Information Processing and Management 56 (2019) 102098

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

ATM: Adversarial-neural Topic Model

Rui Wang[a], Deyu Zhou[*,a], Yulan He[b]

[a] School of Computer Science and Engineering, Southeast University, Nanjing, Jiangsu Province, 210096, China
[b] Department of Computer Science, University of Warwick, Coventry CV4 7AL, UK

ARTICLE INFO

Keywords:
Generative adversarial net
Neural-based topic model
Open domain event extraction
Topic modeling

ABSTRACT

Topic models are widely used for thematic structure discovery in text. But traditional topic models often require dedicated inference procedures for specific tasks at hand. Also, they are not designed to generate word-level semantic representations. To address the limitations, we propose a neural topic modeling approach based on the Generative Adversarial Nets (GANs), called Adversarial-neural Topic Model (ATM) in this paper. To our best knowledge, this work is the first attempt to use adversarial training for topic modeling. The proposed ATM models topics with dirichlet prior and employs a generator network to capture the semantic patterns among latent topics. Meanwhile, the generator could also produce word-level semantic representations. Besides, to illustrate the feasibility of porting ATM to tasks other than topic modeling, we apply ATM for open domain event extraction. To validate the effectiveness of the proposed ATM, two topic modeling benchmark corpora and an event dataset are employed in the experiments. Our experimental results on benchmark corpora show that ATM generates more coherence topics (considering five topic coherence measures), outperforming a number of competitive baselines. Moreover, the experiments on event dataset also validate that the proposed approach is able to extract meaningful events from news articles.

1. Introduction

Topic models (Blei, 2012; Chen, 2017) underpin many successful applications within the field of Natural Language Processing (NLP). Variants of topic models have been proposed for different tasks including content analysis of e-petitions(Hagen, 2018), topic-associated sentiment analysis (Lin & He, 2009), event extraction from social media (Zhou, Chen, & He, 2014; Zhou, Chen, Zhang, & He, 2017; Zhou, Gao, & He, 2016) and product aspect mining (Xiao, Ji, Li, Zhuang, & Shi, 2018). However, topic models typically rely on mean-field variational inference(Asuncion, Welling, Smyth, & Teh, 2009) or collapsed Gibbs sampling for model learning. A small change to the modeling assumption requires the re-derivation of the whole inference algorithm, which is mathematically arduous and time consuming.

In recent years, word embeddings (such as Word2vec(Le & Mikolov, 2014), GloVe(Pennington, Socher, & Manning, 2014), fastText(Bojanowski, Grave, Joulin, & Mikolov, 2017; Grave, Mikolov, Joulin, & Bojanowski, 2017) and probabilistic fastText (Athiwaratkun, Wilson, & Anandkumar, 2018)) have gained an increasing interest thanks to their improved efficiency in representing words as continuous vectors in a low-dimensional space. The resulting embeddings encode numerous semantic relations (similarity or analogies) and are helpful for NLP tasks(Fernández-Reyes, Valadez, & Montes-y-Gómez, 2018; Hsu, Lee, Chang, & Sung, 2018). But the traditional topic models could not generate such word-level semantic representations.

To overcome the limitation that traditional topic model often need sophisticated inference algorithm, Neural Variational

* Corresponding author.
E-mail addresses: rui_wang@seu.edu.cn (R. Wang), d.zhou@seu.edu.cn (D. Zhou), Yulan.He@warwick.ac.uk (Y. He).

https://doi.org/10.1016/j.ipm.2019.102098
Received 15 January 2019; Received in revised form 6 July 2019; Accepted 7 August 2019
Available online 17 August 2019
0306-4573/

Document Model (NVDM) (Miao, Yu, & Blunsom, 2016) was devised based on the Variational Auto-Encoder (VAE)(Kingma & Welling, 2014) and used a hidden layer to reconstruct the document by generating the words independently. However, the usage of gaussian prior over topics in NVDM may lead to incoherent and similar topics being generated. On the contrary, Srivastava and Sutton (2017) proposed LDA-VAE, a neural topic model based on the VAE, in which the logistic normal distribution was employed as the prior over topics for topic generation. To further enhance the quality of the generated topic, Srivastava replaced the mixture assumption with a weighted product of experts at the word-level and proposed the ProdLDA. But both the LDA-VAE and the ProdLDA were not able to produce word-level semantic representations. Besides, the logistic normal prior used in LDA-VAE and ProdLDA also could not capture the multiplicity topical aspects in a document and result in generating bad topics.

To overcome the limitations that the traditional topic models often need sophisticated inference algorithm and the exist neural based topic models could not generate coherent topic words. In this paper, we propose the Adversarial-neural Topic Model (ATM) based on adversarial training. The principle idea is to use a generator network to learn the projection function between the document-topic distribution and the document-word distribution. Instead of providing an analytic approximation, as in traditional topic models, the ATM uses a discriminator network to recognize if the input document is real or fake and its output signal could help the generator to construct a more realistic document from a random noise drawn from a dirichlet distribution. Due to the flexibility of neural networks, the generator is capable of learning complicated non-linear distributions. And the supervision provided by the discriminator in the adversarial training phase will help the generator to capture the semantic patterns embedded in the latent topics. Besides, the connection weights between the embedding layer and the word distribution layer of the generator also encodes the semantic information and naturally provides distributed representations of words as side product.

The objectives of our work in this paper are, more succinctly, as follows:

1. Traditional topic models based on gibbs sampling or variational inference often need sophisticated inference algorithms and obtain incoherent topics. We are interested in devising a novel neural-based topic model which could mine coherent topics from text corpora automatically in an unsupervised manner. To this end, based on the Generative Adversarial Net, we propose the ATM model which could extract the coherent topics among text corpus.
2. From a practical perspective, we would like to devise a neural-based topic model which could be transplanted to other task easily with limited modification. For this purpose, we modify the topic generation process of the proposed ATM and employ it for open domain event extraction task (Zhou, Xu, & He, 2015), experiments on news articles corpus shows that the proposed model is able to extract meaningful events and also verifies the portability of ATM.

The practical significance of this work is that the proposed approach (ATM) could generate more coherent topics than the state-of-the-art topic modeling approaches. Meanwhile, it could also produce semantic representations for each word in the vocabulary as side product, which is currently not supported by the compared models. Besides, the proposed ATM could be easily ported to other NLP task (such as open domain event extraction) with limited modification. The rest of the paper is organized as follows. Section 2 reviews the related literature on neural topic models and generative adversarial nets. In Section 3, we provide the details of the proposed Adversarial-neural Topic Model. Section 4 will introduce our evaluation corpora and our obtained experimental results. Finally, the paper is concluded in Section 5 with suggestions for further work.

2. Related work

Our work is related to two lines of research, neural-based topic modeling and the Generative Adversarial Nets. Thus, we will next briefly introduce the related work in two domain separately.

2.1. Neural-based topic modeling

To overcome the difficult exact inference of topic models based on directed graph, Hinton and Salakhutdinov (2009) modified the Restricted Boltzmann Machines and proposed a replicated softmax model (called RSM). Inspired by the variational autoencoder, Miao et al. (2016) used the multivariate gaussian as the prior distribution of latent space and proposed the Neural Variational Document Model (NVDM) for text modeling. More recently, to deal with the inappropriate gaussian prior of topic distributions in NVDM, Srivastava and Sutton (2017) proposed the LDA-VAE which approximated the dirichlet prior using a logistic normal distribution, and the usage of logistic normal prior could help to generate more coherent and diverse topics. Srivastava and Sutton (2017) replaced the mixture assumption with a weighted product of experts at the word-level and proposed the ProdLDA which further improved topic coherence.

