ConvNTM: Conversational Neural Topic Model

Hongda Sun,[1,*] **Quan Tu,**[1,*] **Jinpeng Li,**[2] **Rui Yan**[1,3,†]

[1] Gaoling School of Artificial Intelligence, Renmin University of China
[2] Wangxuan Institute of Computer Technology, Peking University
[3] Engineering Research Center of Next-Generation Intelligent Search and Recommendation, Ministry of Education
{sunhongda98, quantu}@ruc.edu.cn, lijinpeng@stu.pku.edu.cn, ruiyan@ruc.edu.cn

Abstract

Topic models have been thoroughly investigated for multiple years due to their great potential in analyzing and understanding texts. Recently, researchers combine the study of topic models with deep learning techniques, known as Neural Topic Models (NTMs). However, existing NTMs are mainly tested based on general document modeling without considering different textual analysis scenarios. We assume that there are different characteristics to model topics in different textual analysis tasks. In this paper, we propose a Conversational Neural Topic Model (ConvNTM) designed in particular for the conversational scenario. Unlike the general document topic modeling, a conversation session lasts for multiple turns: each short-text utterance complies with a single topic distribution and these topic distributions are dependent across turns. Moreover, there are roles in conversations, a.k.a., speakers and addressees. Topic distributions are partially determined by such roles in conversations. We take these factors into account to model topics in conversations via the multi-turn and multi-role formulation. We also leverage the word co-occurrence relationship as a new training objective to further improve topic quality. Comprehensive experimental results based on the benchmark datasets demonstrate that our proposed ConvNTM achieves the best performance both in topic modeling and in typical downstream tasks within conversational research (i.e., dialogue act classification and dialogue response generation).

Introduction

Topic models are used to discover abstract topics in a series of documents to understand the latent semantics of a text corpus (Hofmann 1999; Blei, Ng, and Jordan 2003). With the recent development of neural networks and generative models, various neural topic models (NTMs) have been proposed and applied in document classification, retrieval, semantic analysis, etc (Larochelle and Lauly 2012; Dieng et al. 2017; Zhao et al. 2021).

Most existing NTMs are designed for document analysis. Their main modeling scenarios lie in news articles or social platform posts (Lang 1995; Li et al. 2016), with less consideration on various other textual analysis scenarios. However, we assume that there are different characteristics to model topics in different textual analysis tasks. For general topic modeling on long documents, each document is typically assigned a topic distribution, and words in the document are iteratively generated based on the distribution (Blei, Ng, and Jordan 2003; Miao, Grefenstette, and Blunsom 2017; Dieng, Ruiz, and Blei 2020). For short-text topic modeling, since the word co-occurrence information is limited, the sparsity problem should be considered during topic extraction (Cheng et al. 2014; Zhu, Feng, and Li 2018; Lin, Hu, and Guo 2019). While in the conversational scenario, topic modeling is even more complicated with the following two unique properties to discover topics: 1) A conversation session generally consists of multiple turns of short-text utterances (Zhang et al. 2019; Adiwardana et al. 2020), which usually follow different topic distributions (Sun, Loparo, and Kolacinski 2020). A simple operation of utterance concatenation as a long document—which is the way of existing NTMs—leads to the omission of dialogue structural information in topic modeling. As a matter of fact, utterances from different turns are connected and topic distributions are dependent across turns. 2) There are multiple roles within a conversation session, speakers and addressees (Holtgraves, Srull, and Socall 1989). A series of studies indicate that such roles are essential in keeping the topic consistency and content coherence within a conversation (Kim and Vossen 2021; Ma, Zhang, and Zhao 2021). Without the modeling of the conversational structure with multiple roles, it is likely that the topic discovery will be compromised due to the missing consistency and coherence in dialogue understanding.

To this end, we propose a Conversational Neural Topic Model (ConvNTM) which is in particular designed for the conversational scenario with the mentioned characteristics formulated in topic modeling. Specifically, we develop a hierarchical conversation encoder to capture the multi-turn dialogue structure. A sequence encoder is utilized to model the conversation contexts and extract utterance-level representations for the role modeling of speakers and addressees. Then we construct a multi-role interaction graph to model speaker/addressee information from two perspectives. On the one hand, different roles hold personalized topic distributions and they need to integrate the intra-speaker information in their utterances to determine the current topic. All utterances from a particular speaker should be consistent on

[*] These authors contributed equally.

[†] Corresponding author: Rui Yan (ruiyan@ruc.edu.cn)

the topic distribution to avoid contradictions. On the other hand, a speaker can decide whether to keep or change the topic for themselves based on the utterances of other speakers. That is, topic maintenance and switching in a conversation continue under the inter-speaker interaction. We employ a graph neural network to reason the speaker graph and integrate intra-speaker and inter-speaker dependencies among utterances. The graph encoder and the sequence encoder cooperate to adequately capture the hierarchical structure of the conversation. The learned representations of the graph encoder are incorporated into the topic modeling process.

Considering the structural properties of the conversation, we make reasonable assumptions on the topic distribution. First, to prevent confusion from modeling the entire conversation with a single topic, we perform fine-grained topic modeling by assuming that each utterance compiles with a specific topic distribution. These distributions are mutually influenced across multiple turns. Additionally, the topic distribution of each utterance is assumed to rely on both global and local topic information. We assign each speaker a global topic distribution as a specific role. Then the local topic information in each utterance will be extracted and interacted with the global role information to produce final topic distribution. Based on the novel graphical model of ConvNTM, corresponding neural variational inference methods are carried out for model learning. Furthermore, to further improve topic coherence, we leverage the word co-occurrence information as a new training objective, which can be jointly trained with the original objective of neural variational inference. The ConvNTM that grasps the word co-occurrence relationship can make related words tend to be clustered into the same topic, which helps to obtain higher quality topic-word distributions.

