both situations prevent topic models from learning accurate topics from short texts. Thus, sensitivity of word order is essential for topic modelling of short texts. Collocation (phrase or short sequence of adjacent words) is most widely used in bringing word order information into topic models [4, 11, 25–27]. Nevertheless, it is worth noting that topic models utilize word co-occurrence information to reveal topics. The above models, however, might result in reduced co-occurrence information for concatenating successive words into collocations. Therefore, these model are not directly applicable for short texts due to the severe data sparsity problem. On one hand, short text topic modeling needs to consider the ordering of words. On the other hand, exiting models sensitive to word orders are not suitable for short texts. The above consideration motivates us to design a new model, which generates short texts in original sequence, while can alleviate the sparsity of word co-occurrences on short texts.

In this paper, we propose a novel topic model called Pseudo-document-based Topical N-Gram model (PTNG) for short texts. On one hand, PTNG can leverage much less pseudo documents to self aggregate tremendous short texts, which helps it to gain advantages in learning topic distributions from short texts. Besides, PTNG can also automatically determines unigram words and collocations based on context and assign topics to both individual words and collocations, which guarantees the accuracy of learned topics. To our best knowledge, this work is among the earliest studies in this interesting direction. Extensive experiments on three real-world data sets with classic as well as state-of-the-art baselines demonstrate the high quality of topics and more discriminative semantic representations of short texts learned by PTNG, in terms of both topic coherence and classification results.

The remainder of this paper is organized as follows. Section 2 briefly introduces related work. Section 3 describes the details of the PTNG model and its inference way. The experiment results of PTNG can be found in Sec. 4, and we conclude our work in Sec. 5.

2 Related work

2.1 Collocation-based topic analysis

Most topic analysis methods (e.g., probabilistic topic models such as Latent Dirichlet Allocation [1]) rely on *bag-of-words* assumption that words are generated independently, and so ignore potential useful information about word order. To bring word order information into consideration, *collocations* (phrases or short sequences of adjacent words) are most widely used, forming a family of methods for collocation-based topic analysis. In the literature, existing methods on collocation-based topic analysis can be roughly divided into two categories, including topical phrase mining methods and n-gram topic models. Topical phrase mining methods usually adopt a two-phase strategy to mine phrases prior to performing topic modeling [2, 6, 12–14, 30] whereas n-gram topic models make extensions to LDA [1], which directly modifies the generative process for simultaneously discovering n-grams and learning topics. For topical phrase mining methods, phrase mining is first conducted in a pre-processing phase, which either requires complex multi-step statistical framework [2, 12, 13, 30] or is domain/language dependent [6, 14]. Therefore, we focus on n-gram topic models.

Wallach et al. [25] propose the bigram topic model which assumes topic is a distribution of bigrams instead of unigrams, where a word is generated condition on previous word and its topic. One drawback of this model is that it takes every successive two words as bigrams, while unigrams are the major components in a document. Above limitation is addressed in the LDA collocation model (LDACOL) [4] which introduces a set of status variables called the bigram status variable to indicate whether two consecutive words form a bigram or not. Topical n-gram model (TNG) [27] extended LDACOL that the second word of a collocation not only depends on the first word but also depends on semantic contextual information, i.e., topics. One potential drawback of TNG is that words within a topical n-gram are not constrained to share the same topic, which is address in Phrase-Discovering LDA (PDLDA) [11].

2.2 Short text topic models

Short text topic modeling has long suffered from data sparsity. An intuitive way of alleviating data sparseness issue is to make use of auxiliary information when available. For instance, contextual information such as authorship, URL, hashtag, time and location can be utilized as supplementary to textual content in tweets for topic modeling. Research along this line can roughly be divided into two categories. One sort proposes to use auxiliary information for aggregating short texts directly [7, 16, 28], and the other attempts to construct specific models with those information included in the short texts' generative process [9, 15, 18, 22].

Recent study has focused more on the design of customized short text topic models, due to the unavailability of auxiliary information and the too costly deployment in practice. To the best of our knowledge, the biterm topic model (BTM) [31] is one of the pioneer work along this line, which directly models word pairs (i.e. biterms) extracted from short texts. The biterm topic model switches from sparse document-word space to dense word-word space, by which the learned topics are more coherent than topics learnt in LDA. Zuo et al. [35] propose the word network topic model for topic learning from word co-occurrence networks and this model produces more coherent topics than the biterm topic model. These

models, however, are occasionally criticized for lack of directly representing topics of documents. Lin et al. [10] propose to replace LDA's symmetric Dirichlet priors with *Spike and Slab* priors [8, 29], which can produce more coherent topics and better topical representation for documents. The issue with the above models is that they incorporate little additional word co-occurrence information and thus still confront data sparsity problems. Quan et al. [21] introduce a self-aggregated topic model called SATM, capable of aggregating short texts into latent pseudo documents. This aggregation is based on the short texts' own topics rather than auxiliary information. Zuo et al. [36] introduce the pseudo-document-based topic model (PTM), which also utilizes aggregated pseudo documents for topic learning, but is less likely to overfit and performs learning more efficiently compared to SATM.

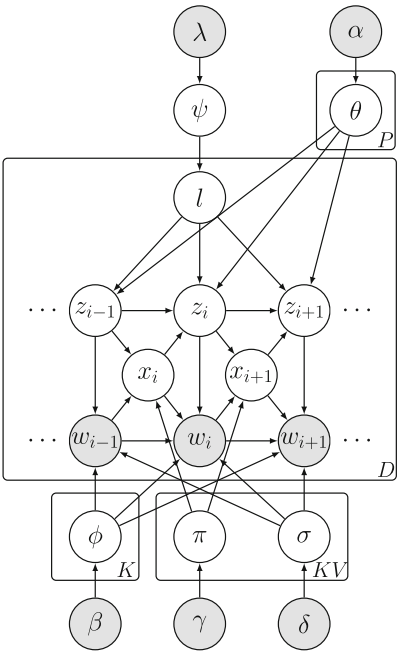
Remark. As we stated in previous section, word order information is essential for accurately revealing topics from short texts. However, existing short text topic models make the bag-of-words assumption that words are generated independently, and so ignore potentially useful information about word order. Topic models that aware of word order are not suitable for short texts due to inability to avoid data sparseness. Our model is the first attempt to make short text topic model beyonds the *bag-of-words* assumption.

3 Model and inference

In this section, we describe the Pseudo-document-based Topical N-Gram model (PTNG) for short text documents. PTNG assumes the observed massive short texts are generated from latent documents, called *pseudo documents*, which are much less yet normal-sized. Moreover, PTNG can automatically detects collocations based on context and assigns topics to them.

