Short Text Topic Modeling with Topic Distribution Quantization and Negative Sampling Decoder

Xiaobao Wu[1] **Chunping Li**[1] **Yan Zhu**[2] **Yishu Miao**[3]
[1]Tsinghua University [2]Southwest Jiaotong University [3]Imperial College London
wxb18@mails.tsinghua.edu.cn, cli@mail.tsinghua.edu.cn
yzhu@swjtu.edu.cn, y.miao20@imperial.ac.uk

Abstract

Topic models have been prevailing for many years on discovering latent semantics while modeling long documents. However, for short texts they generally suffer from data sparsity because of extremely limited word co-occurrences; thus tend to yield repetitive or trivial topics with low quality. In this paper, to address this issue, we propose a novel neural topic model in the framework of autoencoding with a new topic distribution quantization approach generating peakier distributions that are more appropriate for modeling short texts. Besides the encoding, to tackle this issue in terms of decoding, we further propose a novel negative sampling decoder learning from negative samples to avoid yielding repetitive topics. We observe that our model can highly improve short text topic modeling performance. Through extensive experiments on real-world datasets, we demonstrate our model can outperform both strong traditional and neural baselines under extreme data sparsity scenes, producing high-quality topics.

1 Introduction

In addition to formal documents, short texts play an increasingly more important role in the era of information explosion where people could instantly share ideas, feelings, and comments via short text fragments, including tweets, headlines, and product reviews, etc. The latent semantics or topics discovered among these short texts can be utilized in many applications, such as content summarization (Ma et al., 2012), classification(Zeng et al., 2018a), and recommendations (Zeng et al., 2018b; Mehrotra et al., 2013). However, conventional topic models (Blei et al., 2003) work reasonably well on various kinds of long documents, but perform poorly on short texts. The main underlying reason is that the co-occurrence information from short texts is extremely limited as known as the data sparsity problem which hinders the topic models from learning effective semantics and high-quality topics in a pure unsupervised learning fashion. Therefore, several approaches have been proposed to alleviate this issue. One simple approach is to yield pseudo texts (Quan et al., 2015), so that the conventional topic models can apply, e.g., user data (Weng et al., 2010), hashtags (Mehrotra et al., 2013) and external corpora (Zuo et al., 2016), but auxiliary information is not always available. In another vein, extra structural information or semantics are incorporated with the models. For instance, Biterm Topic Model (BTM) (Yan et al., 2013) directly constructs the topic distributions over unordered word-pairs (biterms); Generalized Pólya Urn-DMM (GPUDMM) (Li et al., 2016) applies auxiliary pre-trained word embeddings to introduce external information from other sources. However, the data sparsity problem of short texts remains to be solved, especially resulting in repetitive and trivial topics. For example, as illustrated in Table 1, we can see several repetitive topics about sports including repeated words like "football", "games", and "tennis", and trivial topics composed of incoherent words are discovered from short texts. These topics are of low quality and could impair the performance of downstream tasks.

sports scores games soccer league tennis ncaa players football
sports tennis soccer hockey games football beach match players
sports match cup hockey olympic football players sport league
sports football sport league games tennis champions club
sports football league game tennis players hockey games scores
bad additional abstract aspectj behave displayed customise accept
abstract behave accept additional bad displayed customise
abstract accept behave additional adding long many administration

Table 1: Repetitive and trivial topics from short texts. Repetitive words are underlined.

In this paper, we aim to design a model that can generate high-quality topics from short texts and is more robust to rigorous data sparsity scenarios

Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, pages 1772–1782,
November 16–20, 2020.

without any auxiliary corpus. Different from previous methods, we propose a new **Negative sampling and Quantization Topic Model (NQTM)** in an auto-encoding framework to address the unsupervised short text modeling problem including two essential and novel methods. First, for short texts, we need peakier topic distributions for decoding since short texts cover few primary topics, like Dirichlet Multinomial Mixture (DMM) (Nigam et al., 2000; Yin and Wang, 2014) that assumes each short text only covers one topic. In the autoencoding framework, a possible and straightforward way is using gumbel-softmax (Jang et al., 2016), but its performance is highly determined by the temperature parameter that necessarily needs to be tuned across topic numbers and corpora; therefore, it may not guarantee high-quality topics. Another way is quantizing the latent representations like VQ-VAE (van den Oord and Vinyals, 2017). Unfortunately, the original quantization of VQ-VAE is for image generation and cannot produce peakier distributions for short text topic modeling. Therefore, we propose the novel topic distribution quantization for short texts by separably mapping topic distributions into an appropriate defined embedding space. With this new method, our model can naturally encourage discretization to flexibly yield peakier distributions for decoding, resulting in much better topic quality performance.

Second, we propose a new negative sampling decoder to improve the topic diversity performance. As mentioned previously, short texts are extremely sparse inputs, so the learning signals are too weak to converge to a good local minimum, notably in an unsupervised learning fashion, leading to repetitive topics. Therefore, instead of using a straightforward log-likelihood objective, we propose a negative sampling decoder with the reconstruction by selecting target words from assigned topics and negative words from the topics that are unlikely to be assigned. It acts as an inductive bias that encourages the topic-word distributions to be pushed away from each other, resulting in a better learning objective for generating diverse topics. The main contributions[1] of this paper can be concluded as

- We propose a neural model with a novel topic distribution quantization method to produce peakier distributions for improving short text topic modeling;
- We also propose a negative sampling decoder to enhance the diversity of short text topics instead of conventional log-likelihood maximization;
- We conduct comprehensive experiments on real-world datasets and demonstrate that our model can effectively alleviate the data sparsity problem and generate higher quality topics for short texts (more coherent and diverse);
- We further discuss the trade-off of short text topic models between topic coherence and diversity in detail and show our model outperforms baselines on both these aspects.