We run experiments based on the public benchmark conversational datasets, DailyDialog and EmpatheticDialogues. Our proposed ConvNTM achieves the best performance on topic modeling in terms of topic coherence and quality metrics, which indicates that ConvNTM has better topic interpretability on the dialogue corpora compared against general NTMs. Furthermore, we also conduct experiments on typical downstream tasks for dialogues based on the discovered topics, including dialogue act classification and response generation. The experimental results indicate that with the help of the topics discovered by ConvNTM, the performance is prominently boosted compared against the baselines without topic information and existing topic-aware dialogue methods.

Our overall contributions are summarized as follows:

- To the best of our knowledge, for the first time, we propose ConvNTM, the neural topic model in particular designed for the conversational scenario to formulate the multi-turn structure in dialogues to discover topics.
- Considering the multi-role interactions (speakers and addressees) in conversations, we perform utterance-level fine-grained topic modeling and fuse global and local topic information to determine topic distributions.
- We also leverage the word co-occurrence relationship to constrain the topic-word distribution, which can be coordinated and jointly trained with the neural variational inference objective to further improve topic coherence.

Related Work

Topic Model

Topic modeling has always been a catalyst for other research areas in Natural Language Process (NLP) (Panwar et al. 2020; Jin et al. 2021; Srivastava and Sutton 2016). A classic statistical topic model is Latent Dirichlet Allocation (LDA), which is based on Gibbs sampling to extract topics from documents (Blei, Ng, and Jordan 2003). With the development of deep generative models, it has led to the study of neural topic models (NTMs) (Miao, Grefenstette, and Blunsom 2017; Zhu, Feng, and Li 2018; Wang, Zhou, and He 2019). Variational Autoencoder (VAE) (Kingma and Welling 2013) is the most widely used framework for NTMs. GSM (Miao, Grefenstette, and Blunsom 2017) replaces the prior with a Gaussian softmax function. ProdLDA (Srivastava and Sutton 2017) constructs a Laplace approximation to the Dirichlet prior. ETM (Dieng, Ruiz, and Blei 2020) shares the embedding space between words and topics. GNTM (Shen et al. 2021) adds the document graph into the generative process of topic modeling. With the progress of social platforms (e.g. Microblog and Twitter), application-oriented NTMs keep pouring out. LeadLDA (Li et al. 2016) considers the tree structure based on the re-posts and replying relations. ForumLDA (Chen and Ren 2017) cooperatively models the evolution of a root post, as well as its relevant and irrelevant response posts to detect topics. In these posts, people always discuss a single hot topic. While in our target conversation scenario, speakers with different roles may switch topics in multiple turns.

Multi-Turn Dialogue

The simple concatenation of multi-turn dialogue contexts performs poorly since it makes the latent dialogue structure ignored. Abundant works suggest that the multi-turn dialogue requires specific modeling methods (Qiu et al. 2020a,b). Serban et al. devise the hierarchical LSTM to encode the structure and generate responses. DialoFlow (Li et al. 2021) is another solution, which views the dialogue as a dynamic flow and designs three objectives to capture the information dynamics. Moreover, the speaker feature is also considered as a pivotal factor in the dialogue. He et al. incorporate the turn changes among speakers to capture the fine-grained semantics of dialogue. Gu et al. introduce a speaker-aware disentanglement strategy to tackle the entangled dialogues and improve the performance of multi-turn dialogue response selection. Topic-aware models take the advantage of the related topics to make conversational modeling more consistent. Liu et al. propose two topic-aware contrastive learning objectives to handle information scattering challenges for the dialogue summarization task. Zhu et al. propose a topic-driven knowledge-aware Transformer to deal with the emotion detection in dialogue. We hope that our ConvNTM can better facilitate the development of topic-aware methods.

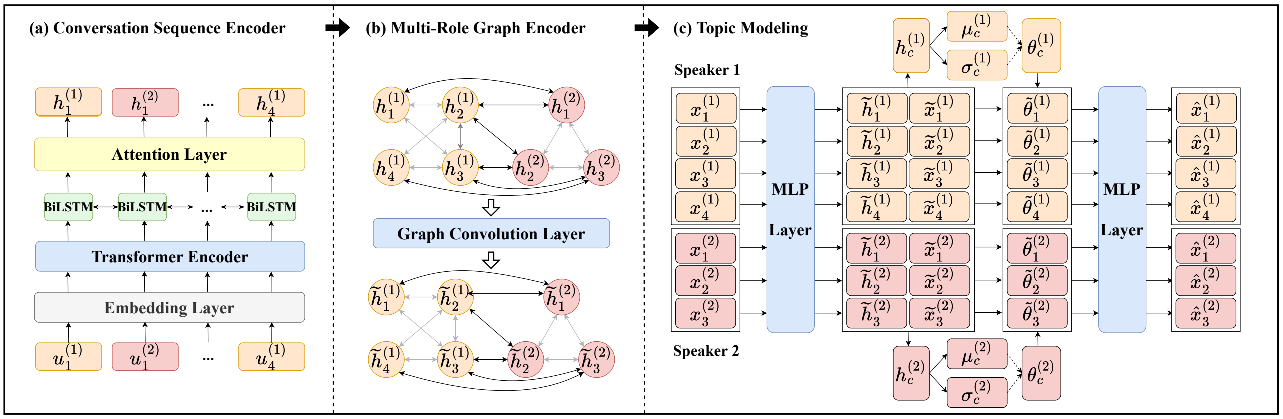

Figure 1: The model overview of ConvNTM: a) The conversation sequence encoder for modeling the multi-turn conversation contexts; b) The multi-role graph encoder for formulating the intra-speaker and inter-speaker dependencies; c) The topic modeling module to reconstruct utterance-level BoWs based on the fusion of global and local topic information.

Conversational Neural Topic Model

In this section, we describe the modules and training objectives of ConvNTM in detail. The model overview of the ConvNTM is illustrated in Figure 1.

Hierarchical Conversation Encoder

To fully extract semantic information in the multi-turn conversation to help topic modeling, we use a hierarchical framework in which both a sequence encoder and a graph encoder cooperatively encode the conversation contexts to better handle cross-utterance dependencies.