3.1 The PTNG model

Now we formally describe the technical details of PTNG. There are D observed short text documents $\{d_s\}_{s=1}^{D}$ and P latent pseudo documents $\{d'_l\}_{l=1}^{P}$. We introduce a multinomial distribution ψ to model the short texts' distribution over pseudo documents. Each short text document is assumed to belong to *one and only one* pseudo document. Furthermore, it is assumed there are K topics $\{\phi_z\}_{z=1}^{K}$, of which each is a multinomial distribution over a vocabulary set of size V. In order to automatically determine unigrams and collocations, we assume there are $K \times V$ Bernoulli distributions $\pi_{z,w'}$ and $K \times V$ multinomial distributions $\sigma_{z,w'}$ over the whole vocabulary. The word w (except for the beginning word) in the s-th short text d_s, is generated by first sample a bigram status x from $\pi_{z',w'}$, where z' is the topic assigned to w' and w' indicates the word occurs previous to w in d_s. As to the beginning word, we fix its bigram status x equals to zero. If the bigram status x of the word w equals to zero, we sample its topic z from θ_l and then generate the word w from ϕ_z, where l indicates the pseudo document d'_l that short text d_s belongs to. If the bigram status x of the word w equals to one, then its topic z is constrained be the same as the topic of w' and the word w is generated from $\sigma_{z,w'}$. In other words, x equals to zero indicates w is generate from multinomial distribution over unigrams, and x equals to one indicates w and w' forms a bigram. Note that continuous bigrams form n-gram, therefore PTNG can detect collocations or phrases.



(a) Plate notation of PTNG.

Notation	Description
K	number of topics
P	number of pseudo documents
D	number of short texts
V	size of vocabulary
l	pseudo document assignment for a short text
z	topic assignment for a word
x	bigram status between two successive words
w	an observed word
ϕ_z	the multinomial distribution of words w.r.t topic z
θ_l	the multinomial distribution of topics w.r.t pseudo document d'_l
ψ	the multinomial distribution of pseudo documents w.r.t given short texts
$\pi_{z,w}$	the Bernoulli distribution of status variables w.r.t. previous topic z and word w
$\sigma_{z,w}$	the multinomial distribution of words w.r.t. previous topic z and word w
λ	Dirichlet prior parameter for ψ
α	Dirichlet prior parameter for θ
β	Dirichlet prior parameter for ϕ
γ	Beta prior parameters for π
δ	Dirichlet prior parameters for σ
$Bern(\cdot)$	Bernoulli distribution with parameter$(\cdot)$
$Beta(\cdot)$	Beta distribution with parameter$(\cdot)$
$Multi(\cdot)$	Multinomial distribution with parameter$(\cdot)$
$Dir(\cdot)$	Dirichlet distribution with parameter$(\cdot)$

(b) Math notations.

Figure 1 Plate notation and math notations of PTNG

The plate notation of PTNG is illustrated in Figure 1-a, with a summary of math notations provided in Figure 1-b. The generative process is thus illustrated as follows:

1. Draw $\psi \sim Dir(\lambda)$
2. For each pseudo document d'_l:
   (a) Draw $\theta_l \sim Dir(\alpha)$
3. For each topic z:
   (a) Draw $\phi_z \sim Dir(\beta)$
4. For each topic z and each word w:
   (a) Draw $\pi_{z,w} \sim Beta(\gamma)$
   (b) Draw $\sigma_{z,w} \sim Dir(\delta)$
5. For each short text d_s:
   (a) Draw a pseudo document $l \sim Multi(\psi)$
   (b) For each word w_i in d_s:
      i. Draw $x_i \sim Bern(\pi_{z_{i-1},w_{i-1}})$
      ii. If $x_i = 0$, draw $z_i \sim Multi(\theta_l)$;
      If $x_i = 1$, $z_i = z_{i-1}$

iii. If $x_i = 0$, draw $w_i \sim Multi(\phi_{z_i})$;
If $x_i = 1$, draw $w_i \sim Multi(\sigma_{z_{i-1}})$

Remark 1 Although the mechanism of topical collocation detection used in PTNG is similar with the one of Topical N-Gram model (TNG) [27], there is a difference which significantly influences model's performance. TNG does not constrain the consistency of topics assigned to words within a collocation (i.e., no arrows $z_{i-1} \rightarrow z_i$ and $x_i \rightarrow z_i$ in Figure 1-b), while PTNG assumes one collocation has only one topic. In our practice, removing arrows $z_{i-1} \rightarrow z_i$ and $x_i \rightarrow z_i$ degrades performance of PTNG significantly.

Remark 2 As we have discussed above, topic model that detects collocations faces much severer data sparsity on short texts, since concatenating successive words into collocation results in fewer co-occurrence information. Thus, before designing a short text topic model that is aware of word order, one needs to address the data sparsity issue first. Short text topic models that directly model word co-occurrences [31, 35] are not suitable because they reorganize short texts into biterms or a word co-occurrence network. Therefore, word order is already lost after the reorganization. Although topic model with sparse priors [10] is capable of utilizing word order during generative process, its performance is relatively poor in practice. Pseudo document based approach is suitable for its good performance and awareness of original word order in short text.

3.2 Inference

Due to the intractable issue of exact posterior inference in our model, we resort to Gibbs sampling algorithm [3] for approximate inference of PTNG. In order to reduce the uncertainty introduced by $\phi, \theta, \psi, \sigma$ and π, we integrate them out with no trouble because of the conjugate prior setting. The latent variables which the sampling algorithm needs to update are pseudo document assignment l, topic assignment z and bigram status x.

3.2.1 Sampling pseudo document assignments *l*.

We give the sampling equation for pseudo document assignments l as follows,

$$\begin{aligned} p(l_{d_s} = l | rest) &\propto \frac{M_{l,\neg d_s}}{D - 1 + P\lambda} \frac{\frac{\prod_{z \in d_s} \Gamma(N_l^z + \alpha)}{\prod_{z \in d_s} \Gamma\left(N_{l,\neg d_s}^z + \alpha\right)}}{\prod_{i=1}^{N_{d_s}} \left(N_{l,\neg d_s} + K\alpha + i - 1\right)} \\ &= \frac{M_{l,\neg d_s}}{D - 1 + P\lambda} \frac{\prod_{z \in d_s} \prod_{j=1}^{N_{d_s}^z} \left(N_{l,\neg d_s}^z + \alpha + j - 1\right)}{\prod_{i=1}^{N_{d_s}} \left(N_{l,\neg d_s} + K\alpha + i - 1\right)}. \end{aligned} \tag{1}$$

The sampling approach is similar to that used in the Dirichlet multinomial mixtures model [33]. Note that M_l is the number of short texts assigned to the l-th pseudo document d'_l. N_{d_s} is the number of tokens in the s-th short text d_s, and $N_{d_s}^z$ is the number of tokens in d_s that are assigned to topic z. Similarly, N_l is the total number of tokens in the l-th pseudo document d'_l, and N_l^z is the number of tokens in d'_l that are assigned to topic z. All counts with $\neg d_s$ mean excluding counting from d_s.

3.2.2 Sampling topic assignments *z* and bigram status *x*.

Instead of sampling z and x separately, we jointly sample them as a block. That is,

$$p(z_{d_s,i} = z_i, x_{d_s,i} = x_i | rest) \propto (\gamma + N^{x_i}_{z_{i-1},w_{i-1}} - 1) \times \begin{cases} (\alpha + N^{z_i}_{l_{d_s}} - 1) \frac{N^{w_i}_{z_i} + \beta - 1}{N_{z_i} + V\beta - 1} & \text{if } x_i = 0 \\ \frac{N^{w_i}_{z_i,w_{i-1}} + \delta - 1}{N_{z_i,w_{i-1}} + V\delta - 1} & \text{if } x_i = 1 \& z_i = z_{i-1} \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

where N^w_z is the number of times w being assigned to topic z and $N^w_{z,w'}$ is the number of times w forming an bigram with previous word w' under topic z. $N_z = \sum_{w=0}^{V} N^w_z$ and $N_{z,w'} = \sum_{w=0}^{V} N^w_{z,w'}$. $N^{x_i}_{z_{i-1},w_{i-1}}$ is the number of times w_{i-1} forming an bigram (or not) with any word under topic z when $x_i = 1$ (or $x_i = 0$).