2.2. Generative adversarial nets

As a neural-based generative model, the Generative Adversarial Nets (Goodfellow et al., 2014) have been extensively researched from both theoretical and practical aspects.

Theoretically, Nowozin, Cseke, and Tomioka (2016) used the Fenchel conjugate to define the F-divergence and proposed the F-GAN to generalize its optimization objective. To precisely measure the distance between two high dimensional distributions, Arjovsky, Chintala, and Bottou (2017) defined the Earth Mover's Distance (Wasserstein distance) and gave a computational method based on the weight clipping mechanism. Along this line, Gulrajani, Ahmed, Arjovsky, Dumoulin, and Courville (2017) improved the

Wasserstein GAN by adding a gradient penalty loss and promoted the stability of adversarial training.

In practical applications, GAN-based models have been extensively researched in computer vision community, especially in image generation scenario. To incorporate the conditional information, Mirza (Isola, Zhu, Zhou, & Efros, 2017) employed the random noise together with label as input and proposed the Conditional-GAN to generate image under the supervision of the annotated label. The deep convolutional neural network were employed as the generator and the discriminator in Radford, Metz, and Chintala (2016) to improve the quality of generated image. And (Ledig et al., 2017) also used the GAN-basd approch to generate super-resolution image. On the other hand, many variants of GAN have been developed for NLP tasks. Such as text generation, a hot research area in NLP. The sequence generative adversarial network (SeqGAN) proposed in Yu, Zhang, Wang, and Yu (2017) incorporated a policy gradient strategy to optimize the generation process. Based on the policy gradient, Lin, Li, He, Zhang, and Sun (2017) proposed the RankGAN to capture the rich structures of language by ranking and analysing a collection of human-written and machine-written sentences. To overcome the mode collapse when dealing with discrete data, Fedus, Goodfellow, and Dai (2018) proposed the MaskGAN which used an actor-critic conditional GAN to fill in missing text conditioned on the surrounding context. Along this line, Wang and Wan (2018) employed multiple generator network (each for one sentiment) and proposed the SentiGAN to generate texts of different sentiment labels. Hu, Yang, Liang, Salakhutdinov, and Xing (2017) incorporated the VAE into GAN framework for text generation. Besides, Miyato, Dai, and Goodfellow (2017) and Li and Ye (2018) improved the performance of semi-supervised text classification using adversarial training. Zeng, Dai, Li, Sherratt, and Wang (2018) designed GAN-based models for distance supervision relation extraction. Wang and Lee (2018) incorporated the generative adversarial net into a encoder-decoder framework and proposed a GAN based model for text summarization. Yang, Hu, Dyer, Xing, and Berg-Kirkpatrick (2018) employed the target domain language model into GAN framework to transfer style of text.

Despite many successful applications using GAN-based approaches, none of these approaches tackles the topic modeling problem. We propose the first GAN-based topic model called ATM, which differs from the existing approaches to neural topic modeling in the following aspects: (1) Unlike the NVDM and the LDA-VAE which use either multivariate gaussian prior or logistic-normal prior for latent topics, ATM uses the dirichlet prior instead. It makes sure that ATM could provide K-dimensional noise and each capture certain semantic patterns in the text corpus; (2) Unlike most GAN-based text generation approaches, a generator network is employed by ATM to learn the projection function between the document-topic distribution and the document-word distribution, which essentially captures the semantic patterns among latent topics rather than generating text sequences; (3) Unlike the traditional topic model, ATM is able to generate meaningful word-level semantic representations as a side product.

3. Adversarial-neural Topic Model

We propose the Advesarial-neural Topic Model (ATM) as shown in Fig. 1. The proposed ATM contains three main components: (1) the document sampling module shown at the top of Fig. 1, which defines the representation mapping function and samples a real document $d_r \in \mathbb{R}^V$ from an input text corpus; (2) the generator G takes a topic distribution $\vec{\theta}$ sampled from a dirichlet prior as input and generates the corresponding fake document d_f; (3) the discriminator D takes d_f and d_r as input and discriminates the fake document from the real ones, whose output is subsequently used as a learning signal to update the parameters of G and D. We explain the design and function of each of these modules in more details below.

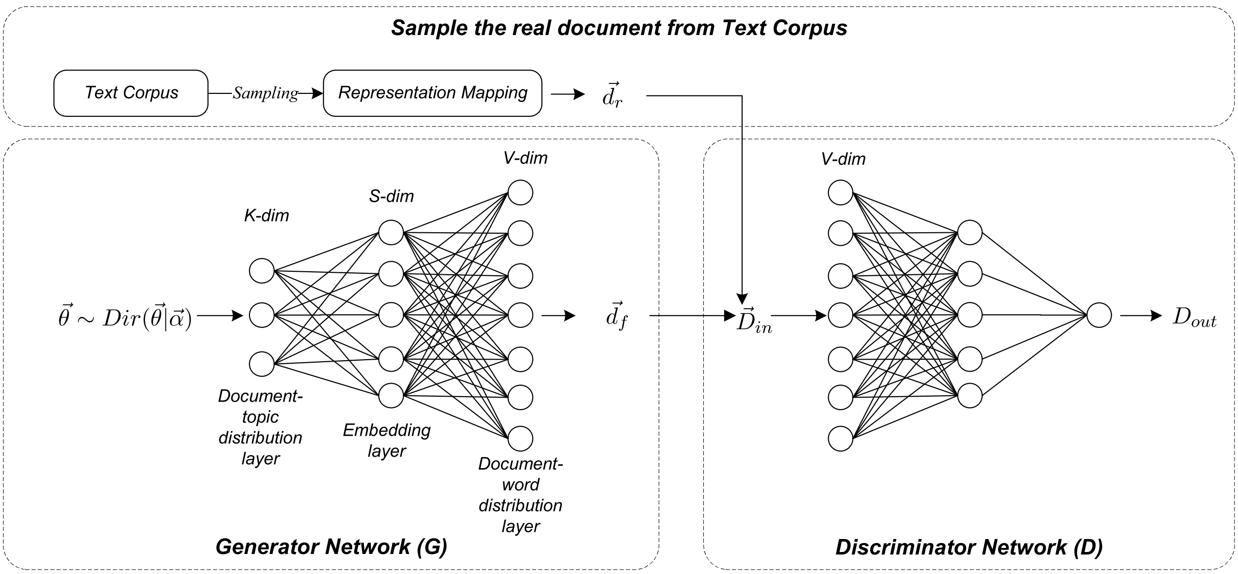

Fig. 1. The framework of the Adversarial-neural Topic Model (ATM).

3.1. Representation mapping

Each document d is represented by a normalized V-dimensional vector weighted by TF-IDF. More concretely:

$$\begin{aligned} tf_{i,d} &= \frac{n_{i,d}}{\sum_v n_{v,d}} \\ idf_i &= \log\frac{|C|}{|C_i|} \\ tf\text{-}idf_{i,d} &= tf_{i,d} \times idf_i \\ d_r^i &= \frac{tf\text{-}idf_{i,d}}{\sum_v tf\text{-}idf_{v,d}} \end{aligned}$$

where V is the vocabulary size, $n_{i,d}$ denotes the number of times the ith word appears in document d, $|C|$ denotes the total number of documents in the corpus, and $|C_i|$ is the number of documents containing the ith word. With this representation, each document in the corpus could be regarded as a multinomial distribution over V words, and each dimension reflects the semantic coherence between the ith word and the document d.