[1]The code is available at https://github.com/bobxwu/NQTM

2 Related Work

Conventional topic models Conventional probabilistic topic models, e.g., Probabilistic Latent Semantic Analysis (PLSA) (Hofmann, 1999) and Latent Dirichlet Allocation (LDA) (Blei et al., 2003), work very well on formal documents with long texts. To improve the performance of short text topic modeling, Biterm Topic Model (BTM) (Yan et al., 2013) and Dirichlet Multinomial Mixture (DMM) model (Nigam et al., 2000; Sadamitsu et al., 2007; Yin and Wang, 2014) are two basic short text probabilistic topic models which employ traditional Bayesian inference methods including Gibbs Sampling (Steyvers and Griffiths, 2007) and Variational Inference (Blei et al., 2017). Several extensions based on BTM and DMM are also proposed, such as Generalized Pólya Urn-DMM (GPUDMM) (Li et al., 2016) with word embeddings and Multiterm Topic Model (Wu and Li, 2019). Besides, Semantics-assisted Non-negative Matrix Factorization (SeaNMF) (Shi et al., 2018) was lately proposed as an NMF topic model incorporating word-context semantic correlations solved by a block coordinate descent algorithm.

Neural topic models More recently, deep neural networks have shown great potential for learning complicated distributions for unsupervised models. Due to the success of Variational AutoEncoder (VAE) (Kingma and Welling, 2014; Rezende et al., 2014), various neural topic models are proposed (Nan et al., 2019; Wu et al., 2020). Neural Variational Document Model (NVDM) (Miao et al., 2016) is the first VAE-based neural topic model that adopts the reparameterization trick of Gaussian distributions and achieves remarkable results

on normal text topic modeling. Some extensions like Gaussian Softmax Construction (GSM) have been explored in (Miao et al., 2017). Product of expert LDA (ProdLDA) is proposed by Srivastava and Sutton (2017) using Logistic Normal distribution due to the difficulty of taking the reparameterization trick for Dirichlet distribution, which is important for topic modeling. Topic Memory Network (TMN) (Zeng et al., 2018a) is proposed for supervised short text topic modeling and classification with pre-trained word embeddings, incorporating the neural topic model (Miao et al., 2016) with memory networks (Weston et al., 2014). Different from these neural topic models, the proposed model aims to improve short text topic modeling without any extra information. Our model relies on the novel topic distribution quantization to discrete the latent representations in the auto-encoding framework instead of the VAE assumption. Meanwhile, a new objective under the negative sampling decoder replaces the traditional log-likelihood maximization objective to especially alleviate the data sparsity of short texts.

3 Negative sampling and Quantization Topic Model

3.1 A Brief Review of Topic Models

LDA (Blei et al., 2003) is one of the most classic probabilistic topic models. In its formulation, a topic is defined as a distribution of words and each word in a text is drawn from a mixture of Multinomial distributions with Dirichlet distribution as the priori. In LDA, the latent variable z denotes the topic assignment of word x_i and $\boldsymbol{\theta}$ is the topic distribution of a text. According to the generation procedure of LDA, the marginal likelihood of a text $\boldsymbol{x}$ is

$$p(\boldsymbol{x}|\boldsymbol{\alpha},\boldsymbol{\beta}) = \int_{\boldsymbol{\theta}} \left(\prod_{i=1}^{N} \sum_{z=1}^{K} p\left(x_i|z,\boldsymbol{\beta}\right) p\left(z|\boldsymbol{\theta}\right) \right) p(\boldsymbol{\theta}|\boldsymbol{\alpha}) d\boldsymbol{\theta}$$

where N refers to the number of words in text $\boldsymbol{x}$, $\boldsymbol{\alpha}$ is the hyperparameter of Dirichlet distribution, $\boldsymbol{\beta}_z$ refers to the topic distribution over words given the topic assignment z and $\boldsymbol{\beta} = (\boldsymbol{\beta}_1, \ldots, \boldsymbol{\beta}_K) \in \mathbb{R}^{V \times K}$ is the matrix of all topic words probability vectors (V is the vocabulary size and K is the topic numbder). Then, approximation methods, like Variational Inference or Gibbs Sampling, are employed to approximate the intractable posterior.

In a different way, with the help of neural variational inference, neural topic models (Miao et al., 2017; Srivastava and Sutton, 2017) have been proposed to simplify the inference and the model can be directly updated by gradient backpropagation. These models adopt a simplification that the discrete latent variable z is integrated out in the marginal likelihood as

$$p(\boldsymbol{x}|\boldsymbol{\alpha},\boldsymbol{\beta}) = \int_{\boldsymbol{\theta}} \left(\prod_{i=1}^{N} p\left(x_i|\boldsymbol{\theta},\boldsymbol{\beta}\right) \right) p(\boldsymbol{\theta}|\boldsymbol{\alpha}) d\boldsymbol{\theta} \quad (1)$$

Based on these preceding neural topic models, we present our proposed model for short text topic modeling.

3.2 Network Architecture

In this section, we detail the proposed Negative sampling and Quantization Topic Model (NQTM). Figure 1 shows the overall architecture including three main parts.