With assignments of l, z and x, we can easily obtain posterior estimates for θ, ϕ, ψ, π and σ. Taken θ and ϕ as examples:

$$\theta_{l,z} = \frac{N^z_l + \alpha}{N_l + K\alpha}, \tag{3}$$

$$\phi_{z,w} = \frac{N^w_z + \beta}{N_z + V\beta}. \tag{4}$$

Note that since PTNG learns topic proportions θ_l for each pseudo document d'_l, we use empirical estimation to obtain θ_s for short text d_s with any single sample of $\vec{z}$:

$$\theta_{s,z} = \frac{N^z_{l_{d_s}} + \alpha}{N_{l_{d_s}} + K\alpha}. \tag{5}$$

For some applications, topic models are sensitive to hyperparameters. In the particular experiments discussed in this paper, we find the sensitivity to hyperparameters is not a big concern. Thus, we skip the inference of hyperparameters, and use empirical values for them.

3.2.3 Complexity analysis

Recall that, when bigram status x_i is on, the generation of word w_i in PTNG is defined by a conditional distribution $p(w_i|z_{i-1}, w_{i-1})$ described by $K \times V \times V$ parameters denoted as $\sigma_{z_{i-1},w_{i-1}}$. One might think $\sigma_{z,w}$ is too large to fit into memory. However, since a word only co-occurs near with a fraction of words in real-life short texts, $\sigma_{z,w}$ is extreme sparse and can be store in HashMap very efficiently. The effective parameters in $\sigma_{z,w}$ is also far less than $K \times V \times V$, therefore, model complexity of PTNG is only slightly increased after the introducing of bigram status x.

Moreover, in Table 1 and Table 2, we have shown time cost and memory cost of our proposed model and several competitive baselines on News collection, respectively. It can be

Table 1 Average time cost (seconds) of each iteration on News

K	50	100	150	200	250
LDA	0.27	0.52	0.78	1.03	1.31
TNG	0.64	1.16	1.59	1.97	2.47
DSTM	3.10	6.71	9.99	13.60	17.80
PTM	6.51	7.12	7.94	8.82	10.16
PTNG	3.05	5.00	7.12	9.13	11.02

Table 2 Maximum memory cost (m) of each iteration on News

K	50	100	150	200	250
LDA	70	81	157	202	246
TNG	426	934	1210	1448	1679
DSTM	472	935	1234	1488	1711
PTM	108	181	258	388	471
PTNG	437	929	1230	1451	1703

seen that in terms of time cost, PTNG is always about 3 times of the topical n-grams based method TNG with different number of K while as for memory cost, PTNG is comparable to TNG. Besides, as for running time, PTNG is comparable to the pseudo-document-based method PTM. Moreover, PTNG is computationally less complex than DSTM in terms of both time cost and memory cost. Especially, DSTM's time cost grows rapidly with increasing number of K. In summary, PTNG is scalable and efficient in handling real-world short text collections.

4 Experiments

In this section, we first give brief descriptions of datasets used in this paper, then introduce baseline methods and evaluation metrics for comparison. Finally, extensive experimental results are presented to show the effectiveness and robustness of PTNG.

4.1 Datasets description

We evaluate our model on three real-world short text datasets, of which the statistics are given in Table 3. The descriptions of these datasets are as follows.

4.1.1 News

This collection[1] is obtained from RSS feeds of three popular newspaper websites, which contains 29,200 English news articles and covers seven categories inlcuding business, sport, U.S., health, world, sci&tech and entertainment. The description of each news is retained as typical short text.

4.1.2 DBLP

This dataset contains 55,290 titles of conference papers extracted from six research fields including computer vision, database, data mining, machine learning, information retrieval and natural language processing. Each short text, *i.e.* the conference title, is labeled with one of the six research fields.

[1] http://acube.di.unipi.it/tmn-dataset/

Table 3 Statistics of data sets

Data set	# Documents	Vocabulary size	Avg. document length
News	29,200	11,007	12.4
DBLP	55,290	7,525	6.4
Tweets	182,671	21,480	8.5

4.1.3 Tweets

Zubiaga et al. [37] crawl a large set of tweets and each tweet contains URL and labeled category of its corresponding web page. The Open Directory Project (ODP) defines the category of each web page. In this paper, we pick nine topic-related categories, under which 182,671 tweets are sampled for experiments.

4.2 Baseline methods

4.2.1 Mixture of Unigrams (MU)

In Mixture of Unigrams (MU) [19], each document is assumed to be generated by one and only one topic, which may sound reasonable for certain short text collections.

4.2.2 Latent Dirichlet Allocation (LDA)

LDA [1], one of the most commonly used probabilistic topic models, may induce sparsity when its Dirichlet prior approximates to zero. The java implementation of LDA with collapsed Gibbs sampling, *i.e.*, jGibbLDA[2], is included for comparison.

4.2.3 Biterm Topic Model (BTM)

BTM [31] directly models the generative process of word pairs (*i.e.*, biterms) in short text collections and is a competitive baseline for short text topic modeling.

4.2.4 Dual Sparse Topic Model (DSTM)

As a sparsity-enhanced topic model, DSTM utilizes Spike and Slab prior for learning documents' focused topics and topics' focused terms. It seems reasonable for DSTM to assume each document in short text collections to be generated by a subset of topics and each topic to be a distribution over a subset of vocabulary.

4.2.5 Topical N-gram model (TNG)

TNG [27] used in our experiments is slightly different from the original one, since we force unigrams in a phrase to share the same topic. This modification helps TNG performs much better in our experiments.

[2]http://jgibblda.sourceforge.net

4.2.6 Pseudo-document-based Topic Model (PTM)

PTM is proposed by [36], which aggregates short texts into pseudo documents without auxiliary information and achieves the state-of-the-art performance on short text topic modeling.

Parameter Settings for Comparison . For MU and LDA we set $\alpha = 0.1$ and $\beta = 0.01$, because they perform better with weak priors in short text collections. For DSTM, we find the setting $\pi = 0.1$, $\gamma = 0.01$ achieves better performance than the setting $\pi = 1.0$, $\gamma = 1.0$ in its original paper. As for $\bar{\pi}$ and $\bar{\gamma}$, we follow the suggestions of the authors. For PTM, PTNG and TNG, we set $\alpha = 0.1$ and $\beta = 0.01$. For PTNG and PTM, we set $\lambda = 0.1$. For PTNG and TNG, we set $\delta = 0.01$. For Beta prior parameters of PTNG and TNG, *i.e*, γ, we set $\gamma_0 = 0.1$ and $\gamma_1 = 0.01$, where γ_0 and γ_1 are the shape parameters of Beta priors. For all models, the number of topics is set to 100. For PTM and PTNG, we set the number of pseudo documents to 1000. As to sampling iterations, we set it to 2000 for all methods except for PTNG and TNG, whereas for PTNG and TNG we set it to 5000 to ensure their convergence. We report all results in this section evaluated over five runs. Detailed discussions about parameter sensitivity of our proposed model will be given in Section 4.8.