3.2. Network architecture

The G network contains three layers, the K-dimensional document-topic distribution layer, the S-dimensional embedding layer and the V-dimensional document-word distribution layer as shown in Fig. 1. First, the G network takes a randomly sampled topic distribution $\vec{\theta}$ as input and transforms it into a document-word distribution. To model the multinomial property of the document-topic distribution, $\vec{\theta}$ is drawn from $Dir(\vec{\theta}|\vec{\alpha})$:

$$p(\vec{\theta}|\vec{\alpha})=Dir(\vec{\theta}|\vec{\alpha}) \triangleq \frac{1}{\triangle(\vec{\alpha})}\prod_{k=1}^{K}\theta_k^{\alpha_k-1} \tag{1}$$

where $\vec{\alpha}$ is the hyper-parameter of the dirichlet distribution, $\triangle(\vec{\alpha}) = \frac{\prod_{k=1}^{K}\Gamma(\alpha_k)}{\Gamma(\sum_{k=1}^{K}\alpha_k)}$, K is the number of topics, $\theta_k \in [0, 1]$ denotes the proportion of topic k in the document and $\sum_{k=1}^{K}\theta_k = 1$.

Then, G projects $\vec{\theta}$ into the S-dimensional (set to 100 in experiments) semantic space through the embedding layer based on equations :

$$\vec{a_s} = \max((W_s\vec{\theta} + \vec{b_s}), leak*(W_s\vec{\theta} + \vec{b_s})) \tag{2}$$

$$\vec{o_s} = BN(\vec{a_s}) \tag{3}$$

where $W_s \in \mathbb{R}^{S\times K}$ is the weight matrix and $\vec{b_s}$ represents the bias term of the embedding layer, $\vec{a_s}$ is the state vector activated by the LeakyReLU function parameterized with $leak$, BN denotes batch normalization and $\vec{o_s}$ is the output of the embedding layer.

Finally, G transforms $\vec{o_s}$ to a V-dimensional multinomial distribution d_f using :

$$\vec{h_w} = W_w\vec{o_s} + \vec{b_w} \tag{4}$$

$$o_w^i = \frac{\exp(h_w^i)}{\sum_{v=1}^{V}\exp(h_w^v)} \tag{5}$$

where $W_w \in \mathbb{R}^{V\times S}$ learns the semantic word embeddings and $\vec{b_w}$ represents the bias term, $\vec{h_w}$ is the state vector and o_w^i denotes the probability of ith word in d_f.

Likewise, we design the discriminator as a three layer fully connected network. The D network employs the d_f and the d_r as input and outputs a scalar as shown in Fig. 1. A higher D_{out} means that the discriminator is prone to consider the input data as a real document and vice versa.

3.3. Training

The fake document d_f and the real document d_r shown in Fig. 1 could be viewed as the random sample from two V-dimensional dirichlet distribution $\mathbb{P}_g$ and $\mathbb{P}_r$. And the training objective of ATM is to let the generated distribution $\mathbb{P}_g$ approximate the real data distribution $\mathbb{P}_r$ as much as possible. Thus, the choice of divergence that measures the distance between two distributions is crucial for effective training of ATM.

The original GAN (Goodfellow et al., 2014) used the Jensen–Shannon divergence as the optimization objective. However, Arjovsky et al. (2017) argued that the divergences which GANs typically minimize are potentially not continuous with respect to the generator's parameters, leading to mode collapse and training difficulty. They proposed instead using the Earth-Mover's distance (also called Wasserstein-1) which is defined as the minimum cost of transporting mass in order to transform the distribution $\mathbb{P}_g$ into

Input: $K, \lambda, n_d, m, \alpha_1, \beta_1, \beta_2$
Output: the trained generator network G.

```
Initial D parameters ω_d and G parameter ω_g
while ω_g has not converged do
    for t = 1, ..., n_d do
        for j = 1, ..., m do
            Sample d_r ~ ℙ_r,
            Sample a random θ⃗ ~ Dir(θ⃗|α⃗)
            Sample a random number ϵ ~ U[0, 1]
            d_f ← G(θ⃗)
            d̂ ← ϵ d_r + (1 − ϵ) d_f
            L_d^(j) = D(d_f) − D(d_r)
            L_gp^(j) = (|| ∇_d̂ D(d̂) || − 1)^2
            L^(j) ← L_d^(j) + λ L_gp^(j)
        end for
        ω_d ← Adam(∇_{ω_d} (1/m) Σ_{j=1}^{m} L^(j), ω_d, p_a)
    end for
    Sample m noise {θ⃗^(j) ~ Dir(θ⃗|α⃗)}
    ω_g ← Adam(∇_{ω_g} (−1/m) Σ_{j=1}^{m} D(G(θ⃗^(j))), ω_g, p_a)
end while
```

Algorithm 1. Training procedure for ATM.

the distribution $\mathbb{P}_r$. Further, Gulrajani et al. (2017) improved the Wassertein-1 with a gradient penalty strategy which performed more stable. We follow their work and define the objective of ATM as:

$$L_d = \mathop{\mathbb{E}}_{d_f \sim \mathbb{P}_g}[D(d_f)] - \mathop{\mathbb{E}}_{d_r \sim \mathbb{P}_r}[D(d_r)] \tag{6}$$

$$L_{gp} = \mathop{\mathbb{E}}_{\hat{d} \sim \mathbb{P}_{\hat{d}}}[(\|\nabla_{\hat{d}} D(\hat{d})\|_2 - 1)^2] \tag{7}$$

$$L = L_d + \lambda L_{gp} \tag{8}$$

where L_d and L_{gp} denote the loss of discriminator D and the gradient penalty, respectively, λ is the gradient penalty coefficient, $\hat{d}$ could be obtained by sampling uniformly along a straight line between a real document d_r and a generated document d_f, and $\mathbb{P}_{\hat{d}}$ is the distribution from which $\hat{d}$ is sampled.

In each training step, the same number of d_r and d_f samples are fed into the Discriminator and the distance between $\mathbb{P}_g$ and $\mathbb{P}_r$ is estimated using Eqs. (6)–(8). Thus, G and D networks could be updated to minimize the distance between $\mathbb{P}_g$ and $\mathbb{P}_r$. Based on the model structure and the optimization objective described above, the training procedure for ATM is given in Algorithm 1. Here, n_d denotes the number of discriminator iterations per generator iteration, m represents the batch size, α_1 is the learning rate, β_1 and β_2 are other hyper-parameters of Adam optimizer (Kingma & Ba, 2015), and p_a denotes $\{\alpha_1, \beta_1, \beta_2\}$. We use the default values of $\lambda = 10$, $n_d = 5$, $m = 512$. Moreover, the α_1, β_1 and β_2 are set to 0.0001, 0 and 0.9 respectively.

3.4. Topic generation

The trained generator G learns the projection function between the document-topic distribution and the document-word distribution. That is, given a topic distribution $\vec{\theta}_d$ for a document d, G is able to generate the corresponding word distribution.