3.2.1 Short Text Encoder

Topic models discover semantic information (topics) among large unlabeled datasets using word co-occurrence, so topic models typically apply the bag-of-words assumption ignoring the sequence for simplification. Thus, we adopt MLPs that are eligible enough for both encoder and decoder. We assume the short text $\boldsymbol{x}$ is in the form of bag-of-words which produces continuous representations through the short text encoder. We adopt the following simple network structure as our short text encoder:

$$\boldsymbol{\pi}_1 = \zeta(\boldsymbol{W}_1\boldsymbol{x} + \boldsymbol{b}_1) \quad (2)$$

$$\boldsymbol{\pi}_2 = \zeta(\boldsymbol{W}_2\boldsymbol{\pi}_1 + \boldsymbol{b}_2) \quad (3)$$

$$\boldsymbol{\theta}_e = \sigma(\boldsymbol{\pi}_2) \quad (4)$$

where $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ are linear transformations, and $\boldsymbol{\pi}_1$ and $\boldsymbol{\pi}_2$ are intermediate outputs; $\sigma(\cdot)$ means softmax function for normalization and $\zeta(\cdot)$ denotes softplus function. After the encoder, we have the lower dimensional representation $\boldsymbol{\theta}_e$ of the short text $\boldsymbol{x}$.

3.2.2 Topic Distribution Quantization

Instead of directly feeding the continuous representation $\boldsymbol{\theta}_e$ to the decoder as previous neural topic models (Miao et al., 2016, 2017; Srivastava and Sutton, 2017), we employ the quantization step ahead. Unfortunately, we find that directly using

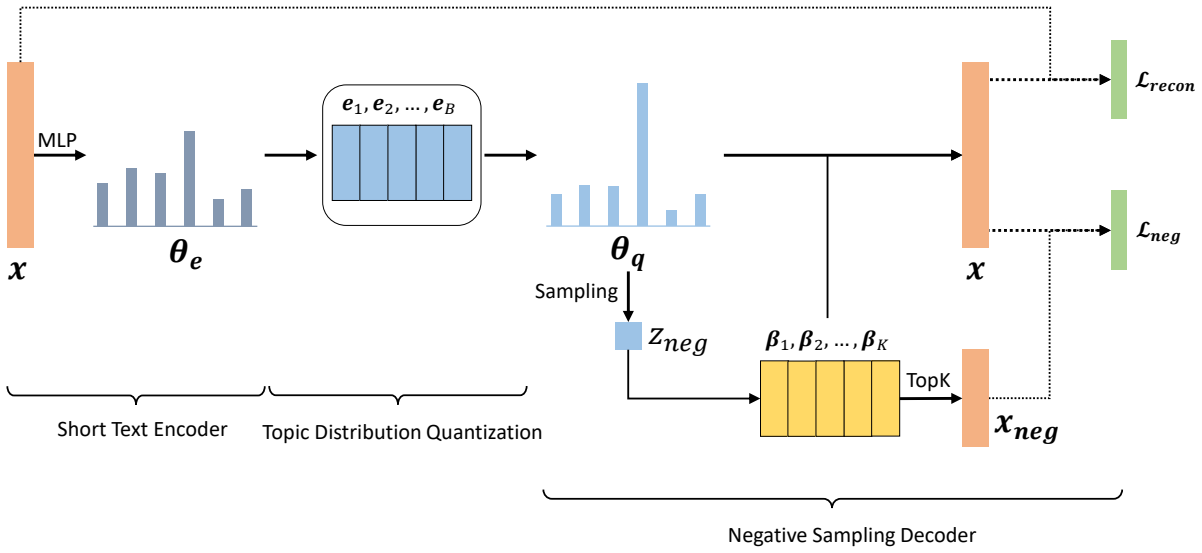


Figure 1: The overall architecture of NQTM with three main components including the short text encoder, the novel topic distribution quantization for short texts, and the new negative sampling decoder.

ordinary quantization is not a guarantee for better topic quality, because the latent representations can not be distinguished during optimization. More precisely, since the original embedding space of VQ-VAE is randomly initialized with uniform distributions, these embedding vectors of VQ-VAE are too close to each other to distinguish. Thus, it is arduous for the model to learn to separably map the latent representations of different topics to the embedding vectors, resulting in extremely repetitive topics.

To this end, we propose a novel topic distribution quantization method to alleviate the data sparsity problem of short texts especially. We first set a discrete embedding space $\boldsymbol{e} = (\boldsymbol{e}_1, \boldsymbol{e}_2, \ldots, \boldsymbol{e}_B) \in \mathbb{R}^{K \times B}$ where B is the size of the embedding space. To encourage the maximum of distances between embedding vectors and have peakier topic distributions, the first K vectors $(\boldsymbol{e}_1 \cdots \boldsymbol{e}_K)$ are initialized with identity matrix and the remaining vectors $(\boldsymbol{e}_{K+1} \cdots \boldsymbol{e}_B)$ are initialized with uniform unit scaling $Uniform\big(-\sqrt{3/K}, \sqrt{3/K}\big)$. Therefore, the embedding space $\boldsymbol{e}$ can be written as

$$\boldsymbol{e} = \begin{bmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \end{bmatrix} \quad \boldsymbol{e}_{K+1} \quad \ldots \quad \boldsymbol{e}_B \qquad (5)$$

which can be seen as an extended identity matrix. The continuous representation $\boldsymbol{\theta}_e$ is mapped to the nearest vector $\boldsymbol{\theta}_q$ of the embedding space $\boldsymbol{e}$ as

$$\boldsymbol{\theta}_q = \boldsymbol{e}_k, \text{ where } k = \operatorname{argmin}_j \|\boldsymbol{\theta}_e - \boldsymbol{e}_j\|_2 . \quad (6)$$

In this way, the proposed new quantizing topic distributions method for short texts can make the latent representations separably map to distinguished embedding vectors and flexibly generate peakier topic distributions, which can stimulate our model to tackle the data sparsity and improve the diversity and coherence of topics.

3.2.3 Negative Sampling Decoder

After the topic distribution quantization, $\boldsymbol{\theta}_q$ is fed to the decoder for reconstruction. It has been found that normalizing topic words probability matrix $\boldsymbol{\beta}$, such as $\sigma(\boldsymbol{\beta})$, results in trivial and less discriminative topics (Srivastava and Sutton, 2017). Hence, according to Equation (1), the reconstruction of a word x_i in the text $\boldsymbol{x}$ is modeled as $x_i \sim \text{Mult}(\sigma(\boldsymbol{\beta}\boldsymbol{\theta}_q))$.