Conversation sequence encoder. To capture the multi-turn structure of the conversation, we employ a sequence encoder that models the conversation contexts from word level to utterance level. Suppose that a conversation session c has J speakers, and the speaker j has n_j utterances: $\{u_1^{(j)}, u_2^{(j)}, \cdots, u_{n_j}^{(j)}\}$. The words in the k-th utterance $u_k^{(j)}$ are first encoded as $e_k^{(j)}$ through an embedding layer f_e. A two-layer Transformer encoder f_{trm} is then used to further process $e_k^{(j)}$ and obtain the utterance-level representation $s_k^{(j)}$ from the `[CLS]` token. In order to enhance the contextual relationship among the multi-turn utterances, we feed the Transformer outputs into a bidirectional LSTM f_{rnn} and a standard self-attention layer f_{attn} successively. Finally, we denote the learned utterance representations for the speaker j as $\{h_1^{(j)}, h_2^{(j)}, \cdots, h_{n_j}^{(j)}\}$. The encoding process for the sequence encoder can be formulated as:

$$e_k^{(j)} = f_e(u_k^{(j)}), \tag{1}$$

$$s_k^{(j)} = f_{trm}(e_k^{(j)})_{\texttt{[CLS]}}, \tag{2}$$

$$\widetilde{s}_k^{(j)} = f_{rnn}(s_1^{(j)}, s_2^{(j)}, \cdots, s_{n_j}^{(j)})_k, \tag{3}$$

$$h_k^{(j)} = f_{attn}(\widetilde{s}_1^{(j)}, \widetilde{s}_2^{(j)}, \cdots, \widetilde{s}_{n_j}^{(j)})_k. \tag{4}$$

Multi-role graph encoder. Considering the impact of speaker information in a conversation, we construct a graph for the conversation to describe the multi-role interactions. We denote each utterance representation $h_k^{(j)}$ as a node, and the two types of edges between nodes reflect the intra-speaker and inter-speaker dependencies. First, the individual roles of each speaker in the dialogue have a significant impact on the continuation of the conversation. The speaker tends to organize what he/she has said in the previous utterances to determine the topic of the current utterance. Therefore, we consider intra-speaker dependency to keep the topic consistency and avoid contradictions. For the speaker j, we add a bidirectional edge between $h_{k_1}^{(j)}$ and $h_{k_2}^{(j)}$ only if $|k_1 - k_2| \leq K_s$, where K_s indicates the window size for aggregating contextual utterances from the same speaker. Second, a speaker will give feedback on the utterance contents from other speakers, and then decide whether to keep or shift the current topic. It is also necessary to construct the inter-speaker dependency in the graph to simulate the dynamic interactions. For two speakers j_1 and j_2, we add a bidirectional edge between $h_{k_{j_1}}^{(j_1)}$ and $h_{k_{j_2}}^{(j_2)}$ only if $|k_{j_1} - k_{j_2}| \leq K_c$, where K_c indicates the absolute distance window size of two utterances in the conversation. Taking Figure 1 as an example, the second speaker has three utterances interspersed with the first speaker's four utterances. In this graph, the intra-speaker edges are in grey while the inter-speaker edges are in black. We utilize a graph convolution network (GCN) f_{gcn} to update the utterance representations under the multi-role interaction relations. Therefore, the learned utterance representation $\widetilde{h}_k^{(j)}$ is given by:

$$\widetilde{h}_k^{(j)} = f_{gcn}(h_k^{(j)}). \tag{5}$$

Topic Modeling

Based on the speaker-oriented utterance representations from the graph encoder, we then introduce our techniques for topic modeling.

Topic distribution assumption. Given a general document, the generative process of existing NTMs is mainly divided into three steps: 1) sample a topic distribution θ for a document or each sentence; 2) sample a topic assignment z_t for each word w_t from the topic distribution θ; 3) generate each word w_t independently from the corresponding topic-word distribution β_{z_t}. However, a conversation contains multiple turns of utterances, the topics in the utterances follow their respective topic distributions and are related to each other. The roles of different speakers also influence the topic determination. Thus, we need to adapt the original assumptions on the topic distribution according to the unique properties of the conversation. Specifically, we assume that each speaker j in the conversation session c holds a global topic information $\theta_c^{(j)}$, and each utterance k has local topic information $\theta_k^{(j)}$, which is fused with the corresponding global topic to determine the eventual topic distribution $\widetilde{\theta}_k^{(j)}$.

NTM framework. We process the n_j utterances of each speaker j into bag-of-words (BoW) representations: $\{x_1^{(j)}, x_2^{(j)}, \cdots, x_{n_j}^{(j)}\}$, where $x_k^{(j)}$ is a $|V|$-dimensional multi-hot encoded vector for the k-th utterance and V is the BoW vocabulary. Note that each g_* mentioned below represents a multilayer perceptron (MLP). We first normalize the BoW vector $x_k^{(j)}$ and then use g_x to extract the representation $\widetilde{x}_k^{(j)}$:

$$\widetilde{x}_k^{(j)} = g_x\left(\frac{x_k^{(j)}}{\sum_{v=1}^{|V|}(x_k^{(j)})_v}\right). \tag{6}$$

In order to introduce multi-role interactions into topic modeling, we concatenate $\widetilde{x}_k^{(j)}$ with the node representation $\widetilde{h}_k^{(j)}$ given by the graph encoder. Then, we obtain the local topic information $\theta_k^{(j)}$ of the utterance through g_s:

$$\theta_k^{(j)} = g_s(\widetilde{x}_k^{(j)} \oplus \widetilde{h}_k^{(j)}). \tag{7}$$