4.3 Evaluation measures

4.3.1 Topic coherence

It is still an open issue for topic model evaluation. Researchers has proven that as for the frequently used metric called *perplexity*, a better perplexity does not necessarily lead to more understandable topics, due to its less correlation to human interpretability. Moreover, it is not a general way for many short text topic models (*e.g.*, PTM and our proposed PTNG model) to evaluate topic learning, since topics can not be revealed directly from short text collections for these models. As such, a more correlated metric named *topic coherence* has been employed and is proven to be more generalized than perplexity. Two types of topic coherence have been proposed, including UMass topic coherence [17] and UCI topic coherence [20]. UMass topic coherence is reported as not a good measure for topic evaluation on short texts while UCI topic coherence needs an external corpus, such as Wikipedia. As such, UCI topic coherence is employed for topic model evaluation on DBLP and News which is well-edited. We do not apply UCI topic coherence on tweets since tweets are less formal text and less appropriate for evaluation.

For each specific topic z, the top-N most probable terms $w_1, w_2, \cdots, w_N$ are used and the average point-wise mutual information (PMI) score for these term pairs are calculated as UCI topic coherence. For consideration of fair comparison, as with all methods, unigrams of each learned topic are examined in calculating topic coherence scores. In our experiments, we set N to 10 and set sliding window size to 30. The UCI topic coherence of topic z is as follows:

$$PMI(z) = \frac{2}{N(N-1)} \sum_{1 \le i \le j \le N} log \frac{p(w_i, w_j)}{p(w_i)p(w_j)}, \tag{6}$$

where $p(w_i, w_j)$ denote the joint probability that term pair w_i and w_j co-occur in the same sliding window, and $p(w_i)$ is the marginal probability that term w_i appears in a sliding window. We estimate these probabilities from the Wikipedia articles' latest dump. After

calculating UCI topic coherence for each topic individually, these topic coherence values are averaged for evaluating the overall performance of topic model.

4.3.2 Classification

Besides evaluation by topic coherence, we also compare the latent topical representations learned by our model and its competitive baseline methods over the task of short text classification. In this classification experiment, we use weighted averaged precision, recall and f-measure as the evaluation metrics.

4.4 Evaluating topic learning with topic coherence

We show the UCI topic coherence results of PTNG and all competitive baseline methods on News and DBLP in Figure 2. Recall that UCI topic coherence is more appropriate than UMass topic coherence for short texts, however, owing to the lack of a suitable reference corpus, we leave out topic coherence comparison on tweets.

From the results as shown in Figure 2, it is observed that PTNG almost outperforms all baselines on both datasets except for the case that BTM performs better than PTNG on DBLP. PTNG's superior performance over PTM demonstrates that it is essential to incorporate word order for accurately revealing topics on short text collections. Moreover, the outperformance of our proposed model over TNG validates that pseudo documents are capable of alleviating the absence of word co-occurrence in short text collections. TNG performs similar to LDA on News and slightly outperforms LDA on DBLP, which indicates solely taking word order into consideration brings little improvements due to data sparsity problem on short texts. It is also interesting to see that LDA with weak prior achieves better topic coherence as compared to DSTM. This might be owing to that DSTM needs to infer large amount of sparse priors, which makes it practically difficult for DSTM to learn a precise model. For MU, poorer performance is achieved on news than on DBLP, which indicates news description might be generated by more than one topic. In contrast, on DBLP, MU produces relatively excellent topic coherence, which could imply that DBLP paper title might involve fewer number of topics than news description.

To look more deeply into the topics learned by BTM and PTNG, following [23], we further employ a new metric for evaluating the diversity of learned topics, which is computed as the average cosine distance of pairwise topics. The results of topic diversity for BTM

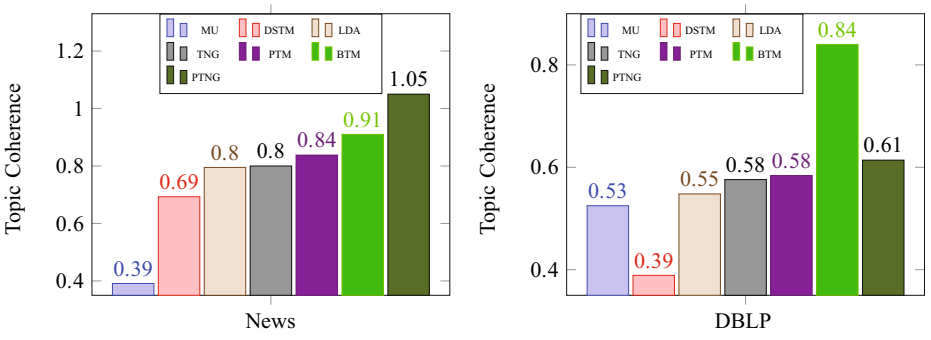


Figure 2 UCI topic coherence on News and DBLP

Table 4 Topic diversity for PTNG and BTM on News and DBLP

The best results are highlighted in bold

	News	DBLP
PTNG	**0.973**	**0.966**
BTM	0.873	0.866

and PTNG are shown in Table 4. We find that on both datasets, topic diversity for PTNG is greater than that for BTM. For example, on DBLP, topic diversity for PTNG is 0.966 while that for BTM is 0.873. This indicates that BTM might be problematic for learning more redundant topics than PTNG and thus shows similar semantics across different topics. As such, the UCI topic coherence of BTM might be overestimated by the high coherence of a few redundant topics.

4.5 Evaluating topic learning via short text classification

Here topic model is used as a utility for dimension reduction, which characterizes each document with a fixed set of topics as input features for document classification. After convergence of each trained topic model, we obtain the latent topical represetation and then conduct five-fold cross-validation on all three datasets. We employ LIBLINEAR[3] for classification.

According to [21], BTM is not suitable for comparison in short text classification experiment. This is because BTM cannot produce direct document representation but need certain post inference strategies, which is thought to give indirect document representation [21]. Details concerning this issue can be referred to the discussion of short text classification experiment conducted in [21]. To empirically validate the effect of different strategies of document representation to the classification performance, we show the results of PTNG and BTM with different document representations in Table 5. Note that, PTNG-d produces document representations with technique described in Section 3.2.2 whereas PTNG-i produces an indirect representation for document d_s as follows:

$$\theta'_{s,z} \overset{chain\ rule}{=} \sum_{i=1}^{I_s} p(z|w_i)p(w_i|d_s) \overset{Bayes'\ formula}{=} \sum_{i=1}^{I_s} \frac{p(z)\phi_{z,w_i}}{\sum_z p(z)\phi_{z,w_i}} p(w_i|d_s),$$

where document d_s has I_s terms, $p(w_i|d_s)$ can be estimated using the relative frequency of w_i in d_s and $p(z)$ can be estimated using the relative frequency of z in all documents. Essentially BTM takes a similar fashion as PTM-i for producing document representations and thus we denote it as BTM-i in Table 5. By comparing PTNG-d with PTNG-i, we can see that the classification performance of PTNG can be dramatically improved by using the indirect representation (*i.e.*, PTNG-i). This indicates the competitive classification performance of BTM (*i.e.*, BTM-i) does not necessarily mean that BTM can produce more meaningful document representation than other models.