Negative sampling algorithm In contrast to the standard decoder with log-likelihood maximization objective function, we propose to take advantage of the negative sampling scheme and formulate a new decoder to generate more diverse topics. Similar ideas are mentioned in some data sparsity fields like collaborative filtering (Liang et al., 2018) where if for a short text , the negative samples simply are all the words that do not exist in it. But this method is unable to distinguish the words from different topics explicitly.

Thus, instead of applying this simple solution, we further propose the negative sampling decoder. We take the words with high probabilities in the other topics but not assigned to the current text fragment as negative samples. The intuition is to strengthen the discrimination between the words drawn from the assigned topic distribution and a negative draw from other topics that are not assigned to the text. Therefore, we introduce an inductive bias that prompts the topic-word distributions to be pushed away from each other. In the meantime, the neural model benefits from a better learning signal other than the ordinary softmax loss. As shown in Figure 1, given a short document and its topic distribution, we first remove the top t probable topics and sample one negative topic z_{neg} from the left $(K-t)$ topics with equal probability, which is

$$z_{neg} \sim \text{Mult}(\boldsymbol{p}, 1) \tag{7}$$

where $\boldsymbol{p} = (p_1, p_2, \ldots, p_K)$ and p_k is the probability of choosing topic k, defined as

$$p_k = \begin{cases} 0 & \text{topic } k \text{ is included in top } t \text{ topics} \\ \frac{1}{K-t} & \text{otherwise} \end{cases}$$

Therefore, z_{neg} represents a topic that the document is unlikely to cover because of its low probability to be assigned. Then, we generate M words from $\boldsymbol{\beta}_{z_{neg}}$ by TopK function as

$$\boldsymbol{x}_{neg} = \text{TopK}(\boldsymbol{\beta}_{z_{neg}}, M) \tag{8}$$

where $\boldsymbol{x}_{neg}$ denotes the M words that topic z_{neg} is more likely to contain. But since the document is supposed to not cover z_{neg}, the decoder should avoid generating them during reconstruction. This heuristic acts as a positive bias to help the model discover high-quality topics and the negative samples $\boldsymbol{x}_{neg}$ can amplify the learning signals for better optimizing the neural model and improving topic diversity.

Objective function With the negative sampling decoder, we can then construct our objective function. The reconstruction error and the negative sampling error are

$$\mathcal{L}_{recon}(\boldsymbol{x}^{(i)}) = -\boldsymbol{x}^{(i)} \cdot \log\big(\sigma(\boldsymbol{\beta}\boldsymbol{\theta}_q^{(i)})\big) \tag{9}$$

$$\mathcal{L}_{neg}(\boldsymbol{x}^{(i)}) = -\boldsymbol{x}_{neg}^{(i)} \cdot \log\big(1 - \sigma(\boldsymbol{\beta}\boldsymbol{\theta}_q^{(i)})\big) \tag{10}$$

where $\boldsymbol{x}^{(i)}$ refers to the i-th short text in the corpus. As indicated previously, $\boldsymbol{\theta}_e^{(i)}$ means the latent representation outputted by the encoder for $\boldsymbol{x}^{(i)}$ and $\boldsymbol{\theta}_q^{(i)}$ is the discrete representation after the new topic distribution quantization part. We apply the cross-entropy between inputs $\boldsymbol{x}^{(i)}$ and $\sigma(\boldsymbol{\beta}\boldsymbol{\theta}_q^{(i)})$ to calculate the reconstruction error. For the negative sampling error, we also use the cross-entropy between $\boldsymbol{x}_{neg}^{(i)}$ and $(1 - \sigma(\boldsymbol{\beta}\boldsymbol{\theta}_q^{(i)}))$ to enrich learning signals. Therefore, the overall training objective with the negative sampling decoder can be written as

$$\mathcal{L}(\boldsymbol{\Theta}) = \sum_{i=1}^{D} \Big[\mathcal{L}_{recon}(\boldsymbol{x}^{(i)}) + \mathcal{L}_{neg}(\boldsymbol{x}^{(i)}) + \big\| sg(\boldsymbol{\theta}_e^{(i)}) - \boldsymbol{\theta}_q^{(i)} \big\|_2^2 + \lambda \big\| sg(\boldsymbol{\theta}_q^{(i)}) - \boldsymbol{\theta}_e^{(i)} \big\|_2^2 \Big]$$

where $\boldsymbol{\Theta}$ means all parameters and D is the number of texts in a corpus. In order to minimize the distance between the embedding vector $\boldsymbol{\theta}_q^{(i)}$ and the encoder output $\boldsymbol{\theta}_e^{(i)}$, training objective includes the l_2 regularization between them. In detail, λ is a hyper parameter and $sg(\cdot)$ operator means the stop-gradient operation defined as

$$sg(x) = \begin{cases} x & \text{forward pass} \\ 0 & \text{backward pass} \end{cases}$$

that blocks gradients from flowing into its argument.

The above is the architecture of our proposed model NQTM and moreover, we name a simple variant of NQTM without the negative sampling error $\mathcal{L}_{neg}$ as **Quantization Topic Model (QTM)**. From the above description, our model NQTM differs from the VQ-VAE in two aspects. First, instead of a standard decoder, our model includes the new negative sampling decoder. Second, a novel topic distribution quantization method is proposed particularly for short texts to yield sharper distributions. These approaches are both to alleviate the data sparsity issue and we demonstrate the effectiveness of these two technical contributions in the next sections.