Next, all the utterances of each speaker j are integrated to derive the global speaker-aware representation $h_c^{(j)}$, which can be used to estimate the prior variables $\mu_c^{(j)}$ and $\log \sigma_c^{(j)}$ via two separate networks g_μ and g_σ:

$$h_c^{(j)} = \tanh\left(\sum_{k=1}^{n_j} g_c(\widetilde{x}_k^{(j)} \oplus \widetilde{h}_k^{(j)}) \cdot \theta_k^{(j)}\right), \tag{8}$$

$$\mu_c^{(j)} = g_\mu(h_c^{(j)}), \quad \log \sigma_c^{(j)} = g_\sigma(h_c^{(j)}). \tag{9}$$

With the reparameterisation trick (Kingma and Welling 2013), we can sample a latent variable $z_c^{(j)} \sim \mathcal{N}(\mu_c^{(j)}, \sigma_c^{(j)})$. Then we use g_θ to generate the global topic distribution $\theta_c^{(j)}$:

$$\theta_c^{(j)} = \text{softmax}(g_\theta(z_c^{(j)})). \tag{10}$$

Finally, we can use g_f to fuse local and global topic information to derive the eventual topic distribution $\widetilde{\theta}_k^{(j)}$:

$$\widetilde{\theta}_k^{(j)} = g_f(\theta_k^{(j)} \oplus \theta_c^{(j)}). \tag{11}$$

Assuming that the number of topics is K, all the above topic distributions are K-dimensional vectors. To reconstruct the BoWs for each utterance in the conversation, we leverage a weighted matrix $\beta \in \mathbb{R}^{K \times |V|}$ to represent K topic-word distributions. The reconstructed utterance BoW can be derived as:

$$\hat{x}_k^{(j)} = \text{softmax}(\widetilde{\theta}_k^{(j)} \beta). \tag{12}$$

Generative process. Based on the above definitions, we summarize the generative process of ConvNTM as follows.

1. For each speaker j in the conversation session c:

   i) Sample the latent variable $z_c^{(j)} \sim \mathcal{N}(\mu_c^{(j)}, \sigma_c^{(j)})$;

   ii) Draw $\theta_c^{(j)} = \text{softmax}(g_\theta(z_c^{(j)}))$ as the global topic distribution.

2. For each utterance $u_k^{(j)}$ of the speaker j:

   i) Draw $\theta_k^{(j)}$ as the local topic information;

   ii) Draw $\widetilde{\theta}_k^{(j)}$ by fusing $\theta_c^{(j)}$ and $\theta_k^{(j)}$;

   iii) For each word w in the utterance $u_k^{(j)}$: draw $w \sim \text{softmax}(\widetilde{\theta}_k^{(j)} \beta)$.

The Joint Training Objective

Neural variational inference objective. Under the generative process of ConvNTM, the marginal likelihood of the conversation session c is decomposed as:

$$p(c|\mu, \sigma, \beta) = \prod_{j=1}^{J} \int_{\theta_c^{(j)}} p(\theta_c^{(j)}|\mu_c^{(j)}, \sigma_c^{(j)}) \cdot \left(\prod_{k=1}^{n_j} \prod_{w} p(w|\beta, \theta_c^{(j)})\right) d\theta_c^{(j)}. \tag{13}$$

Inspired by the success of VAE-based NTMs (Miao, Grefenstette, and Blunsom 2017; Dieng, Ruiz, and Blei 2020), we also employ a VAE framework for the utterance-level BoW reconstruction process. The posterior global topic distribution $p(\theta_c^{(j)})$ for each speaker j can be approximated by the inference network $q(\theta_c^{(j)}|\mu_c^{(j)}, \sigma_c^{(j)})$. We can formulate parameter updates from the variational evidence lower bound (ELBO). From the perspective of ELBO, the training objective for the log-likelihood of the conversation consists of two terms. The first term is to minimize the cross entropy between the input normalized BoW and reconstructed BoW, and the second Kullback–Leibler (KL) divergence term is to minimize the distance between the variational posterior and true posterior of latent variables. This part of the training loss can be formulated as:

$$\mathcal{L}_c^{(j)} = -\mathbb{E}_{q(\theta_c^{(j)}|\mu_c^{(j)}, \sigma_c^{(j)})}\left(\sum_{k=1}^{n_j} \sum_{w} \log p(w|\theta_c^{(j)}, \beta)\right) + w_{kl} \cdot D_{KL}(q(\theta_c^{(j)}|\mu_c^{(j)}, \sigma_c^{(j)}) || p(\theta_c^{(j)})), \tag{14}$$

where w_{kl} is the hyper-parameter for the weight of the KL term.

Controllable word co-occurrence objective. In addition to the ELBO commonly used in general NTMs, we further leverage the word co-occurrence information of the training corpus to improve the topic quality. For the topic-word distribution matrix $\beta \in \mathbb{R}^{K \times |V|}$, its i-th row represents a multinomial distribution on the i-th topic over the vocabulary V. We expect that the top words in each topic are highly correlated and tend to co-occur in the same real conversations. Thus, we count the co-occurrence frequencies of all word pairs in all conversations in the training corpus, and construct a co-occurrence matrix $M \in \mathbb{R}^{|V| \times |V|}$. Next, we add such a constraint on β, which can be described as the following loss:

$$\mathcal{L}_{co} = -\sum_{w_1=1}^{|V|} \sum_{w_2=1}^{|V|} M_{w_1,w_2} \log(\beta^{\mathrm{T}}\beta)_{w_1,w_2}. \tag{15}$$

Intuitively, we make the β-derived matrix as close as possible to the reference co-occurrence matrix M. We set a target co-occurrence distance as d_{co}, and then design a controllable weight w_{co} for the trade-off between $\mathcal{L}_c$ and $\mathcal{L}_{co}$. Suppose that there are C conversations in the training set, the overall training loss of ConvNTM is given by:

$$\mathcal{L} = (1 - w_{co}) \sum_{c=1}^{C} \sum_{j=1}^{J} \mathcal{L}_c^{(j)} + w_{co}\mathcal{L}_{co}. \tag{16}$$

The controllable factor w_{co} is dynamically adjusted as:

$$w_{co} = \begin{cases} 0, & \mathcal{L}_{co} \le d_{co}, \\ \min\left(1, \dfrac{\mathcal{L}_{co} - d_{co}}{W_{co}}\right), & \mathcal{L}_{co} > d_{co}, \end{cases} \tag{17}$$

where W_{co} is another hyper-parameter of the correcting factor for the proportional signal.

