What You Say and How You Say it: Joint Modeling of Topics and Discourse in Microblog Conversations

Jichuan Zeng[1]* **Jing Li**[2]* **Yulan He**[3] **Cuiyun Gao**[1] **Michael R. Lyu**[1] **Irwin King**[1]

[1]Department of Computer Science and Engineering
The Chinese University of Hong Kong, HKSAR, China
[2]Tencent AI Lab, Shenzhen, China
[3]Department of Computer Science, University of Warwick, UK
[1]{jczeng, cygao, lyu, king}@cse.cuhk.edu.hk
[2]ameliajli@tencent.com, [3]yulan.he@warwick.ac.uk

Abstract

This paper presents an *unsupervised* framework for jointly modeling topic content and discourse behavior in microblog conversations. Concretely, we propose a *neural* model to discover word clusters indicating what a conversation concerns (i.e., *topics*) and those reflecting how participants voice their opinions (i.e., *discourse*).[1] Extensive experiments show that our model can yield both coherent topics and meaningful discourse behavior. Further study shows that our topic and discourse representations can benefit the classification of microblog messages, especially when they are jointly trained with the classifier.

1 Introduction

The last decade has witnessed the revolution of communication, where the ‘‘kitchen table conversations’’ have been expanded to public discussions on online platforms. As a consequence, in our daily life, the exposure to new information and the exchange of personal opinions have been mediated through microblogs, one popular online platform genre (Bakshy et al., 2015). The flourish of microblogs has also led to the sheer quantity of user-created conversations emerging every day, exposing individuals to superfluous information. Facing such an unprecedented number of conversations relative to limited attention of individuals, how shall we automatically extract the critical points and make sense of these microblog conversations?

Toward key focus understanding of a conversation, previous work has shown the benefits of discourse structure (Li et al., 2016b; Qin et al., 2017; Li et al., 2018), which shapes how messages interact with each other, forming the discussion flow, and can usefully reflect salient topics raised in the discussion process. After all, the topical content of a message naturally occurs in context of the conversation discourse and hence should not be modeled in isolation. Conversely, the extracted topics can reveal the purpose of participants and further facilitate the understanding of their discourse behavior (Qin et al., 2017). Further, the joint effects of topics and discourse will contribute to better understanding of social media conversations, benefiting downstream tasks such as the management of discussion topics and discourse behavior of social chatbots (Zhou et al., 2018) and the prediction of user engagements for conversation recommendation (Zeng et al., 2018b).

To illustrate how the topics and discourse interplay in a conversation, Figure 1 displays a snippet of Twitter conversation. As can be seen, the content words reflecting the discussion topics (such as ‘‘*supreme court*’’ and ‘‘*gun rights*’’) appear in context of the discourse flow, where participants carry the conversation forward via making a statement, giving a comment, asking a question, and so forth. Motivated by such an observation, we assume that *a microblog conversation can be decomposed into two crucially different components: one for topical content and the other for discourse behavior*. Here, the topic components indicate what a conversation is centered around and reflect the important discussion points put forward in the conversation process. The discourse components signal the **discourse roles** of messages, such as making

[*] This work was partially conducted in Jichuan Zeng’s internship in Tencent AI Lab. Corresponding author: Jing Li.

[1] Our data sets and code are available at: `http://github.com/zengjichuan/Topic_Disc`.

Transactions of the Association for Computational Linguistics, vol. 7, pp. 267–281, 2019. Action Editor: Jianfeng Gao.
Submission batch: 12/2018; Revision batch: 2/2019; Published 5/2019.

...
M$_1$ [*Statement*]: Just watched **HRC** openly endorse a **gun-control measure** which will fail in front of the **Supreme Court**. This is a train wreck.
M$_2$ [*Comment*]: People said the same thing about **Obama**, and nothing took place. **Gun laws** just aren't being enforced like they should be. :/
M$_3$ [*Question*]: Okay, hold up. What do you think I'm referencing here? It's not what you're talking about.
M$_4$ [*Agreement*]: Thought it was about **gun control**. I'm in agreement that **gun rights** shouldn't be stripped.
...