To generate the word distribution of each topic, we use $\vec{t}s_{(k)}$, a K-dimensional vector, as the one-hot encoding of the kth topic. For example, $\vec{t}s_{(1)} = [1, 0, 0, 0, 0]^{\mathsf{T}}$ in the five topic number setting. We could then obtain the word distribution $\vec{\phi}_k$ for topic k using:

$$\vec{\phi}_k = G(\vec{t}s_{(k)}) \tag{9}$$

4. Experiments

We evaluate our proposed ATM on two tasks, topic extraction and open domain event extraction. We first describe the datasets and the baseline approaches, and then present the topic coherence evaluation results for the topic extraction task. Finally, we discuss the results of using ATM for open domain event extraction to validate the feasibility of applying ATM for tasks other than topic modeling.

4.1. Experimental setup

Two publicly accessible datasets, Grolier[1] and NYtimes[2] datasets, are used for topic coherence evaluation, and an event dataset built based on the Global Database of Events, Language, and Tone (GDELT)[3] is used for event extraction. Details are summarized below:

- *Grolier dataset*[1] is built from Grolier Multimedia Encyclopedia, and its content covers almost all the fields in the world, such as sports, economics, politics and etc. It contains 29,762 documents and is a benchmark text corpora in topic modeling.
- *NYtimes dataset*[2] is a collection of newswire articles written and published by New York Times between January 1, 1987 and June 19, 2007 with article metadata provided by the New York Times Newsroom. This corpus also has a wide range of topics in real world, such as politics and entertainment.
- *Event dataset*. This dataset is the subset of GDELT which is released by Google. we crawl the Database[3] and built the event dataset by selecting the articles published on the first day of May in 2014. It contains many real events occurred at that day, such as MH370 and Indian Election.

We choose the following five models as the baselines:

- **LDA** (Blei, Ng, & Jordan, 2003), is a topic model that generates topics based on word the co-occurrence patterns from documents. With the usage of dirichlet prior topic distribution and word distribution, LDA could capture the multiplicity topic aspects from document collections in an unsupervised manner. We implement the LDA model and set the dirichlet prior of the document-topic distribution $\alpha = 50/K$ and the dirichlet prior of the topic-word distributions $\beta = 0.01$, following what have been suggested in Griffiths and Steyvers (2004).
- **NVDM** (Miao et al., 2016), is an neural based approach which models topics using variational auto-encoder. In NVDM, multi-variate gaussian distribution is used as prior distribution of the latent space, and it is trained under the supervision of evidence lower bound (ELBO). We use the original implementation.[4]
- **LDA-VAE** (Srivastava & Sutton, 2017), is a neural topic model based on variational auto-encoder. To obtain readable topics, LDA-VAE substitute multivariate gaussian with a logistic normal distribution as the prior of the latent space. In this paper, the original implementation[5] of LDA-VAE is employed to obtain the compared results.
- **ProdLDA** (Srivastava & Sutton, 2017), is a variant of LDA-VAE which also uses logistic normal as the prior of the latent space. Beside, it assumes that the distribution over individual words is a product of experts rather than the mixture model used in LDA. The original implementation is used in this paper.
- **LEM** (Zhou et al., 2014), is a Bayesian modeling approach for open domain event extraction. It treats an event as a latent variable and models the generation of an event as a joint distribution of its individual event elements (organization, location, person, keyword).[6] We implement the algorithm with the default configuration.

For the NYtimes dataset, we random select 100,000 articles and remove the low frequent words. For the Event dataset, we use the Stanford Named Entity Recognizer[7] (Finkel, Grenager, & Manning, 2005) for identifying the named entities (Location, Organization and Person). In addition, we remove common stopwords and only keep the recognized name entities and the tokens which are verbs, nouns, or adjectives from these event documents. The statistics of the processed corpora are shown in Table 1.

4.2. Topic coherence evaluation

Typically topic models are evaluated based on the likelihood of held-out documents. However, as pointed out in Chang, Gerrish, Wang, Boyd-Graber, and Blei (2009), higher likelihood of held-out document does not necessarily correspond to human judgement of topic coherence. In this subsection, we follow (Röder, Both, & Hinneburg, 2015) and choose five coherence metrics to evaluate the topics generated by models. They are C_P (a metric based on a sliding window, a one-preceding segmentation of the given words and the confirmation measure of Fitelson's coherence), C_A (a metric based on a context window, a pairwise comparison of the given words and an indirect confirmation measure that uses normalized pointwise mutual information and the cosine similarity), UCI (a metric based on a sliding window and the pointwise mutual information of all word pairs of the given topics), NPMI (an enhanced version of UCI using the normalized pointwise mutual information) and UMass (Mimno, Wallach, Talley, Leenders, & McCallum, 2011) (a metric based on document cooccurrence counts, a one-preceding segmentation and a logarithmic conditional probability as confirmation measure). For all these five metrics, higher value implies more coherent topic. In our evaluation, we choose the top 10

[1] https://cs.nyu.edu/~roweis/data/ .
[2] http://archive.ics.uci.edu/ml/datasets/Bag + of + Words .
[3] http://data.gdeltproject.org/events/index.html
[4] https://github.com/ysmiao/nvdm
[5] https://github.com/akashgit/autoencoding_vi_for_topic_models
[6] means organization, location, person and keywords.
[7] https://nlp.stanford.edu/software/CRF-NER.html

Table 1
The statistics of datasets.

Dataset	#Document	#Words
Grolier	29,762	15,276
NYtimes	99,992	12,604
Event	20,199	9,346

Table 2
Average topic coherence on Grolier and NYtimes corpus with five topic settings [20, 30, 50, 75, 100].

Dataset	Model	C_P	C_A	NPMI	UCI	UMass
	NVDM	−0.187746	0.145684	−0.061911	−2.114927	−4.291624
	LDA-VAE	−0.220548	0.150469	−0.065378	−2.479750	−4.755522
Grolier	ProdLDA	−0.037436	0.173391	−0.019347	−1.639878	−4.542689
	LDA	0.190845	0.200942	0.049753	−0.050336	−2.918612
	ATM	**0.210448**	**0.218898**	**0.058167**	**0.105086**	**−2.765081**
	NVDM	−0.413086	0.134154	−0.143711	−4.307269	−5.931614
	LDA-VAE	−0.157560	0.148221	−0.061418	−2.420816	−4.640276
NYtimes	ProdLDA	−0.003455	0.196395	−0.028223	−1.917367	−4.193377
	LDA	0.308336	0.212750	0.077278	0.516503	−2.420221
	ATM	**0.356771**	**0.237524**	**0.089874**	**0.658218**	**−2.324093**

words to represent each topic and compute the topic coherence using the Palmetto library.[8]

To compare the performance of the proposed approach, experiments are conducted on Grolier and NYtimes with five topic number settings [20, 30, 50, 75, 100]. The average coherence values are listed in Table 2 and each value is computed by averaging the average topic coherences (all the topics are used) over five topic number settings. Besides, we calculate the average topic coherence among topics whose coherence values are ranked at the top 50%, 70%, 90%, 100% positions. For example, to calculate the average UCI coherence of ATM @ 70%, we first compute the average UCI coherence with the select topics whose UCI values are ranked at the top 70% positions for each topic number setting, and then average the five averaged coherence values. The corresponding results are shown in Fig. 2. It can be observed from Fig. 2 that the proposed model outperforms the LDA, NVDM, LDA-VAE and ProdLDA in general. This maybe caused by following factors: i) ATM models the multinomial distribution over topics using a Dirichlet prior, which is more proper than the Gaussian prior (used in NVDM) and logistic normal prior (used in LDA-VAE and ProdLDA). The usage of the Dirichlet prior in ATM make it could capture the multiplicity topic aspects from document collection and further obtain more coherent topics. 2.) The strong representation ability of the neural network makes the ATM could fit the true data distribution better than the traditional topic model and generate more coherent topics.