Experiments

Experimental Setup

Datasets. We conduct the experiments on two widely used multi-turn dialogue datasets, DailyDialog[1] and EmpatheticDialogues[2]. DailyDialog (Li et al. 2017) totally contains 13,118 high-quality open-domain daily conversations, and covers various topics about daily life. It has 7.9 average speaker turns per conversation, and each speaker has enough utterances for multi-turn modeling. We use the official splits, i.e., 11,118/1,000/1,000. EmpatheticDialogues (Rashkin et al. 2019) contains about 25k personal conversations with rich emotional expressions and topic situations. Speakers discuss emotional topics and tend to interact with empathy. We also employ the official splits data, i.e. 19,533/2,770/2,547 for train/val/test respectively.

Evaluation metrics. To evaluate the quality of topics generated by topic models, we adopt topic coherence (TC) and topic diversity (TD) metrics. TC measures the semantic consistency of top words within each topic. A higher TC metric indicates more relevant keywords within each topic and better topic interpretability. Following the previous work (Shen et al. 2021), we choose two TC measurements, CV and normalized pointwise mutual information (NPMI), to provide a robust evaluation. The NPMI of the word pair (w_i, w_j) is calculated as equation (18). CV score stands for a widely used Content Vector-based coherence metric, adopted by (Röder, Both, and Hinneburg 2015). Both of these TC metrics can be obtained in the gensim library (Rehurek and Sojka 2011). TD measures the diversity across different topics. It is defined as the percentage of unique words among the top words. A higher TD metric indicates more topic variability. Pursuing either a high TC value or a high TD value independently does not guarantee the topic quality. Inspired by (Dieng, Ruiz, and Blei 2020), we regard CV as the TC score and measure the topic quality score (TQ) as the product of TC and TD.

$$\mathrm{NPMI}(w_i, w_j) = \frac{\log \frac{p(w_i,w_j)+\epsilon}{P(w_i)P(w_j)}}{-\log(p(w_i, w_j) + \epsilon)}. \tag{18}$$

Baselines. We compare our model with the mainstream and state-of-the-art topic models as baselines. The baselines include: 1) **LDA** (Blei, Ng, and Jordan 2003), the most representative statistical topic model using Gibbs sampling; 2) **GSM** (Miao, Grefenstette, and Blunsom 2017), a VAE-based NTM introducing Gaussian softmax for generating latent variables; 3) **ProdLDA** (Srivastava and Sutton 2017), an NTM constructing Laplace approximation to the Dirichlet prior; 4) **ETM** (Dieng, Ruiz, and Blei 2020), an NTM projecting topics and words into the same embedding space; 5) **GNTM** (Shen et al. 2021), a recent NTM designing a document graph and introducing it into the generative process of topic modeling. For all baselines, we employ their officially reported parameter settings.

Implementation details. For the multi-role interaction graph, we set the window sizes K_s and K_c to 2. The BoW dictionary size is set to 6,500 in DailyDialog and 7,533 in EmpatheticDialogues. The embedding size and hidden size of the Transformer, LSTM and GCN are all set to 64. For the loss function, w_{kl} and W_{co} are set to 0.01 and 0.05, while the value of d_{co} is determined by the number of topics and the dataset. In our main results, d_{co} is recommended to be set to 32 in DailyDialog and 31.375 in EmpatheticDialogues. The training process has 100 epoches using the Adam optimizer with the base learning rate of 0.001. We implement the experiments on a Nvidia A40 GPU.[3]

Main Results

For all baselines, one conversation is treated as one document for topic modeling. Here we set the number of topics to 20, and analyze the impact of the number of topics later. To properly evaluate the learned topics, we follow the previous works (Kim et al. 2012; Shen et al. 2021) and select the top 10 words with the highest probability under each topic as the representative word list to calculate topic quality metrics. The comparison results are available in Table 1. Our ConvNTM outperforms all baselines on two TC metrics (i.e. CV and NPMI) on two datasets, which indicates that with the help of formulating the specific multi-turn and multi-role information in the conversation, the topics discovered by ConvNTM have the best topic interpretability. GNTM achieves the highest on TD, while ConvNTM is slightly behind. This reason may be that GNTM generates words and edges based

[1] http://yanran.li/dailydialog

[2] https://github.com/facebookresearch/EmpatheticDialogues

[3] Our code and data are available at https://github.com/ssshddd/ConvNTM.

Dataset	DailyDialog				EmpatheticDialogues			
Method	**TD**	**CV**	**NPMI**	**TQ**	**TD**	**CV**	**NPMI**	**TQ**
LDA	0.390	0.4308	-0.0083	0.1680	0.510	0.4230	0.0011	0.2158
GSM	0.445	0.4931	-0.0040	0.2194	0.530	0.4486	0.0055	0.2378
ProdLDA	0.720	0.5363	-0.0007	0.3861	0.736	0.4610	0.0173	0.3393
ETM	0.690	0.5688	0.0364	0.3925	0.713	0.4690	0.0130	0.3342
GNTM	**0.810**	0.5916	0.0588	0.4792	**0.812**	0.4809	0.0289	0.3905
ConvNTM	0.750	**0.6542**	**0.0831**	**0.4907**	0.790	**0.5136**	**0.0495**	**0.4057**

Table 1: Comparison results of topic quality on DailyDialog and EmpatheticDialogues.

on topics at the same time, which may indirectly increase the sparsity among topic proportions. ETM and ProdLDA also have moderate TC metrics, but their TD is relatively low, which is prone to generate redundant topics on the conversation dataset. Comprehensively considering the impact of TC and TD, our ConvNTM which integrates multiple turns and speaker roles can achieve state-of-the-art performance on the TQ score.