We show the results of PTNG and other baselines with weighted precision, recall and f-measure in Table 6. The best results are highlighted in bold while the second bests are in italic. It can be seen that among all baseline methods PTM performs the best on News, DBLP and Tweets. This shows PTM's excellent performance over other baseline methods for learning semantic representations in short text collections. It is also in line with our motivation that short text topic modeling may greatly benefit from aggregating massive

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

Table 5 Classification results of PTNG and BTM with different document representations in five-fold cross validation

	News			DBLP			Tweets		
	precision	recall	f-measure	precision	recall	f-measure	precision	recall	f-measure
PTNG-d	0.770	0.770	0.769	0.662	0.669	0.664	0.597	0.604	0.596
PTNG-i	**0.788**	**0.791**	**0.789**	**0.697**	**0.704**	**0.699**	0.647	0.657	0.646
BTM-i	0.786	0.789	0.787	0.690	0.697	0.692	**0.649**	**0.658**	**0.649**

The best results are highlighted in bold

Table 6 Classification results of different methods in five-fold cross validation

	News			DBLP			Tweets		
	precision	recall	f-measure	precision	recall	f-measure	precision	recall	f-measure
MU	0.697	0.617	0.626	0.640	0.643	0.638	**0.634**	0.546	0.546
LDA	0.727	0.732	0.728	0.613	0.624	0.614	0.553	0.560	0.546
TNG	0.710	0.715	0.710	0.608	0.619	0.610	0.559	*0.571*	0.558
DSTM	0.720	0.724	0.720	0.619	0.628	0.620	0.539	0.547	0.535
PTM	*0.755*	*0.757*	*0.754*	**0.667**	**0.672**	**0.668**	0.561	0.568	*0.559*
PTNG	**0.770**	**0.770**	**0.769**	*0.662*	*0.669*	*0.664*	*0.597*	**0.604**	**0.596**

The best results are highlighted in bold

Table 7 UCI topic coherence for PTNG and PTNG-U

	News	DBLP
PTNG	1.05	0.61
PTNG-U	0.86	0.58

short texts into much less yet normal size pseudo documents. Our proposed PTNG significantly outperforms PTM on News and Tweets, while performs comparable with PTM on DBLP. This promising result suggests that word order information plays an important role in learning accurate topics, which further helps learning of more discriminative semantic representations of short texts. TNG performs slightly worse than LDA on News and DBLP, which indicates order information alone is not enough for short text topic modeling. The incompetence of MU and DSTM suggests applying sparse constraints to topic models is not reliable on short texts as compared to our approach.

4.6 Ablation experiment

In PTNG, we impose a constraint to assign the topics of words within a collocation to be the same one. To study the contribution of this topic constraint, we design a new model called PTNG-U to eliminate the constraint taken on the topics of consecutive n-grams. The UCI topic coherence results and classification results are shown in Table 7 and Table 8, respectively. It is obvious to see that in terms of both UCI topic coherence and classification metrics, after eliminating the constraint taken on the topics of consecutive n-grams, PTNG's performance on all datasets degrades significantly. This well demonstrates that the topic constraint within a phrase is indeed crucial to the success of short text topic modeling.

4.7 Evaluating topic learning by semantics

We further employ PTNG on DBLP and show four topics learned from a 100-topic run in Table 9, with comparison to the corresponding closest TNG topics. The number of pseudo documents is set to 1,000.

The "Collaborative Filtering" topic learned by PTNG provides an extremely clear summary of the corresponding research area by forming n-gram phrases (such as "matrix factorization", "low rank", *etc.*) that are only high probable in collaborative filtering. In contrast, although TNG can also form some meaningful phrases (*e.g.*, "non negative" and "recommender systems"), more unrelated phrases are introduced by TNG in the top 20 word list, *e.g.*, "non rigid", "non redundant".

In other three topics of Table 9, we can find similar results as well, *i.e.*, PTNG tends to form more semantically coherent phrases than TNG. For example, in "Outlier Detection"

Table 8 Classification results of five-fold cross validation for PTNG and PTNG-U

	News			DBLP			Tweets		
	precision	recall	f-measure	precision	recall	f-measure	precision	recall	f-measure
PTNG	0.770	0.770	0.769	0.662	0.669	0.664	0.597	0.604	0.596
PTNG-U	0.762	0.763	0.762	0.651	0.656	0.652	0.584	0.591	0.583

Table 9 The four topics learned by PTNG and TNG on DBLP

TNG		PTNG	
n-gram(1)	n-gram(2+)	n-gram(1)	n-gram(2+)
Collaborative Filtering			
matrix	non negative	matrix	matrix factorization
filtering	recommender systems	sparse	low rank
collaborative	cold start	clustering	high dimensional
factorization	non parametric	robust	matrix completion
recommendation	non rigid	low	missing data
based	non linear	subspace	non negative
completion	cold start recommendations	high	tensor factorization
aware	rolling shutter	estimation	non parametric
personalized	non uniform	kernel	singular value
tensor	user friendly	spectral	high order
non	user centric	regression	matrix approximation
user	non overlapping	matrices	matrix decomposition
context	non redundant	non	non linear
nonnegative	non repeating	tensor	matrix factorizations
tag	non euclidean	norm	low dimensional
preference	non existence	covariance	divide conquer
item	recommender semantically	nonnegative	missing values
probabilistic	non monotonic	pca	high dimensions
ranking	non player	missing	near optimal
recommender	non concatenative	means	low level
Outlier Detection			
detection	detection tracking	detection	scan statistic
object	detection crowded scenes	outlier	scan statistics
outlier	haar features	anomaly	land cover
change	detection videos	based	scan window
anomaly	detection natural	change	anomalous behaviour
pedestrian	pointer swizzling	concept	anomalous window
saliency	detection evolving	event	remotely sensed
robust	detection viewpoint	network	analog vlsi
event	detection unfamiliar	application	land vehicle
edge	detection segment	drift	restructuring sparse
detectors	anomalous behaviour	novelty	anomalous particles
fast	detection benchmark dataset	ensemble	quantum computation
boundary	anomalous window	intrusion	quantum mechanics
novelty	anomalous behavior	anomalies	analog resistive
near	epileptic seizure	streaming	anomalous samples
duplicate	detection result	scan	entropic estimator
cascade	detection gpgpu	point	jet engine vibration
detector	detection ecommerce	classifier	defence ransac
scan	detection hungarian	online	white functionals
rapid	detection frontal	detect	matches clickthrough

Table 9 (continued)

TNG		PTNG	
n-gram(1)	n-gram(2+)	n-gram(1)	n-gram(2+)
Boltzmann Machines			
support	vector machines	neural	boltzmann machines
vector	vector machine	networks	boltzmann machine
learning	restricted boltzmann	network	feed forward
classification	vector regression	deep	short term
kernel	vector space model	learning	multilayer perceptrons
kernels	vector quantization	convolutional	sum product
machines	vector fields	recurrent	short categorization
algorithm	vector space models	belief	higgs boson
incremental	vector method	boltzmann	short texts
adaptive	vector space	propagation	sum parts
application	vector valued	restricted	loosely synchronized
training	vector spaces	recursive	feed opinions
vectors	vector based	bayesian	rolling incorporating
svm	vector representations	classification	boltzmann perceptron
restricted	vector classifiers	training	rolling plate
regularization	von mises	nets	boltzmann dynamics
fisher	vector compression	continuous	boltzmann units collective
structural	vector quantizer	architecture	loosely organized
boltzmann	restricted training	representations	short conversation
convolution	vector space retrieval	connectionist	short snippets
Natural Language Processing			
extraction	distant supervision	entity	entity recognition
entity	pronoun resolution	word	entity linking
resolution	zero anaphora	sense	distant supervision
relation	zero pronouns	disambiguation	fine grained
named	zero shot	named	entity disambiguation
recognition	zero pronoun resolution	extraction	entity extraction
coreference	pronominal anaphora	relation	entity resolution
event	zero crossings	entities	entity search
entities	zero loss	identification	entity transliteration
linking	slot filling	unsupervised	entity based
features	zero knowledge	embeddings	entity type
joint	pronoun referent	wikipedia	entity set expansion
exploiting	zero pronoun detection	exploiting	entity detection
reference	pronoun interpretation	joint	entity centric
zero	identical events	senses	entity mentions
relations	identical mrf	inference	entity types
semantic	distant spelling	distant	entity matching
chinese	pronominal anaphor	bootstrapping	entity evolution
discourse	zero adnominal	biomedical	entity classification
wikipedia	pronominal translation	person	entity normalization