To explore how topic coherence results vary with different topic numbers, we show in Fig. 3 the average topic coherence of two datasets vs. different topic number settings. We can observe that ATM achieves better results compared to other baselines most of the time with 20, 30, 50 or 75 topics. However, when the topic number is 100, the performance gap between ATM and LDA diminishes and in some cases (e.g., C_P and C_A for the Grolier dataset), ATM gives slightly worse results compared to LDA, though it still largely outperforms all the other baselines. This might attribute to the increased network complexity due to the larger topic number setting.

From the above topic coherence evaluation results, it is clear that ATM is able to extract more coherence topics compared to baselines. To verify this qualitatively, we show examples of topics from all the models in Table 3. These topics correspond to *'airline', 'trade', 'music', 'election'* and *'film'* respectively. Words that do not seem to belong to its corresponding topic are highlighted in italic. It can be observed that the number of less semantically relevant words somewhat correlates with the coherence results observed earlier in Table 2 and Fig. 2.

Unlike traditional topic models, the proposed ATM could learn the semantic embeddings of words apart from generating coherent topics. The weights matrix $W_w \in \mathbb{R}^{V \times S}$ contains the word-level semantic information, and each row could be viewed as the corresponding word embedding. Thus, we select the topic words of six topics from a 50-topic run on the NYtimes corpus and use the Principal Component Analysis (PCA) to project their word embeddings into a two-dimensional space. The visualization of these topic words is shown as Fig. 4. We can clearly see that the words related to the *'trade'* topic are grouped at the lower right corner, and the topic words of *'religious'* are displayed at the top region. Besides, the words related to the topics *'music'* and *'film'* are close to each other, which is not surprising, since these topics are closely related.

4.3. Open domain event extraction

To further prove the feasibility of porting ATM to tasks other than topic modeling, we apply it for open domain event extraction. For this task, an event is represented in a structured form as $<$ *org, loc, per, key* $>$ [6] (Zhou et al., 2015), with each of the elements in the quadruples represented by a list of words.

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

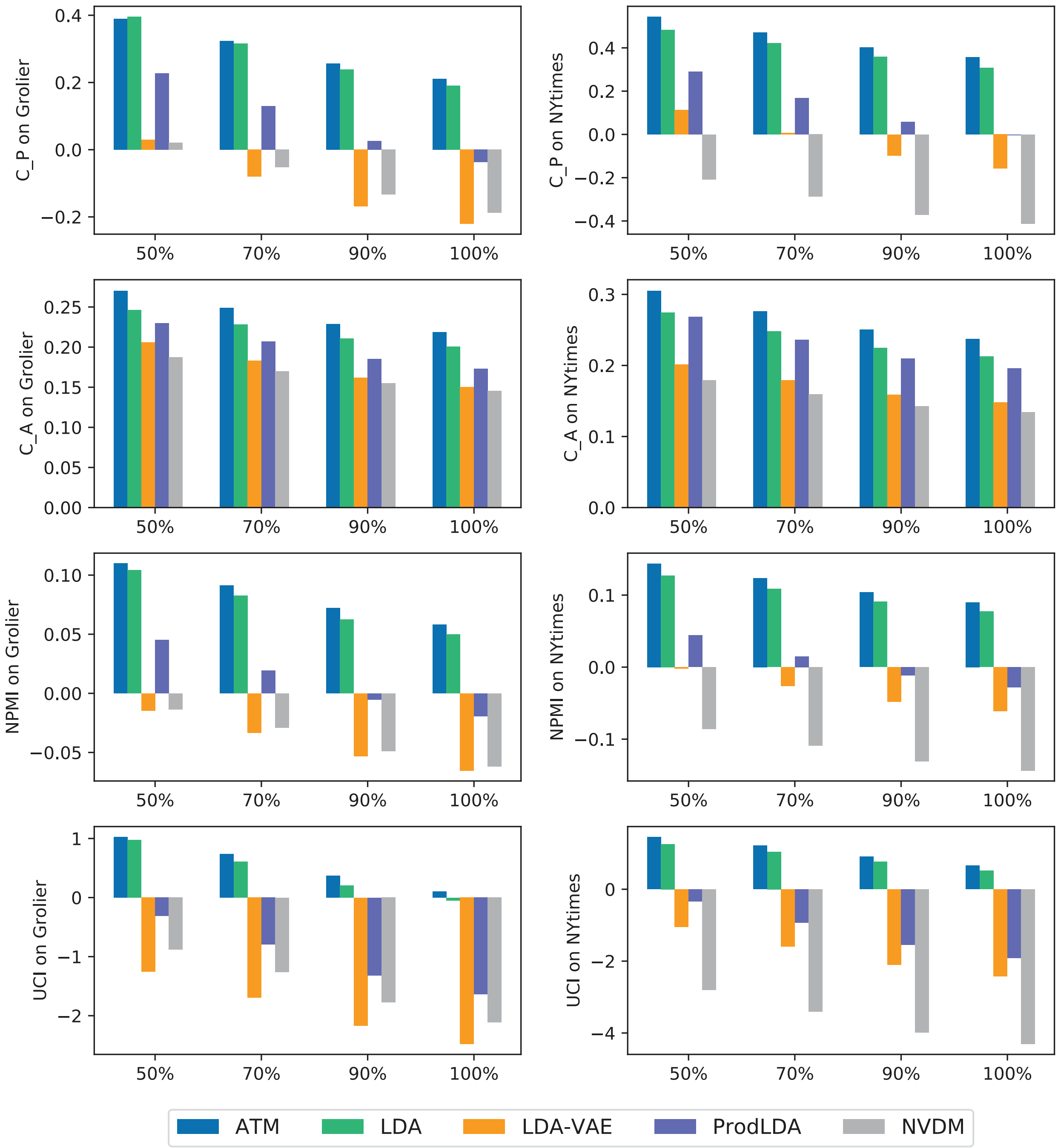

Fig. 2. Average topic coherence on Grolier and NYtimes with five topic settings [20, 30, 50, 75, 100] among topics whose coherence values are ranked at the top 50%, 70%, 90% and 100% positions.

We use the pre-identified named entities,[7] verbs, nouns and adjectives to construct the word set of organization, location, person and keywords. When using ATM for event extraction, these four word sets and the event-specific word distribution are used to generate the related topics. For example, the organization topic of an event could be obtained by sorting the words in the organization word set based on the corresponding probabilities in the event-specific word distribution learned by ATM. Table 4 shows the example events extracted by ATM and LEM where the relevant words are highlighted in bold. It can be observed that ATM performs comparably with LEM. However, while LEM required the model-specific inference algorithm to be derived, ATM did not need any modification of its network architecture or parameter estimation procedure.