4 Experiments Setup

4.1 Datasets

Several real-world short text datasets are adopted in our experiment. The details are listed as

- **StackOverflow**[2] This dataset originates from

[2] https://github.com/jacoxu/StackOverflow

Datasets	# of docs	Average length	# of labels	Vocabulary size
StackOverflow	19,901	4.6	20	2,607
TagMyNews Title	31,223	5.2	7	6,391
Snippet	10,053	10.3	8	4,004
Yahoo Answer	19,027	4.1	10	3,243

Table 2: Statistics of datasets after preprocessing. Labels refer to the class labels of the corpus.

the challenge data published in Kaggle[3] . We use the dataset containing randomly selected 20,000 question titles provided by Xu et al. (2015). Each question title is annotated with an information technology name like "matlab", "osx" and "visual studio" as labels.

- **TagMyNews Title**[4] This dataset contains titles and contents of Engish news released by Vitale et al. (2012). We utilize the news titles as short texts in our experiment. Each news is assigned with a ground-truth label, e.g., "sci-tech", and "business", etc.

- **Snippet**[5] This dataset is provided by Phan et al. (2008) composed of the web content from Google search snippets. Eight labels are included in this dataset, such as "Culture-Arts-Entertainment" and "Computers", etc.

- **Yahoo Answer**[6] We obtained this dataset from Zhang et al. (2015) through the Yahoo Webscope program, including question titles, contents, and best answers. We adopt the question titles for topic modeling, totally containing ten labels.

To preprocess the raw content, we conduct the following steps: (1) tokenize each text and remove non-Latin characters and stop words by using NLTK[7] ; (2) filter out short texts with length less than 2; (3) remove words with document frequency less than 5; (4) convert all letters into lower cases. The statistics of each dataset after preprocessing are summarized in Table 2.

[3] https://www.kaggle.com/c/predict-closed-questions-on-stack-overflow/download/train.zip

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

[5] http://jwebpro.sourceforge.net/data-web-snippets.tar.gz

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

[7] https://nltk.org

4.2 Baseline Models

We take both conventional and neural topic models as baselines for comparison. For traditional topic models, we consider LDA (Blei et al., 2003), BTM[8] (Yan et al., 2013), DMM[9] (Yin and Wang, 2014), GPUDMM[10] (Li et al., 2016), and SeaNMF[11] (Shi et al., 2018). Note that SeaNMF is the state-of-the-art conventional model. In terms of neural topic models, we compare our model with NVDM[12] (Miao et al., 2016), GSM (Miao et al., 2017) and ProdLDA[13] (Srivastava and Sutton, 2017). Recently proposed supervised model TMN[14] (Zeng et al., 2018a) is also taken into consideration. We also compare our model with VQ-VAE to demonstrate the effectiveness of our proposed topic distribution quantization method.

5 Experimental Results

5.1 Topic Quality Evaluation

Topic Quality Metrics As mentioned before, the challenge of data sparsity in short texts results in two problems: generated topic words tend to be incoherent (trivial topics), and highly similar topics with repeated words are also yielded (repetitive topics). Therefore, we focus on the evaluation of topic quality referring to these two aspects, topic coherence and diversity. Topic coherence metrics depend on co-occurrences of topic words learned by models in the external corpus assuming that coherent words should co-occur within a certain distance. A new topic coherence metric C_V was introduced by Röder et al. (2015), which has been proven to perform better than other coherence metrics like widely-used NPMI (Bouma, 2009; Newman et al., 2010; Chang et al., 2009) and UMASS(Mimno et al., 2011). According to Krasnashchok and Jouili (2018), given a topic z and its top T words $(x_1, x_2, ..., x_T)$ sorted by the probability, the definition of C_V is

$$C_V(z) = \frac{1}{T}\sum_{i=1}^{T} s_{\cos}(\boldsymbol{v}_{\text{NPMI}}(x_i), \boldsymbol{v}_{\text{NPMI}}(\boldsymbol{x}_{1:T}))$$

where $s_{cos}(\cdot)$ means cosine similarity function and the vectors are defined as

[8] https://github.com/xiaohuiyan/BTM

[9] https://github.com/jackyin12/GSDMM

[10] https://github.com/NobodyWHU/GPUDMM

[11] https://github.com/tshi04/SeaNMF

[12] https://github.com/ysmiao/nvdm

[13] https://github.com/akashgit/autoencoding_vi_for_topic_models

[14] https://github.com/zengjichuan/TMN

Models	StackOverflow				TagMyNews Title				Snippet				Yahoo Answer			
	K=20		K=50		K=20		K=50		K=20		K=50		K=20		K=50	
Unsupervised	C_V	TU	C_V	TU	C_V	TU	C_V	TU	C_V	TU	C_V	TU	C_V	TU	C_V	TU
LDA	0.353	0.675	0.352	0.639	0.355	0.845	0.352	0.789	0.389	0.747	0.396	0.699	0.327	0.690	0.334	0.689
BTM	0.377	0.530	0.378	0.379	0.412	0.765	0.415	0.681	0.426	0.625	0.420	0.574	0.389	0.560	**0.392**	0.454
DMM	0.370	0.561	0.366	0.409	0.367	0.788	0.383	0.742	0.392	0.590	0.401	0.585	0.326	0.628	0.341	0.595
GPUDMM	0.372	0.568	0.362	0.496	0.378	0.798	0.391	0.744	0.405	0.604	0.409	0.600	0.332	0.633	0.351	0.626
SeaNMF	0.371	0.770	0.368	0.703	0.397	0.935	0.415	0.925	0.439	0.922	0.436	0.923	0.346	0.773	0.361	0.811
NVDM	0.386	0.982	0.376	0.905	0.458	0.995	0.421	0.964	0.434	0.986	0.391	0.937	0.387	0.988	0.370	0.915
GSM	0.365	0.658	0.356	0.482	0.357	0.807	0.351	0.612	0.399	0.781	0.399	0.649	0.325	0.668	0.321	0.470
ProdLDA	0.385	0.926	0.378	0.868	0.415	0.969	0.397	0.929	0.439	0.811	**0.440**	0.653	0.385	0.968	0.390	0.885
QTM	0.412	0.993	0.390	0.942	0.499	**1.000**	0.430	0.975	**0.442**	0.999	0.426	0.957	0.392	**0.997**	0.371	0.956
NQTM	**0.416**	**0.998**	**0.394**	**0.953**	**0.502**	**1.000**	**0.432**	**0.985**	**0.442**	**1.000**	0.431	**0.968**	**0.406**	**0.997**	0.373	**0.977**
Supervised																
TMN	0.423	0.397	0.420	0.269	0.464	0.453	0.428	0.347	0.465	0.613	0.427	0.516	0.343	0.527	0.322	0.220
VQ-VAE	0.457	0.303	0.363	0.435	0.477	0.693	0.483	0.444	0.419	0.737	0.407	0.447	0.382	0.423	0.383	0.463