We give the semantic summary of each topic on top of the word list, *i.e.*, Collaborative Filtering, Outlier Detection, Boltzmann Machines and Natural Language Processing. For each topic, we show the unigrams and n-grams ($n > 1$), which are the topic's top-20 most probable words.

topic, some generic word like "detection" or "outlier" is achieved to rank very high in TNG's unigrams, however, TNG fails to deliver meaningful phrases for outlier detection. On the contrary, PTNG succeeds in capturing rarely-mentioned phrases in the related literatures of outlier detection, *e.g.* "scan statistics" for anomalous window discovery and "land cover" for change detection. In the topic of "Boltzmann Machines", TNG cannot well separate the phrase "restricted boltzmann" from "vector machines", while in contrast PTNG accurately relates "boltzmann machines" with neural networks. Moreover, in the topic of "Natural Language Processing", PTNG achieves to form several key phrases in NLP such as "entity recognition", "entity linking", "entity extraction", *etc.*, whereas TNG cannot form these phrases.

4.8 Parameter sensitivity

4.8.1 Impact of number of pseudo documents

Since our proposed PTNG model learns topics from P pseudo documents, we take P as the key parameter for our method. Specifically, we set different numbers of topics $K = \{50, 100, 150, 200\}$ and under these different values of K, we vary P in $\{50, 100, 500, 1000, 1500, 2000\}$. We show the performance of PTNG in terms of UCI topic coherence on News and DBLP in Figure 3. Then we give the performance of PTNG in terms of classification f-measure in 5-fold cross validation on News, DBLP and Tweets in Figure 4.

Firstly, topic model with small value of P produces less coherent topics and thus results in less accurate short text classification performance. As illustrated in Figure 3, on both datasets under different values of K, when P equals to 50, PTNG produces less coherent topics and better topics since $P \geq 500$. Such topic coherence results are in accordance with the classification performance as shown in Figure 4 except for the case on DBLP when $K = 50$. Another interesting observation is that when K is large enough, *e.g.*, $K = 200$, a small P can weaken the performance of topic models in terms of both topic coherence and classification accuracy to a greater extent than the case when K is small, *e.g.*, $K = 50$. This may indicate that large value of pseudo documents is essential to the success of topic models when large number of topics need to be learned from the corpus.

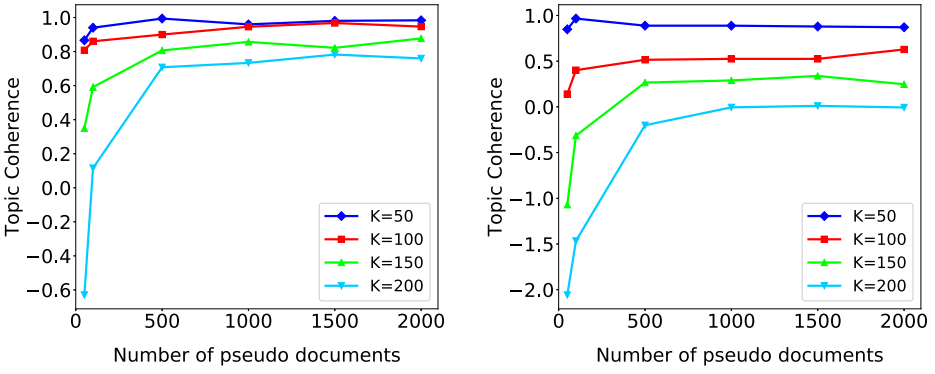


Figure 3 Variation of UCI topic coherence on News and DBLP with the number of pseudo documents

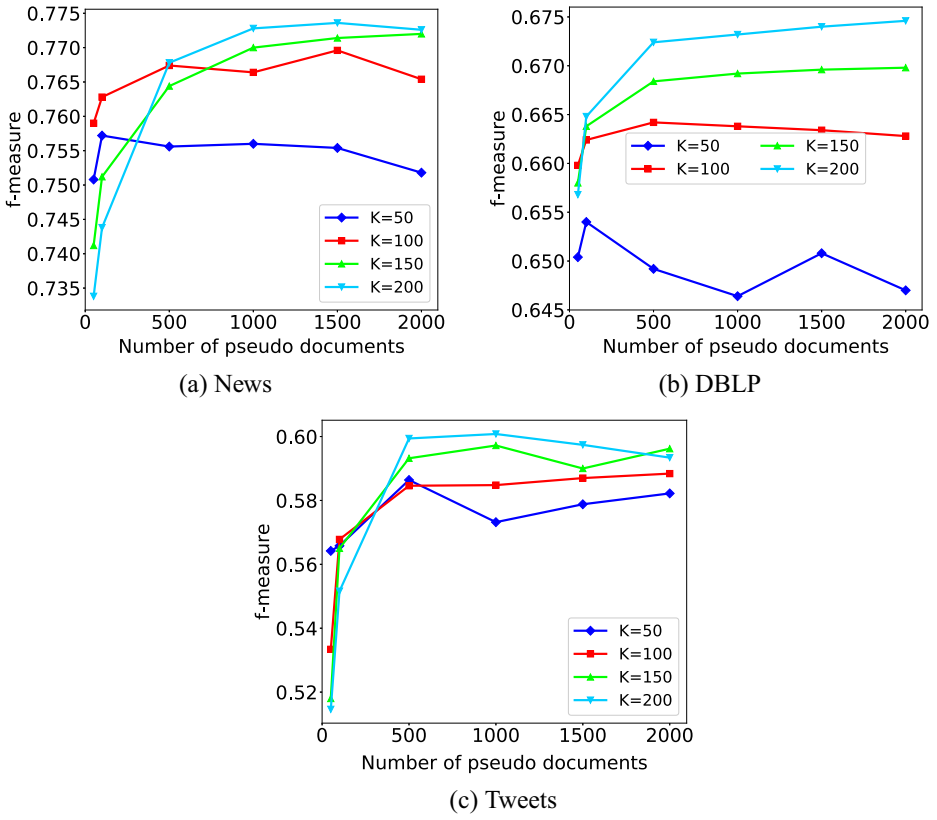
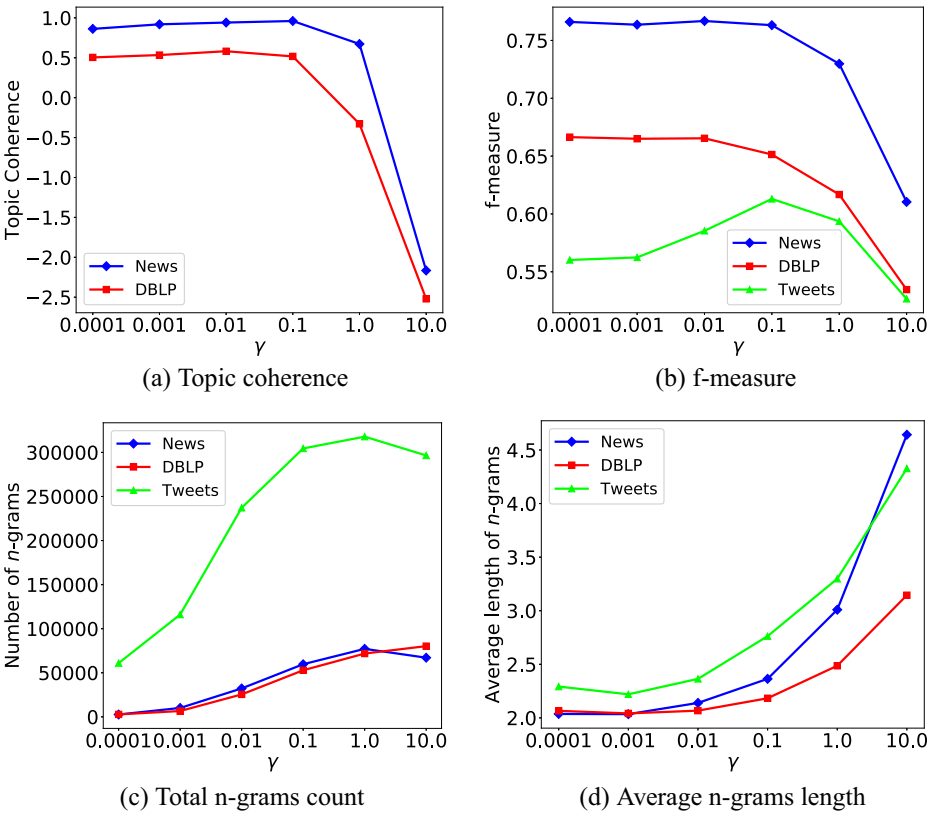