To validate the correctness of the extracted events, we retrieve the title of articles using the event-related words from ATM and obtain the following results:

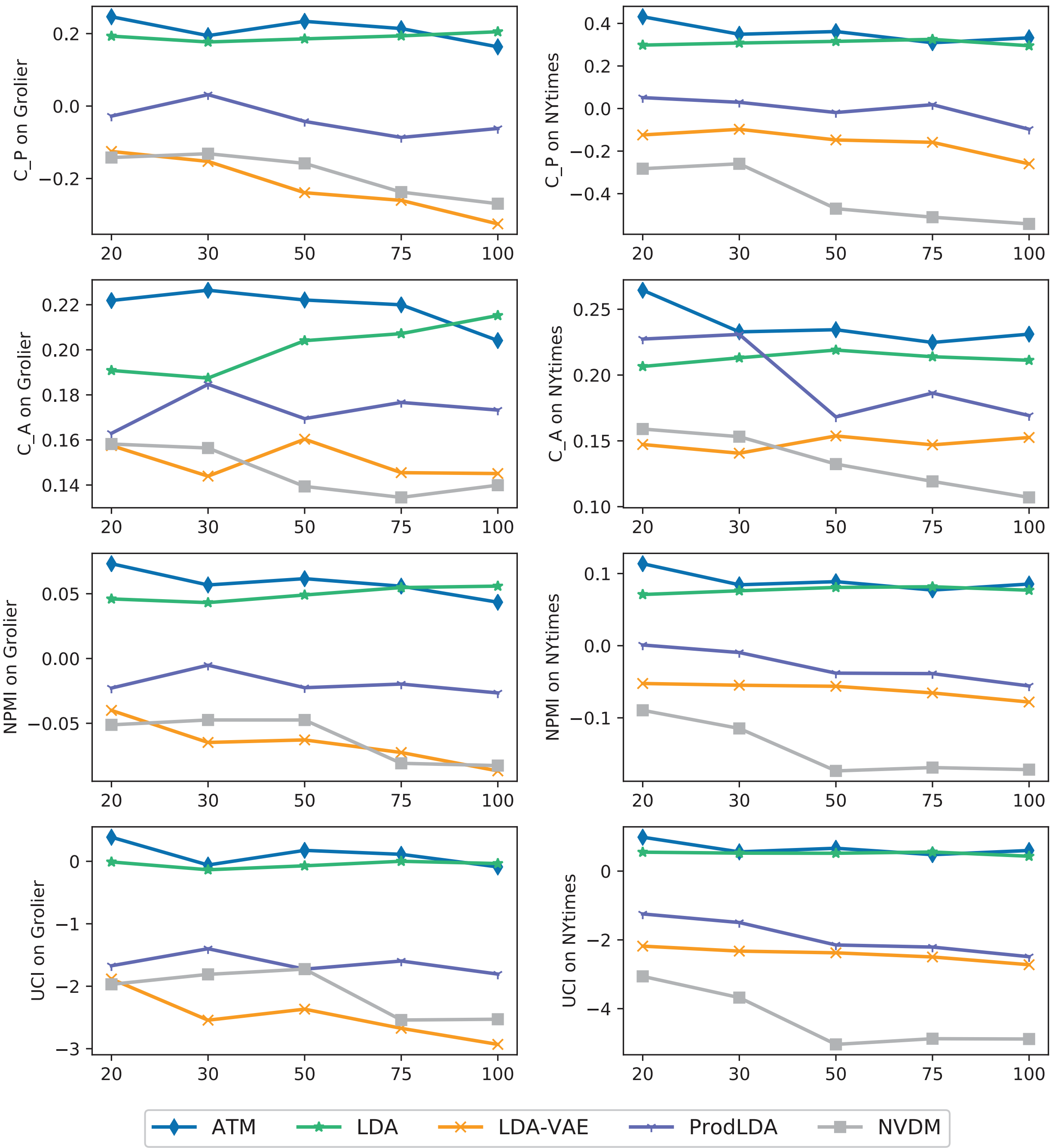

Fig. 3. Average topic coherence (100%) on Grolier and NYtimes datasets vs. different topic setting [20, 30, 50, 75, 100].

- *Missing Malaysia Airlines flight MH370: Government report suggests official search for plane did not begin until four hours after disappearance.*
- *Saudi Arabia finds 26 more cases of MERS, Egypt reports first sufferer.*
- *India's defence experts and politicos condemn Pak Army Chief's Kashmir statement.*
- *Top BJP leaders, Rajnath Singh, MM Joshi, Sushma Swaraj to campaign for Narendra Modi in Varanasi.*
- *Turkey May Day protests hit by tear gas near Taksim Square - Panorama.*

It is clear that the retrieved titles indeed correspond well with the extracted events by ATM.

Table 3
Topic examples of all the models, italics means out-of-topic.

Model	Topics
	jet flight airline hour plane passenger trip plan travel pilot
	stock market companies money investor technology fund investment company business
ATM	**music song musical album jazz band record recording mp3 composer**
	voter vote poll republican race primary percent election campaign democratic
	film movie actor director award movies character theater production play
	flight plane *ship crew* **air pilot hour** *boat* **passenger airport**
	stock market *percent* **investor analyst** *quarter* **investment shares share fund**
LDA	**music song band sound record artist album show musical rock**
	voter vote poll election campaign primary candidates republican race party
	film movie character play actor director movies *minutes* **theater** *cast*
	wireless **customer** *telecommunication* **airlines** *broadband satellites phones subscriber* **airline** *provider*
	brokerage securities broker lender buyer transaction investor investment stock borrower
ProdLDA	**musical album** *playwright* **composer** *choreographer onstage* **songwriter song guitarist repertory**
	voter vote votes election electoral polling poll presidential primaries *turnout*
	film comedy *beginitalic enditalic sci* **filmmaker cinematic filmmaking movie starring**
	passenger destination traveler *fares booking* **airlines luggage routes** *rider excursion*
	acquisition shareholder merge *takeover acquire* **merger** *consolidated* **stockholder** *suitor consolidation*
LDA-VAE	**soloist operatic composer** *repertory* **troupe** *choreographer choreography* **sung** *dances* **recital**
	balloting nominating election elect incumbent victor primaries *contested* **electoral vote**
	moviegoer studios filmmaker movies film filming *vh1* **studio stardom** *rapper*
	nesting instructor ranchers wingspan veteran **fly** *manager* **pilot** *ecosystems* **flight**
	company billion companies *production equipment processed processing producer manufacturing products*
NVDM	**conducting conductor instrumental** *interval staff* **discography** *knighted radioactive charge* **director**
	degrees **national party** *billion* **nations** *decrease university exceed disorder nuclear*
	bay **film** *indian french company* **novel** *dec lake explorer travels*