Table 3: Topic coherence (C_V) and unique score (TU) of the top 15 words. K is the topic number. **QTM** means the variant of NQTM without negative sampling. The best in each unsupervised topic model is in bold.

$$\boldsymbol{v}_{\text{NPMI}}(x_i) = \{\text{NPMI}(x_i, x_j)\}_{j=1,\dots,T}$$
$$\boldsymbol{v}_{\text{NPMI}}(\boldsymbol{x}_{1:T}) = \left\{\sum_{i=1}^{T} \text{NPMI}\,(x_i, x_j)\right\}_{j=1,\dots,T}.$$

Then, the NPMI is calculated as

$$\text{NPMI}(x_i, x_j) = \frac{\log \frac{p(x_i,x_j)+\epsilon}{p(x_i)p(x_j)}}{-\log(p(x_i, x_j) + \epsilon)}$$

where $p(x_i)$ is the probability of x_i, $p(x_i, x_j)$ the coocurrance probability of x_i, x_j within a window in the reference corpus and ϵ is used to avoid zero. We use the public tool[15] to compute C_V provided by Röder et al. (2015).

Besides C_V score, we employ the topic unique metric (TU) (Nan et al., 2019) for topic diversity evaluation. For the top T words of topic z, it is defined as

$$TU(z) = \frac{1}{T}\sum_{i=1}^{T}\frac{1}{\text{cnt}(x_i)}$$

where $\text{cnt}(x_i)$ is the total number of times that word x_i appears in the top T words of all topics. Therefore TU score ranges from $1/K$ to 1 and a higher value means the generated topics are more diverse due to fewer duplicated words across other topics. It is crucial to note that in general, higher TU scores tend to cause lower C_V scores because coherent words seldom repeat, and higher C_V scores often lead to lower TU scores because coherent words frequently repeat across topics. We show our model can achieve significantly better performance on both aspects in the following.

Result Analysis Table 3 reports the topic coherence (C_V) and unique scores (TU) of the top 15 words under topic number $K = 20$ and 50. To be more specific, when topic number $K = 20$, NQTM can achieve significantly higher C_V scores, and we notice that TU scores of NQTM reach the highest on all datasets. When $K = 50$, NQTM still surpasses all unsupervised baselines on StackOverflow and TMN title in terms of both TU and C_V scores. Although C_V scores of ProdLDA and BTM are higher on Snippet and Yahoo Answer, TU scores of NQTM are much better than them. As mentioned earlier, the reason is that the C_V scores can be easily tricked by the repetitive topics composed of prominent words while with low topic diversity performance (further illustrated in Section 5.4). This issue is evenly severer for TMN. Notably, we can see TU scores of TMN are among the worst of all baselines, which is because the diversity of topics learned from TMN is not encouraged with the strong learning signal from the classification loss. Although some discovered topics seem coherent from the above baselines, unfortunately, many repetitive and less informative topics are ineffective in downstream applications; thus, their higher C_V scores are meaningless. On the contrary, we can observe the topic diversity performance of NQTM is

[15] https://github.com/dice-group/Palmetto

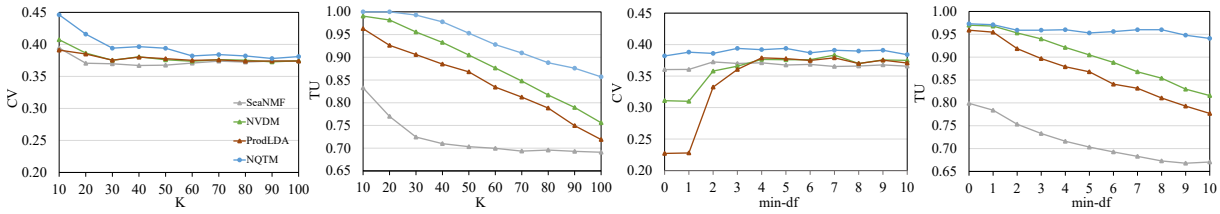


Figure 2: Topic coherence (C_V) and diversity (TU) performance with various topic numbers(K) (a, b) and minimum document frequencies (min-df) (c, d).

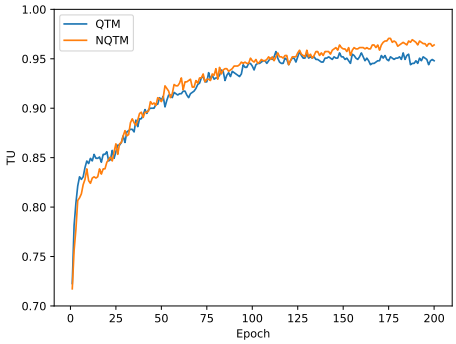


Figure 3: Change of TU scores along training epochs.

clearly superior with high coherence performance at the same time, which demonstrates the effectiveness of our model to alleviate the data sparsity problem.