Ablation Study

In order to verify the effectiveness of key modules of our model, we compare ConvNTM with the following four model variants: 1) **ConvNTM (w/o contexts)** removes the conversation sequence encoder used to model multi-turn dialogue contexts; 2) **ConvNTM (w/o graph)** removes the multi-role graph encoder used to model interactions between speakers; 3) **ConvNTM (w/o speaker)** sets the number of speakers to 1 that completely ignores the effect of the roles; 4) **ConvNTM (w/o $\mathcal{L}_{co}$)** remove the loss term $\mathcal{L}_{co}$ for the word co-occurrence objective.

Table 2 shows the comparison results of these different ablation methods on DailyDialog. Compared with the full model, both ConvNTM (w/o contexts) and ConvNTM (w/o graph) decrease on TC and TD, indicating that both the multi-turn context structure and multi-role interaction information of the conversation have a significant impact on the topic quality. The performance of ConvNTM (w/o speaker) is further degraded when the speaker's role is not modeled and the utterances in the conversation are treated as sentences in the general document. This reflects the superiority of ConvNTM over general NTMs for topic modeling on the unique properties of the conversation. In addition, when removing the word co-occurrence training objective, ConvNTM (w/o $\mathcal{L}_{co}$) improves slightly on TD, while it drops more significantly on TC, making the overall topic quality worse. It means that considering word-occurrence information can help improve the coherence and interpretability of learned topics.

Method	TD	TC	NPMI	TQ
ConvNTM (w/o contexts)	0.715	0.6240	0.0619	0.4462
ConvNTM (w/o graph)	0.705	0.6282	0.0657	0.4429
ConvNTM (w/o speaker)	0.650	0.6099	0.0548	0.3964
ConvNTM (w/o $\mathcal{L}_{co}$)	**0.780**	0.6237	0.0645	0.4865
ConvNTM	0.750	**0.6542**	**0.0831**	**0.4907**

Table 2: Ablation results for ConvNTM on DailyDialog.

Methods	Top Words
ConvNTM	food time eat great pizza rice love dinner restaurant cake friend happy family birthday nice year home people party kid work hard boss job coworker project today manager lot late car happen bad hurt drive traffic terrible police left accident
GNTM	food eat restaurant **people** pizza **day** real kid tonight birthday **work** week hear nervous family **help day** ago remember job home man left terrible **friend** stop door police **people help** car money new drive thing **people work friend** vote happen

Figure 2: Visualization of an example for discovered topics (one topic per line). Repeated words are in bold.

Analysis on Discovered Topic Examples

We also perform a qualitative analysis on discovered topics, comparing ConvNTM and the strong baseline GNTM. Figure 2 shows several representative topics learned by ConvNTM and GNTM. We display the top 10 words under each topic per line. For our ConvNTM, we can see that the top words in each line have strong associations and focus on a certain topic. This means that each learned topic has good internal coherence. The selected 4 topics can be summarized as *food*, *family & friends*, *work*, and *traffic accidents*. Meanwhile, ConvNTM has fewer repeated words, indicating less redundancy in the learned topics. While for GNTM, these topic words are mixed together, and some non-topic words are repeated in different topics. For instance, "people" are shown in multiple topics, and "work" and "family" appear in the same topic in GNTM, which destroys the topic diversity, coherence and interpretability.

Analysis on Number of Topics

Since the number of topics is an important factor of the topic model, we compare the topic quality performance of ConvNTM and several strong baselines. We set the varying number of topics from 10 to 100, and the comparison results are shown in Figure 3. Our ConvNTM achieves the highest TC and TQ under all number of topics, which indicates the robustness of our method on topic quality. All models have high topic quality when the number of topics is between 20 and 50. When the number of topics exceeds

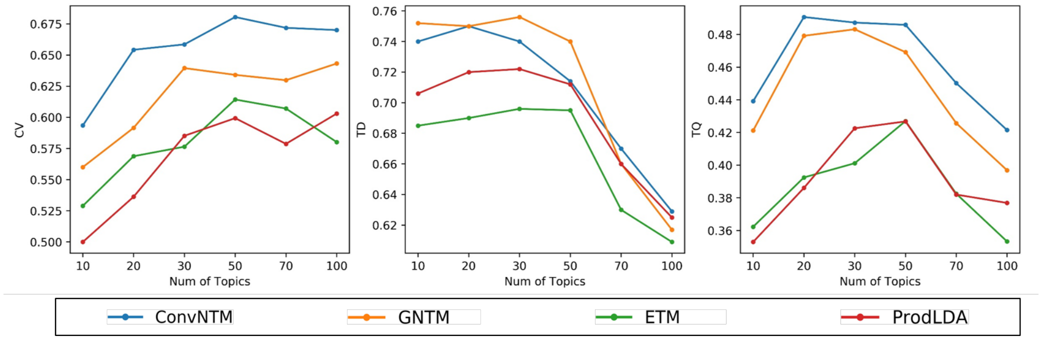

Figure 3: Comparison results of the varying number of topics on DailyDialog.

Method	**Accuracy**
JAS	75.9
DAH-CRF+MANUALconv	86.5
DAH-CRF+LDAconv	86.4
DAH-CRF+LDAuttr	88.1
STM	87.1
STM+GNTM	87.2
STM+ConvNTM	**88.9**

Table 3: Comparison results of topic-aware models for the dialogue act classification task on DailyDialog.

Method	**PPL ↓**	**BLEU-1 ↑**	**Distinct-1 ↑**
HERD	41.38	6.40	4.42
TA-Seq2Seq	38.98	15.84	6.79
DAWnet	39.36	16.90	7.78
THERD+LDA	36.46	18.26	7.90
THERD+GNTM	36.68	18.53	8.26
THERD+ConvNTM	**34.14**	**20.14**	**8.79**

Table 4: Comparison results of topic-aware models for the dialogue response generation task on DailyDialog.

50, TC tends to be stable or slightly decreases, and TD decreases significantly. ConvNTM can achieve the highest TD when the number of topics is large, and hold the best topic quality under any number of topics.