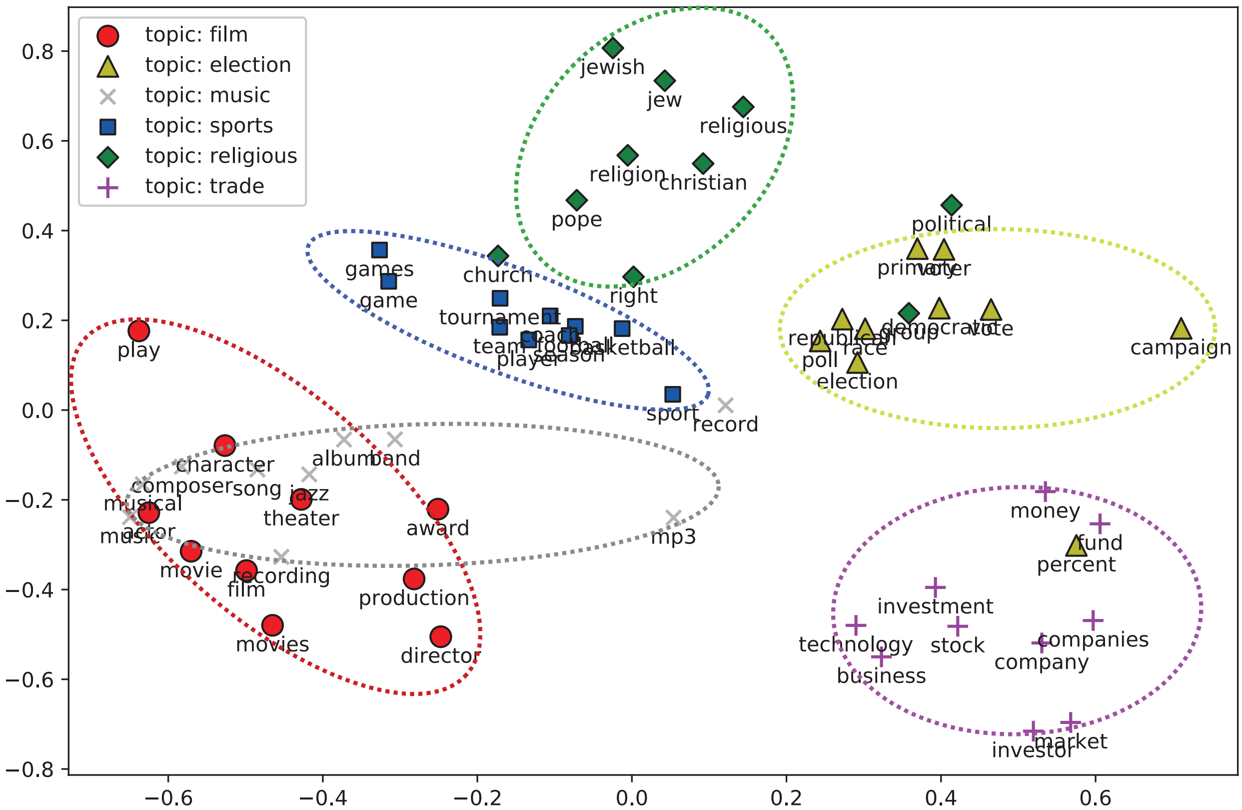

Fig. 4. Visualization of the topic words from the six selected topics.

Table 4
The event examples extracted by ATM and LEM.

Events	Method	Representative words
MH370	ATM	org: **air airlines** ministry transport international loc: **malaysia** beijing france vietnam dubai per: **hishammuddin hussein** najib kerry lee key: **search flight aircraft air plane**
	LEM	org: **airlines air** international transport government loc: **malaysia** south korea beijing us per: **hussein hishammuddin** fitch long park key: **flight airlines plane** preliminary **search**
Saudi MERS	ATM	org: **community** ministry **saudi healthcare** government loc: **saudi** ontario iran canada jeddah per: **president** obama jordan kerry walker key: **health hospital patients disease medical**
	LEM	org: **saudi** jordan army eastern state loc: east **saudi** jordan egypt israel per: jordan **president** frank rob geldof key: east middle **respiratory syndrome health**
Pakistan vs. India	ATM	org: **army kashmir** sharif taliban afghanistan loc: **pakistan kashmir india** afghanistan islamabad per: **sharif** kerry khan president lovell key: **army peace chief region** province
	LEM	org: **army kashmir** sharif government congress loc: **pakistan kashmir** islamabad **india** delhi per: **sharif** tsvangirai morgan dube biti key: **army chief** vein news **peace**
Indian Election	ATM	org: **bjp party congress singh gandhi** loc: **gujarat india varanasi delhi** seemandhra per: **modi singh** gandhi naidu khan key: **congress election candidate minister leader**
	LEM	org: **bjp congress party commission delhi** loc: **delhi gujarat** modis **varanasi india** per: **modi** gandhi **singh** modis president key: **prime candidate election ministerial congress**
Taksim Clash	ATM	org: **police** city **government** erdogan **union** loc: **taksim istanbul** city **turkey union** per: **erdogan** park walker quinn hall key: **square protesters tear demonstrators street**
	LEM	org: **police** international **labor** central greenpeace loc: **istanbul taksim turkey** rotterdam **union** per: mark **erdogan** geldof park hall key: **protesters square** international **gas water**

5. Conclusions

We have proposed a novel topic modeling approach based on adversarial training. The proposed approach, ATM, models the topics with Dirichlet prior and employs the generator network to learn the semantic patterns among latent topics. Apart from automatically generating latent topics from a text corpus, it could also produce word-level semantic representations as a side product. The experimental comparison with the state-of-the-art methods show that ATM achieves improved topical coherence results. Moreover, the feasibility of porting ATM for tasks other than topic modeling has been verified for open domain event extraction. In the future, we want to incorporate the sequential information contained in texts into GAN based topic modeling approaches and devise a topic driven sentence generation model. And an extension to cope with the data sparsity in short text is also our future work. Besides, another direction we are interested in exploring is to develop dynamic and correlated topic models based on adversarial training.

Acknowledgments

We would like to thank anonymous reviewers for their valuable comments and helpful suggestions. This work was funded by the National Key Research and Development Program of China (2016YFC1306704), the National Natural Science Foundation of China (61772132), the Natural Science Foundation of Jiangsu Province of China (BK20161430).

References

Arjovsky, M., Chintala, S., & Bottou, L. (2017). *Wasserstein generative adversarial networks. Proceedings of the 34th international conference on machine learning, Vol. 70 of proceedings of machine learning research, PMLR, 465 international convention centre, Sydney, Australia*214–223.