Figure 4 Variation of classification results on News, DBLP and Tweets with the number of pseudo documents

4.8.2 Impact of the γ value

As another key parameter, γ consists of the shape parameters γ_0, γ_1 and denote the Beta prior parameters for π, which governs the probability of forming collocations or phrases in PTNG. Specifically, the larger the value of γ_0 is, the larger the probability of x being 0 is, which means unigrams are more likely to be formed. On the contrary, the larger the value of γ_1 is, the larger the probability of x being 1 is, which means n-grams ($n > 1$) are more likely to be formed. To study the impact of γ, we fix the value of γ_0 to 0.1 and vary the value of γ_1 in $\{0.0001, 0.001, 0.01, 0.1, 1.0, 10.0\}$. We study the performance of PTNG on News, DBLP and Tweets, of which the results are shown in Figure 5.

From Figure 5-a, we can see that the UCI topic coherence first increases slightly with increasing value of γ_1, however with further increasing γ_1, when γ_1 is greater than 0.1, PTNG starts to show significant decline in UCI topic coherence on all datasets. Figure 5-b shows that short text classification performance of PTNG is insensitive to the fluctuation of γ_1 when γ_1 is less than 1.0 on all datasets except for Tweets and with further increasing value of γ_1, the classification performance of PTNG also follows a sharp decrease. This indicates that when γ_1 is small, increasing the value of γ_1 forms more n-grams, which can indeed help to produce more coherent topics. To further exploit the reason of the sharp



(a) Topic coherence (b) f-measure

(c) Total n-grams count (d) Average n-grams length

Figure 5 Variation of PTNG performance on News, DBLP and Tweets with γ. **a**: UCI topic coherence; **b**: f-measure for short text classification; **c**: Total number of n-grams; **d**: Average length of n-grams

decrease of both metrics when γ_1 is large, we plot the total number of n-grams and the average length of n-grams formed by PTNG with varying values of γ_1 in Figure 5-c and Figure 5-d, respectively. From these two figures, we can observe that the PTNG is likely to form more n-grams with increasing value of γ_1 when γ_1 is small, whereas when γ_1 is large enough (*i.e.*, $\gamma_1 > 0.1$), PTNG prefers to form longer phrases rather than more phrases. This might weaken the performance of topic models in terms of both UCI topic coherence and f-measure, since PTNG may produce unreasonably and unexpectedly long phrases in large number of documents when γ_1 is large. As such, the value of γ_1 is recommended not to be set to larger than 0.1, and in our experiments, we set γ_1 to 0.01 as our default setting.

5 Conclusion

In this paper, we propose a Pseudo-document-based Topical N-Gram model (PTNG) for short text topic modeling. PTNG benefits tremendously from self aggregating massive short texts into much less pseudo documents without incorporating any auxiliary contextual information. Moreover, PTNG also includes latent bigram status variables into the generative process to detect collocations automatically. This way of modeling makes PTNG sensitive

to word order in short texts, and helps PTNG significantly outperform the model that solely relies on pseudo documents, e.g. PTM, and other state-of-the-art short text topic models. The advantages of PTNG have been justified by the experiments on three real-world short text corpora in terms of both topic coherence and classification evaluation measures. To our best knowledge, this work is among the earliest studies in using self aggregation and topical n-grams simultaneously for short text topic modeling.

Acknowledgements Dr. Junjie Wu's work was partially supported by the National Key R&D Program of China (2019YFB2101804), and the National Natural Science Foundation of China (U1636210, 71725002, 71531001). Dr. Guannan Liu was supported in part by NSFC under Grants 71701007. Dr. Yuan Zuo was partially supported by the National Natural Science Foundation of China (NSFC) under Grant 71901012, and by the China Postdoctoral Science Foundation under Grant 2018M640045. Dr. Hong Li was partially supported by NSFC under Grants 71471009. Dr. Zhiang Wu was supported by Industry Projects in Jiangsu S&T Pillar Program under Grant No. BE201910.