5.2 Ablation Study

To conduct an ablation study, we also compare NQTM with VQ-VAE and QTM in Table 3. We can notice VQ-VAE sometimes has higher C_V scores, but as indicated in Section 5.1, it is useless because of its much lower TU scores. However, we can see QTM clearly has higher TU scores than VQ-VAE. This is because our new topic distribution quantization can separably distinguish topic distributions from different topics, while VQ-VAE cannot and leads massive texts under different topics to map to the same embedding vector. This contrast shows the effectiveness of our new topic distribution quantization method. Moreover, compared to QTM, we can see NQTM performs comparatively better on C_V scores and achieves obvious improvements on TU scores. This is because our negative sampling decoder provides extra learning signals to encourage topic-word distributions to differ from each other, bringing about better topic diversity performance. The change of TU scores of QTM and NQTM along training epochs is shown in Figure 3 that illustrates the TU score of NQTM gradually becomes higher than QTM during the training process. It is necessary to note that one advantage of QTM over NQTM is that QTM is faster on training since the negative sampling error is not required.

According to the above comparisons between VQ-VAE, QTM and NQTM, we can observe that our proposed new topic distribution quantization and negative sampling decoder are effective in improving the topic quality of short texts.

5.3 Data Sparsity Analysis

Since data sparsity is the essential challenge of short text topic modeling, to further demonstrate the advantages of our model, we explore the topic coherence and diversity performance under varying data sparsity degrees regarding two aspects, topic numbers (K) and minimum document frequencies (min-df) in preprocessing (see Section 4.1). Experimental results of NVDM, ProdLDA, SeaNMF are reported as these baselines perform relatively better in traditional and neural topic models respectively. Figures 2a and 2b show the C_V and TU scores of StackOverflow with topic number K ranging from 10 to 100. We can see although the TU scores of all models tend to decline due to the lack of word co-occurrences, NQTM decreases much slower than others by a large margin and also surpasses other baseline models in terms of C_V. Figures 2c and 2d present the C_V and TU scores of StackOverflow preprocessed by different min-df, from 0 to 10 under $K = 50$. Note that data sparsity becomes severer when preprocessing corpora with a bigger min-df. We can see that NQTM remains higher C_V scores than others and especially, TU scores of baselines fall sharply while NQTM still obviously keeps up.

Based on the above results under various data sparsity conditions, we can conclude that NQTM is grossly more robust in tackling the data sparsity challenge of short texts, which means NQTM can be better utilized in downstream applications.

Models	Topic Word Examples
DMM	able abort absolute abstract accept accepts able abort absolute abstract accept accepts wiki wikipedia encyclopedia film article movie movie movies film com imdb news reviews oscar academy movies movie picture winners
GPUDMM	qt library using matlab project use widget mac os qt osx windows application using oscar academy awards com movie winners award movie film com movies news reviews films movie movies imdb film title celebs encyclopedia
SeaNMF	cocoa window text menu button item focus application cocoa context without getting running oscar academy awards com winners award movie movie film com movies news reviews films movie movies imdb film title celebs encyclopedia
NVDM	featuring conducts homes hole creates aspects hand hear serve spanning compliance topix breakthrough continually rule progressive remedy ankle yet dry gum pink interview added lamp construct natural arrows width correct
ProdLDA	music romantic pop rock movie comedy movies music movie romantic pop movies comedy movie celebrity movies favorite youtube episode intel duo athlon core parallel processor memory intel processor memory cache ram pentium core
NQTM	mac os leopard snow installing osx installer qt widget signal slot signals creator slots cocoa interface builder events nsview app movie movies character actor scripts actors core intel processor pentium dual processors

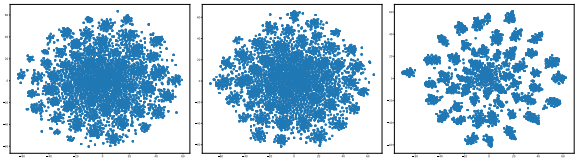
Table 4: Topic words examples under $K = 50$. Repetitive words are underlined.

5.4 Topic Examples Evaluation

To qualitatively illustrate the high-quality topics generated by our model, Table 4 presents the examples of topic words yielded by DMM, GPUDMM, SeaNMF, NVDM, ProdLDA, and NQTM in one experiment. We can observe that baseline models generate some repetitive topics with repeated words, such as "movie", "qt" and "processor", and although the topics of NVDM seem diverse, they're less informative. However, we can see that NQTM only generates a single coherent topic for each corresponding topic and the topic quality of NQTM is apparently higher.

5.5 Visualization of Latent Space

Figure 4 shows the t-SNE (van der Maaten and Hinton, 2008) visualization for topic distributions of texts under $K = 50$. It obviously illustrates that the points of NQTM are relatively more aggregated as



(a) NVDM (b) ProdLDA (c) **NQTM**

Figure 4: tSNE visualization of topic distributions.

groups and well separately scattered in the canvas, which is because NQTM can generate peakier topic distributions for short text topic modeling. The discretization and separation of the latent space can explain why NQTM is able to achieve higher topic coherence and diversity performance.

6 Conclusion

In this paper, for short text topic modeling, we propose the Negative sampling and Quantization Topic Model (NQTM) with a novel topic distribution quantization mechanism to yield peakier distributions and a new negative sampling decoder to enrich the learning signals. Experiments on benchmark datasets quantitatively and qualitatively show our model significantly outperforms baselines to overcome the data sparsity problem of short texts. Future works could focus on employing the proposed model in more downstream tasks.