Asuncion, A., Welling, M., Smyth, P., & Teh, Y. W. (2009). *On smoothing and inference for topic models. Proceedings of the twenty-fifth conference on uncertainty in artificial intelligence*. Montreal, Canada: AUAI Press27–34.
Athiwaratkun, B., Wilson, A., & Anandkumar, A. (2018). *Probabilistic fasttext for multi-sense word embeddings. Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)*. Melbourne, Australia: Association for Computational Linguistics1–11.
Blei, D. M. (2012). Probabilistic topic models. *Communications of the ACM, 55*(4), 77–84. https://doi.org/10.1145/2133806.2133826.
Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research, 3*(Jan), 993–1022.
Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics, 5*, 135–146. https://doi.org/10.1162/tacl_a_00051.
Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., & Blei, D. M. (2009). *Reading tea leaves: How humans interpret topic models. Advances in neural information processing systems*288–296.
Chen, L. (2017). An effective lda-based time topic model to improve blog search performance. *Inf. Process. Manage. 53*(6), 1299–1319. https://doi.org/10.1016/j.ipm.2017.08.001.
Fedus, W., Goodfellow, I., & Dai, A. M. (2018). *Maskgan: Better text generation via filling in the _. 6th international conference on learning representations (iclr). Vancouver, Canada*.
Fernández-Reyes, F. C., Valadez, J. H., & Montes-y-Gómez, M. (2018). A prospect-guided global query expansion strategy using word embeddings. *Inf. Process. Manage. 54*(1), 1–13. https://doi.org/10.1016/j.ipm.2017.09.001.
Finkel, J. R., Grenager, T., & Manning, C. (2005). *Incorporating non-local information into information extraction systems by gibbs sampling. Proceedings of the 43rd annual meeting on association for computational linguistics*. Association for Computational Linguistics363–370. https://doi.org/10.3115/1219840.1219885.
Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... Bengio, Y. (2014). Generative adversarial nets. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, & K. Q. Weinberger (Eds.). *Advances in neural information processing systems 27* (pp. 2672–2680). Curran Associates, Inc.
Grave, E., Mikolov, T., Joulin, A., & Bojanowski, P. (2017). *Bag of tricks for efficient text classification. Proceedings of the 15th conference of the european chapter of the association for computational linguistics, eacl*3–7. https://doi.org/10.18653/v1/e17-2068.
Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National academy of Sciences, 101*(suppl 1), 5228–5235. https://doi.org/10.1073/pnas.0307752101.
Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., & Courville, A. C. (2017). *Improved training of wasserstein gans. Advances in neural information processing systems 30*. Curran Associates, Inc.5767–5777.
Hagen, L. (2018). Content analysis of e-petitions with topic modeling: How to train and evaluate LDA models? *Inf. Process. Manage. 54*(6), 1292–1307. https://doi.org/10.1016/j.ipm.2018.05.006.
Hinton, G. E., & Salakhutdinov, R. R. (2009). Replicated softmax: an undirected topic model. In Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, & A. Culotta (Eds.). *Advances in neural information processing systems 22* (pp. 1607–1614). Curran Associates, Inc.
Hsu, F., Lee, H., Chang, T., & Sung, Y. (2018). Automated estimation of item difficulty for multiple-choice tests: An application of word embedding techniques. *Inf. Process. Manage. 54*(6), 969–984. https://doi.org/10.1016/j.ipm.2018.06.007.
Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., & Xing, E. P. (2017). *Toward controlled generation of text. Proceedings of the 34th international conference on machine learning-volume 70*. JMLR. org1587–1596.
Isola, P., Zhu, J.-Y., Zhou, T., & Efros, A. A. (2017). *Image-to-image translation with conditional adversarial networks. 2017 ieee conference on computer vision and pattern recognition (cvpr)*. IEEE5967–5976. https://doi.org/10.1109/CVPR.2017.632.
Kingma, D. P., & Ba, J. (2015). *Adam: A method for stochastic optimization. 3rd international conference on learning representations (ICLR). San Diego, USA*.
Kingma, D. P., & Welling, M. (2014). *Auto-encoding variational bayes. 2nd international conference on learning representations (ICLR). Alberta, Canada*.
Le, Q., & Mikolov, T. (2014). *Distributed representations of sentences and documents. Proceedings of the 31 st international conference on machine learning, Beijing, China*1188–1196.
Ledig, C., Theis, L., Huszár, F., Caballero, J., Cunningham, A., Acosta, A., ... Wang, Z., et al. (2017). *Photo-realistic single image super-resolution using a generative adversarial network. 2017 ieee conference on computer vision and pattern recognition (cvpr)*. IEEE105–114. https://doi.org/10.1109/CVPR.2017.19.
Li, Y., & Ye, J. (2018). *Learning adversarial networks for semi-supervised text classification via policy gradient. Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*. ACM1715–1723. https://doi.org/10.1145/3219819.3219956.
Lin, C., & He, Y. (2009). *Joint sentiment/topic model for sentiment analysis. Proceedings of the 18th acm conference on information and knowledge management*. Hong Kong, China: ACM375–384. https://doi.org/10.1145/1645953.1646003.
Lin, K., Li, D., He, X., Zhang, Z., & Sun, M.-t. (2017). Adversarial ranking for language generation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.). *Advances in neural information processing systems 30* (pp. 3155–3165). Curran Associates, Inc.
Miao, Y., Yu, L., & Blunsom, P. (2016). *Neural variational inference for text processing. Proceedings of the 33rd international conference on machine learning, New York, NY, USA*1727–1736.
Mimno, D., Wallach, H. M., Talley, E., Leenders, M., & McCallum, A. (2011). *Optimizing semantic coherence in topic models. Proceedings of the conference on empirical methods in natural language processing*. Association for Computational Linguistics262–272.
Miyato, T., Dai, A. M., & Goodfellow, I. (2017). *Adversarial training methods for semi-supervised text classification. 5th international conference on learning representations (ICLR), Toulon, France*.
Nowozin, S., Cseke, B., & Tomioka, R. (2016). f-gan: training generative neural samplers using variational divergence minimization. In D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, & R. Garnett (Eds.). *Advances in neural information processing systems 29* (pp. 271–279). Curran Associates, Inc.
Pennington, J., Socher, R., & Manning, C. (2014). *Glove: Global vectors for word representation. Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP 2014), Doha, Qatar*1532–1543. https://doi.org/10.3115/v1/d14-1162.
Radford, A., Metz, L., & Chintala, S. (2016). *Unsupervised representation learning with deep convolutional generative adversarial networks. 4th international conference on learning representations (ICLR), San Juan, Puerto Rico*.
Röder, M., Both, A., & Hinneburg, A. (2015). *Exploring the space of topic coherence measures. Proceedings of the eighth ACM international conference on web search and data mining*. ACM399–408. https://doi.org/10.1145/2684822.2685324.
Srivastava, A., & Sutton, C. (2017). *Autoencoding variational inference for topic models. 5th international conference on learning representations (ICLR), Toulon, France*.
Wang, K., & Wan, X. (2018). *Sentigan: Generating sentimental texts via mixture adversarial networks. Proceedings of the twenty-seventh international joint conference on artificial intelligence, Stockholm, Sweden*4446–4452. https://doi.org/10.24963/ijcai.2018/618.
Wang, Y., & Lee, H.-y. (2018). *Learning to encode text as human-readable summaries using generative adversarial networks. Proceedings of the 2018 conference on empirical methods in natural language processing*. Brussels, Belgium: Association for Computational Linguistics4187–4195.
Xiao, D., Ji, Y., Li, Y., Zhuang, F., & Shi, C. (2018). Coupled matrix factorization and topic modeling for aspect mining. *Information Processing Management, 54*(6), 861–873. https://doi.org/10.1016/j.ipm.2018.05.002.
Yang, Z., Hu, Z., Dyer, C., Xing, E. P., & Berg-Kirkpatrick, T. (2018). *Unsupervised text style transfer using language models as discriminators. Advances in neural information processing systems 31*. Curran Associates, Inc.7287–7298.
Yu, L., Zhang, W., Wang, J., & Yu, Y. (2017). *Seqgan: Sequence generative adversarial nets with policy gradient. Proceedings of the thirty-first AAAIconference on artificial intelligence, San Francisco, USA*2852–2858.
Zeng, D., Dai, Y., Li, F., Sherratt, R. S., & Wang, J. (2018). Adversarial learning for distant supervised relation extraction. *Computers, Materials & Continua, 55*(1), 121–136. https://doi.org/10.3970/cmc.2018.055.121.
Zhou, D., Chen, L., & He, Y. (2014). *A simple Bayesian modelling approach to event extraction from twitter. Proceedings of the 52nd annual meeting of the 390 association for computational linguistics, ACL 2014, Vol. 2, Baltimore, Maryland, USA*700–705. https://doi.org/10.3115/v1/p14-2114.
Zhou, D., Chen, L., Zhang, X., & He, Y. (2017). Unsupervised event exploration from social text streams. *Intelligent Data Analysis, 21*(4), 849–866. https://doi.org/10.3233/IDA-160048.

Zhou, D., Gao, T., & He, Y. (2016). *Jointly event extraction and visualization on twitter via probabilistic modelling. Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)1. Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* 269–278.

Zhou, D., Xu, H., & He, Y. (2015). *An unsupervised Bayesian modelling approach for storyline detection on news articles. Proceedings of the 2015 conference on empirical methods in natural language processing, Lisbon, Portugal*1943–1948. https://doi.org/10.18653/v1/d15-1225.