Downstream Tasks

The essence of topic modeling is an unsupervised learning process for latent semantic structure, and we expect that not only ConvNTM can achieve state-of-the-art topic quality, but we can leverage the topic information learned by ConvNTM to help improve downstream dialogue tasks. Here, we take dialogue act classification and response generation as examples to verify that ConvNTM is helpful for improving both classification and generation tasks. Specifically, we choose GNTM as a strong baseline and respectively add topic information learned by GNTM and ConvNTM into topic-aware models for comparison.

We use different topic extraction approaches for different tasks. For dialogue act classification, we borrow the framework of (He et al. 2021b) (named STM), which utilized topic labels for each utterance when modeling speaker turns. We extract the topic labels using our ConvNTM and GNTM for comparison, and replace original topic labels with them. We also compare other topic-aware models in this task including JAS (Wallace et al. 2013) and DAH (Li et al. 2019). The comparison results on DailyDialog are shown in Table 3. This indicates that ConvNTM can indeed help improve this task and it performs better than all topic-aware baselines and GNTM. For dialogue response generation, we borrow the framework of THERD (Dziri et al. 2019), which proposes a topical hierarchical recurrent framework for multi-turn response generation. THERD utilizes LDA to extract the top 100 topic words for each conversation. Here LDA can be directly replaced by GNTM to extract topic words. While for our ConvNTM, we first label all the utterances of a conversation, and then extract the top 100 words with the highest probability under these topics. We also compare other topic-aware models in this task including HERD (Serban et al. 2016b), TA-Seq2Seq (Xing et al. 2017) and DAWnet (Wang et al. 2018). The comparison results on DailyDialog are shown in Table 4. THERD+ConvNTM can achieve better performance than all topic-aware baselines and GNTM on multiple metrics.

Conclusion

In this work, we propose the first Conversational Neural Topic Model (ConvNTM) specifically for the conversation scenario. We develop a hierarchical conversation encoder to capture the multi-turn dialogue structure. Considering the impact of roles of different speakers in a conversation, we construct a multi-role interaction graph to formulate the intra-speaker and inter-speaker dependencies. We then perform utterance-level fine-grained topic modeling by fusing global and local topic information. Furthermore, we leverage the word co-occurrence relationship as a new training objective, which can be jointly trained with the neural variational inference objective and further improve topic quality.

Acknowledgements

This work was supported by National Natural Science Foundation of China (NSFC Grant No. 62122089 and No. 61876196), Beijing Outstanding Young Scientist Program NO. BJJWZYJH012019100020098, and Intelligent Social Governance Platform, Major Innovation & Planning Interdisciplinary Platform for the "Double-First Class" Initiative, Renmin University of China. We also wish to acknowledge the support provided and contribution made by Public Policy and Decision-making Research Lab of RUC. Rui Yan is supported by Beijing Academy of Artificial Intelligence (BAAI) and CCF-Tencent Rhino-Bird Open Research Fund.

References

Adiwardana, D.; Luong, M.-T.; So, D. R.; Hall, J.; Fiedel, N.; Thoppilan, R.; Yang, Z.; Kulshreshtha, A.; Nemade, G.; Lu, Y.; et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Blei, D. M.; Ng, A. Y.; and Jordan, M. I. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan): 993–1022.

Chen, C.; and Ren, J. 2017. Forum latent Dirichlet allocation for user interest discovery. *Knowledge-Based Systems*, 126: 1–7.

Cheng, X.; Yan, X.; Lan, Y.; and Guo, J. 2014. Btm: Topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, 26(12): 2928–2941.

Dieng, A. B.; Ruiz, F. J.; and Blei, D. M. 2020. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8: 439–453.

Dieng, A. B.; Wang, C.; Gao, J.; and Paisley, J. 2017. TopicRNN: A Recurrent Neural Network with Long-Range Semantic Dependency. In *International Conference on Learning Representations*.

Dziri, N.; Kamalloo, E.; Mathewson, K.; and Zaiane, O. 2019. Augmenting Neural Response Generation with Context-Aware Topical Attention. In *Proceedings of the First Workshop on NLP for Conversational AI*, 18–31. Florence, Italy: Association for Computational Linguistics.

Gu, J.-C.; Li, T.; Liu, Q.; Ling, Z.-H.; Su, Z.; Wei, S.; and Zhu, X. 2020. Speaker-aware BERT for multi-turn response selection in retrieval-based chatbots. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, 2041–2044.

He, Z.; Tavabi, L.; Lerman, K.; and Soleymani, M. 2021a. Speaker Turn Modeling for Dialogue Act Classification. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 2150–2157. Punta Cana, Dominican Republic: Association for Computational Linguistics.

He, Z.; Tavabi, L.; Lerman, K.; and Soleymani, M. 2021b. Speaker Turn Modeling for Dialogue Act Classification. *arXiv preprint arXiv:2109.05056*.

Hofmann, T. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, 50–57.

Holtgraves, T.; Srull, T. K.; and Socall, D. 1989. Conversation memory: The effects of speaker status on memory for the assertiveness of conversation remarks. *Journal of Personality and Social Psychology*, 56(2): 149.

Jin, Y.; Zhao, H.; Liu, M.; Du, L.; and Buntine, W. 2021. Neural Attention-Aware Hierarchical Topic Model. *arXiv preprint arXiv:2110.07161*.

Kim, H.; Sun, Y.; Hockenmaier, J.; and Han, J. 2012. Etm: Entity topic models for mining documents associated with entities. In *2012 IEEE 12th International Conference on Data Mining*, 349–358. IEEE.

Kim, T.; and Vossen, P. 2021. Emoberta: Speaker-aware emotion recognition in conversation with roberta. *arXiv preprint arXiv:2108.12009*.

Kingma, D. P.; and Welling, M. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

Lang, K. 1995. Newsweeder: Learning to filter netnews. In *Machine Learning Proceedings 1995*, 331–339. Elsevier.