Acknowledgement

We want to thank all anonymous reviewers for their helpful comments. This work is supported by China NSFC under Grant 61672309, Sichuan Science and Technology Program under Grant 2019YFSY0032, and MOST Fundamental Research Project under Grant 2017FY201407.

References

David M Blei, Alp Kucukelbir, and Jon D McAuliffe. 2017. Variational inference: A review for statisticians. *Journal of the American statistical Association*, 112(518):859–877.

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

Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of GSCL*, pages 31–40.

Jonathan Chang, Sean Gerrish, Chong Wang, Jordan L Boyd-Graber, and David M Blei. 2009. Reading tea leaves: How humans interpret topic models. In

Advances in neural information processing systems, pages 288–296.

Thomas Hofmann. 1999. Probabilistic latent semantic analysis. In *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence*, pages 289–296. Morgan Kaufmann Publishers Inc.

Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*.

Diederik P Kingma and Max Welling. 2014. Auto-encoding variational bayes. In *The International Conference on Learning Representations (ICLR)*.

Katsiaryna Krasnashchok and Salim Jouili. 2018. Improving Topic Quality by Promoting Named Entities in Topic Modeling. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, volume 2, pages 247–253.

Chenliang Li, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 2016. Topic modeling for short texts with auxiliary word embeddings. In *Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval*, pages 165–174. ACM.

Dawen Liang, Rahul G Krishnan, Matthew D Hoffman, and Tony Jebara. 2018. Variational autoencoders for collaborative filtering. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 689–698. International World Wide Web Conferences Steering Committee.

Zongyang Ma, Aixin Sun, Quan Yuan, and Gao Cong. 2012. Topic-driven reader comments summarization. In *Proceedings of the 21st ACM international conference on Information and knowledge management*, pages 265–274. ACM.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research*, 9(Nov):2579–2605.

Rishabh Mehrotra, Scott Sanner, Wray Buntine, and Lexing Xie. 2013. 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*, pages 889–892. ACM.

Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering discrete latent topics with neural variational inference. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2410–2419. JMLR. org.

Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *International conference on machine learning*, pages 1727–1736.

David Mimno, Hanna M Wallach, Edmund Talley, Miriam Leenders, and Andrew McCallum. 2011. 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.

Feng Nan, Ran Ding, Ramesh Nallapati, and Bing Xiang. 2019. Topic Modeling with Wasserstein Autoencoders. *arXiv preprint arXiv:1907.12374*.

David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. 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.

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

Aaron van den Oord and Oriol Vinyals. 2017. Neural discrete representation learning. In *Advances in Neural Information Processing Systems*, pages 6306–6315.

Xuan-Hieu Phan, Le-Minh Nguyen, and Susumu Horiguchi. 2008. Learning to classify short and sparse text & web with hidden topics from large-scale data collections. In *Proceedings of the 17th international conference on World Wide Web*, pages 91–100. ACM.

Xiaojun Quan, Chunyu Kit, Yong Ge, and Sinno Jialin Pan. 2015. Short and sparse text topic modeling via self-aggregation. In *Twenty-fourth international joint conference on artificial intelligence*.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. *In Proceedings ofthe 31th International Conference on Machine Learning*.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the eighth ACM international conference on Web search and data mining*, pages 399–408. ACM.

Kugatsu Sadamitsu, Takuya Mishina, and Mikio Yamamoto. 2007. Topic-based language models using dirichlet mixtures. *Systems and Computers in Japan*, 38(12):76–85.

Tian Shi, Kyeongpil Kang, Jaegul Choo, and Chandan K Reddy. 2018. Short-text topic modeling via non-negative matrix factorization enriched with local word-context correlations. In *Proceedings of the 2018 World Wide Web Conference*, pages 1105–1114. International World Wide Web Conferences Steering Committee.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *ICLR*.

Mark Steyvers and Tom Griffiths. 2007. Probabilistic topic models. *Handbook of latent semantic analysis*, 427(7):424–440.

Daniele Vitale, Paolo Ferragina, and Ugo Scaiella. 2012. Classification of short texts by deploying topical annotations. In *European Conference on Information Retrieval*, pages 376–387. Springer.

Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. 2010. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of the third ACM international conference on Web search and data mining*, pages 261–270. ACM.

Jason Weston, Sumit Chopra, and Antoine Bordes. 2014. Memory networks. *arXiv preprint arXiv:1410.3916*.

Xiaobao Wu and Chunping Li. 2019. Short Text Topic Modeling with Flexible Word Patterns. In *International Joint Conference on Neural Networks*.

Xiaobao Wu, Chunping Li, Yan Zhu, and Yishu Miao. 2020. Learning Multilingual Topics with Neural Variational Inference. In *International Conference on Natural Language Processing and Chinese Computing*.

Jiaming Xu, Peng Wang, Guanhua Tian, Bo Xu, Jun Zhao, Fangyuan Wang, and Hongwei Hao. 2015. Short Text Clustering via Convolutional Neural Networks. In *VS@ HLT-NAACL*, pages 62–69.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1445–1456. ACM.

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

Jichuan Zeng, Jing Li, Yan Song, Cuiyun Gao, Michael R Lyu, and Irwin King. 2018a. Topic memory networks for short text classification. In *Proceedings of the conference on empirical methods in natural language process*.

Xingshan Zeng, Jing Li, Lu Wang, Nicholas Beauchamp, Sarah Shugars, and Kam-Fai Wong. 2018b. Microblog Conversation Recommendation via Joint Modeling of Topics and Discourse. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*, volume 1, pages 375–385.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. In *Advances in neural information processing systems*, pages 649–657.

Yuan Zuo, Junjie Wu, Hui Zhang, Hao Lin, Fei Wang, Ke Xu, and Hui Xiong. 2016. Topic modeling of short texts: A pseudo-document view. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 2105–2114. ACM.