References

1. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. J. Mach. Learn. Res. **3**, 993–1022 (2003)
2. El-Kishky, A., Song, Y., Wang, C., Voss, C.R., Han, J.: Scalable topical phrase mining from text corpora. Proc. VLDB Endow. **8**(3), 305–316 (November 2014)
3. Griffiths, T.L., Steyvers, M.: Finding scientific topics. Proc. Natl. Acad. Sci. **101**, 5228–5235 (2004)
4. Griffiths, T.L., Tenenbaum, J.B., Steyvers, M.: Topics in semantic representation. Psychol. Rev. **114**, 2007 (2007)
5. Huang, J., Peng, M., Wang, H., Cao, J., Gao, W., Zhang, X.: A probabilistic method for emerging topic tracking in microblog stream. World Wide Web **20**(2), 325–350 (March 2017). https://doi.org/10.1007/s11280-016-0390-4. https://doi.org/10.1007/s11280-016-0390-4
6. He, Y.: Extracting topical phrases from clinical documents. In: Thirtieth AAAI Conference on Artificial Intelligence, AAAI'16, pp. 2957–2963 (2016)
7. Hong, L., Davison, B.D.: Empirical study of topic modeling in twitter. In: Proceedings of the first workshop on social media analytics, pp. 80–88 (2010)
8. Ishwaran, H., Rao, J.S.: Spike and slab variable selection: frequentist and bayesian strategies. Ann. Stat. **33**(2), 730–773 (2005)
9. Jin, O., Liu, N.N., Zhao, K., Yu, Y., Yang, Q.: Transferring topical knowledge from auxiliary long texts for short text clustering. In: Proceedings of the 20th ACM international conference on Information and knowledge management, pp. 775–784 (2011)
10. Lin, T., Tian, W., Mei, Q., Cheng, H.: The dual-sparse topic model: Mining focused topics and focused terms in short text. In: Proceedings of the 23rd international conference on World wide web, pp. 539–550 (2014)
11. Lindsey, R.V., Headden, W.P., Stipicevic, M.J.: A phrase-discovering topic model using hierarchical pitman-yor processes. In: Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL '12, pp. 214–222 (2012)
12. Li, B., Yang, X., Zhou, R., Wang, B., Liu, C., Zhang, Y.: An efficient method for high quality and cohesive topical phrase mining. IEEE Trans. Knowl. Data Eng. **31**(1), 120–137 (2019Jan). https://doi.org/10.1109/TKDE.2018.2823758
13. Li, B., Wang, B., Zhou, R., Yang, X., Liu, C.: Citpm: A cluster-based iterative topical phrase mining framework. In: Database Systems for Advanced Applications, pp 197–213. Springer International Publishing, Cham (2016)
14. Lau, J.H., Baldwin, T., Newman, D.: On collocations and topic models. ACM Trans. Speech Lang. Process. **10**(3), 10:1–10:14 (July 2013). https://doi.org/10.1145/2483969.2483972. http://doi.acm.org/10.1145/2483969.2483972
15. Li, C., Duan, Y., Wang, H., Zhang, Z., Sun, A., Ma, Z.: Enhancing topic modeling for short texts with auxiliary word embeddings. ACM Trans. Inf. Syst. **36**(2), 11:1–11:30 (August 2017)
16. Mehrotra, R., Sanner, S., Buntine, W., Xie, L.: 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 in information retrieval, pp. 889–892 (2013)

17. Mimno, D., Wallach, H.M., Talley, E., Leenders, M., McCallum, A.: Optimizing semantic coherence in topic models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 262–272 (2011)
18. Nugroho, R., Zhao, W., Yang, J., Paris, C., Nepal, S.: Using time-sensitive interactions to improve topic derivation in twitter. World Wide Web **20**(1), 61–87 (January 2017). https://doi.org/10.1007/s11280-016-0417-x. https://doi.org/10.1007/s11280-016-0417-x
19. Nigam, K., McCallum, A., Thrun, S., Mitchell, T.: Text classification from labeled and unlabeled documents using em. Mach. Learn. **39**(2-3), 103–134 (2000)
20. Newman, D., Lau, J.H., Grieser, K., Baldwin, T.: Automatic evaluation of topic coherence. In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pp. 100–108 (2010)
21. Quan, X., Kit, C., Ge, Y., Pan, S.J.: Short and sparse text topic modeling via self-aggregation. In: Proceedings of the 24th International Conference on Artificial Intelligence, pp. 2270–2276 (2015)
22. Tang, J., Zhang, M., Mei, Q.: 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, pp. 5–13 (2013)
23. Tang, J., Li, C., Zhang, M., Mei, Q.: Less is more: Learning prominent and diverse topics for data summarization. arXiv:1611.09921 (2016)
24. Wang, X., McCallum, A.: 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, pp. 424–433 (2006)
25. Wallach, H.M.: Topic modeling: Beyond bag-of-words. In: Proceedings of the 23rd International Conference on Machine Learning, ICML '06, pp. 977–984 (2006)
26. Wang, X., McCallum, A.: A note on topical n-grams. Tech. rep., MASSACHUSETTS UNIV AMHERST DEPT OF COMPUTER SCIENCE (2005)
27. Wang, X., McCallum, A., Wei, X.: Topical n-grams: Phrase and topic discovery, with an application to information retrieval. In: Proceedings of the 2007 Seventh IEEE International Conference on Data Mining, ICDM '07, pp 697–702. IEEE Computer Society, Washington, DC, USA (2007)
28. Weng, J., Lim, E.-P., Jiang, J., He, Q.: Twitterrank: Finding topic-sensitive influential twitterers. In: Proceedings of the third ACM international conference on Web search and data mining, pp. 261–270 (2010)
29. Wang, C., Blei, D.M.: Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process. In: Advances in neural information processing systems, pp 1982–1989. Curran Associates Inc. (2009)
30. Wang, C., Danilevsky, M., Desai, N., Zhang, Y., Nguyen, P., Taula, T., Han, J.: A phrase mining framework for recursive construction of a topical hierarchy. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '13, pp 437–445. ACM, New York, NY, USA (2013). http://doi.acm.org/10.1145/2487575.2487631
31. Yan, X., Guo, J., Lan, Y., Cheng, X.: A biterm topic model for short texts. In: Proceedings of the 22Nd International Conference on World Wide Web, WWW '13, pp 1445–1456. ACM, New York, NY, USA (2013). http://doi.acm.org/10.1145/2488388.2488514
32. Yang, Y., Wang, F., Zhang, J., Xu, J., Yu, P.S.: A topic model for co-occurring normal documents and short texts. World Wide Web **21**(2), 487–513 (March 2018). https://doi.org/10.1007/s11280-017-0467-8. https://doi.org/10.1007/s11280-017-0467-8
33. Yin, J., Wang, J.: A dirichlet multinomial mixture model-based approach for short text clustering. In: Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 233–242 (2014)
34. 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: Advances in Information Retrieval, pp. 338–349 (2011)
35. Zuo, Y., Zhao, J., Xu, K.: Word network topic model: a simple but general solution for short and imbalanced texts. Knowl. Inf. Syst. **48**(2), 379–398 (2016)
36. Zuo, Y., Wu, J., Zhang, H., Lin, H., Wang, F., Xu, K., Xiong, H.: Topic modeling of short texts: A pseudo-document view. In: Proceedings of the 22Nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, pp 2105–2114. ACM, New York, NY, USA (2016)
37. Zubiaga, A., Ji, H.: Harnessing web page directories for large-scale classification of tweets. In: Proceedings of the 22nd international conference on World Wide Web companion, pp. 225–226 (2013)

Affiliations

Hao Lin[1] · Yuan Zuo[1] · Guannan Liu[1] · Hong Li[1] · Junjie Wu[1,2,3] · Zhiang Wu[4]

Hao Lin
linhao2014@buaa.edu.cn

Guannan Liu
liugn@buaa.edu.cn

Hong Li
hong_lee@buaa.edu.cn

Junjie Wu
wujj@buaa.edu.cn

Zhiang Wu
zawu@seu.edu.cn

1 School of Economics and Management, Beihang University, Beijing, 100191, China

2 Beijing Advanced Innovation Center for Big Data and Brain Computing, Beihang University, Beijing, 100191, China

3 Beijing Key Laboratory of Emergency Support Simulation Technologies for City Operations, Beihang University, Beijing, 100191, China

4 Jiangsu Provincial Key Laboratory of E-Business, Nanjing University of Finance and Economics, Nanjing, China