Larochelle, H.; and Lauly, S. 2012. A neural autoregressive topic model. *Advances in Neural Information Processing Systems*, 25.

Li, J.; Liao, M.; Gao, W.; He, Y.; and Wong, K.-F. 2016. Topic Extraction from Microblog Posts Using Conversation Structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2114–2123. Berlin, Germany: Association for Computational Linguistics.

Li, R.; Lin, C.; Collinson, M.; Li, X.; and Chen, G. 2019. A Dual-Attention Hierarchical Recurrent Neural Network for Dialogue Act Classification. In *CoNLL*.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of The 8th International Joint Conference on Natural Language Processing (IJCNLP 2017)*.

Li, Z.; Zhang, J.; Fei, Z.; Feng, Y.; and Zhou, J. 2021. Conversations Are Not Flat: Modeling the Dynamic Information Flow across Dialogue Utterances. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 128–138. Online: Association for Computational Linguistics.

Lin, T.; Hu, Z.; and Guo, X. 2019. Sparsemax and relaxed Wasserstein for topic sparsity. In *Proceedings of the twelfth ACM international conference on web search and data mining*, 141–149.

Liu, J.; Zou, Y.; Zhang, H.; Chen, H.; Ding, Z.; Yuan, C.; and Wang, X. 2021. Topic-Aware Contrastive Learning for Abstractive Dialogue Summarization. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, 1229–1243. Punta Cana, Dominican Republic: Association for Computational Linguistics.

Ma, X.; Zhang, Z.; and Zhao, H. 2021. Enhanced Speaker-aware Multi-party Multi-turn Dialogue Comprehension. *arXiv preprint arXiv:2109.04066*.

Miao, Y.; Grefenstette, E.; and Blunsom, P. 2017. Discovering discrete latent topics with neural variational inference. In *International Conference on Machine Learning*, 2410–2419. PMLR.

Panwar, M.; Shailabh, S.; Aggarwal, M.; and Krishnamurthy, B. 2020. TAN-NTM: Topic attention networks for neural topic modeling. *arXiv preprint arXiv:2012.01524*.

Qiu, L.; Zhao, Y.; Shi, W.; Liang, Y.; Shi, F.; Yuan, T.; Yu, Z.; and Zhu, S.-C. 2020a. Structured Attention for Unsupervised Dialogue Structure Induction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1889–1899. Online: Association for Computational Linguistics.

Qiu, L.; Zhao, Y.; Shi, W.; Liang, Y.; Shi, F.; Yuan, T.; Yu, Z.; and Zhu, S.-C. 2020b. Structured attention for unsupervised dialogue structure induction. *arXiv preprint arXiv:2009.08552*.

Rashkin, H.; Smith, E. M.; Li, M.; and Boureau, Y.-L. 2019. Towards Empathetic Open-domain Conversation Models: a New Benchmark and Dataset. In *ACL*.

Rehurek, R.; and Sojka, P. 2011. Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2).

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

Serban, I. V.; García-Durán, A.; Gulcehre, C.; Ahn, S.; Chandar, S.; Courville, A.; and Bengio, Y. 2016a. Generating Factoid Questions With Recurrent Neural Networks: The 30M Factoid Question-Answer Corpus. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 588–598. Berlin, Germany: Association for Computational Linguistics.

Serban, I. V.; Sordoni, A.; Bengio, Y.; Courville, A.; and Pineau, J. 2016b. Building End-to-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI'16, 3776–3783. AAAI Press.

Shen, D.; Qin, C.; Wang, C.; Dong, Z.; Zhu, H.; and Xiong, H. 2021. Topic Modeling Revisited: A Document Graph-based Neural Network Perspective. *Advances in Neural Information Processing Systems*, 34: 14681–14693.

Srivastava, A.; and Sutton, C. 2016. Neural variational inference for topic models. *ArXiv Preprint*, 1(1): 1–12.

Srivastava, A.; and Sutton, C. 2017. Autoencoding Variational Inference For Topic Models. In *International Conference on Learning Representations*.

Sun, Y.; Loparo, K.; and Kolacinski, R. 2020. Conversational structure aware and context sensitive topic model for online discussions. In *2020 IEEE 14th International Conference on Semantic Computing (ICSC)*, 85–92. IEEE.

Wallace, B. C.; Trikalinos, T. A.; Laws, M. B.; Wilson, I. B.; and Charniak, E. 2013. A generative joint, additive, sequential model of topics and speech acts in patient-doctor communication. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, 1765–1775.

Wang, R.; Zhou, D.; and He, Y. 2019. Atm: Adversarial-neural topic model. *Information Processing & Management*, 56(6): 102098.

Wang, W.; Huang, M.; Xu, X.-S.; Shen, F.; and Nie, L. 2018. Chat more: Deepening and widening the chatting topic via a deep model. In *The 41st international acm sigir conference on research & development in information retrieval*, 255–264.

Xing, C.; Wu, W.; Wu, Y.; Liu, J.; Huang, Y.; Zhou, M.; and Ma, W.-Y. 2017. Topic Aware Neural Response Generation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence*, AAAI'17, 3351–3357. AAAI Press.

Zhang, Y.; Sun, S.; Galley, M.; Chen, Y.-C.; Brockett, C.; Gao, X.; Gao, J.; Liu, J.; and Dolan, B. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.

Zhao, H.; Phung, D.; Huynh, V.; Le, T.; and Buntine, W. 2021. Neural Topic Model via Optimal Transport. In *International Conference on Learning Representations*.

Zhu, L.; Pergola, G.; Gui, L.; Zhou, D.; and He, Y. 2021. Topic-Driven and Knowledge-Aware Transformer for Dialogue Emotion Detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 1571–1582. Online: Association for Computational Linguistics.

Zhu, Q.; Feng, Z.; and Li, X. 2018. GraphBTM: Graph enhanced autoencoded variational inference for biterm topic model. In *Conference on Empirical Methods in Natural Language Processing (EMNLP 2018)*.