Figure 1: A Twitter conversation snippet about the gun control issue in U.S. **Topic words** reflecting the conversation focus are in boldface. The *italic* words in [ ] are our interpretations of the messages' discourse roles.

a statement, asking a question, and other dialogue acts (Ritter et al., 2010; Joty et al., 2011), which further shape the discourse structure of a conversation.[2] To distinguish the above two components, we examine the conversation contexts and identify two types of words: **topic words**, indicating what a conversation focuses on, and **discourse words**, reflecting how the opinion is voiced in each message. For example, in Figure 1, the topic words "*gun*" and "*control*" indicate the conversation topic while the discourse word "*what*" and "*?*" signal the question in M$_3$.

Concretely, we propose a neural framework built upon topic models, enabling the joint exploration of word clusters to represent topic and discourse in microblog conversations. Different from the prior models trained on annotated data (Li et al., 2016b; Qin et al., 2017), our model is fully unsupervised, not dependent on annotations for either topics or discourse, which ensures its immediate applicability in any domain or language. Moreover, taking advantages of the recent advances in neural topic models (Srivastava and Sutton, 2017; Miao et al., 2017), we are able to approximate Bayesian variational inference without requiring model-specific derivations, whereas most existing work (Ritter et al., 2010; Joty et al., 2011; Alvarez-Melis and Saveski, 2016; Zeng et al., 2018b; Li et al., 2018) require expertise involved to customize model inference algorithms. In addition, our neural nature enables end-to-end training of topic and discourse representation learning with other neural models for diverse tasks.

For model evaluation, we conduct an extensive empirical study on two large-scale Twitter data sets. The intrinsic results show that our model can produce latent topics and discourse roles with better interpretability than the state-of-the-art models from previous studies. The extrinsic evaluations on a tweet classification task exhibit the model's ability to capture useful representations for microblog messages. Particularly, our model enables an easy combination with existing neural models for end-to-end training, such as convolutional neural networks, which is shown to perform better in classification than the pipeline approach without joint training.

2 Related Work

Our work is in the line with previous studies that use *non-neural* models to leverage discourse structure for extracting topical content from conversations (Li et al., 2016b; Qin et al., 2017; Li et al., 2018). Zeng et al. (2018b) explore how discourse and topics jointly affect user engagements in microblog discussions. Different from them, we build our model in a *neural network* framework, where the joint effects of topic and discourse representations can be exploited for various downstream deep learning tasks in an end-to-end manner. In addition, we are inspired by prior research that only models topics or conversation discourse. In the following, we discuss them in turn.

Topic Modeling. Our work is closely related with the topic model studies. In this field, despite the huge success achieved by the springboard topic models (e.g., pLSA [Hofmann, 1999] and LDA [Blei et al., 2001]), and their extensions (Blei et al., 2003; Rosen-Zvi et al., 2004), the applications of these models have been limited to formal and well-edited documents, such as news reports (Blei et al., 2003) and scientific articles (Rosen-Zvi et al., 2004), attributed to their reliance on document-level word collocations. When processing short texts, such as the messages on microblogs, it is likely that the performance of these models will be inevitably compromised, due to the severe data sparsity issue.

To deal with such an issue, many previous efforts incorporate the *external* representations, such as word embeddings (Nguyen et al., 2015; Li et al.,

[2] In this paper, the discourse role refers to a certain type of dialogue act (e.g., *statement* or *question*) for each message. And the discourse structure refers to some combination of discourse roles in a conversation.

2016a; Shi et al., 2017) and knowledge (Song et al., 2011; Yang et al., 2015; Hu et al., 2016), pre-trained on large-scale high-quality resources. Different from them, our model learns topic and discourse representations only with the internal data and thus can be widely applied on scenarios where the specific external resource is unavailable.

In another line of the research, most prior work focuses on how to enrich the context of short messages. To this end, biterm topic model (BTM) (Yan et al., 2013) extends a message into a biterm set with all combinations of any two distinct words appearing in the message. On the contrary, our model allows the richer context in a conversation to be exploited, where word collocation patterns can be captured beyond a short message.

In addition, there are many methods using some heuristic rules to aggregate short messages into long pseudo-documents, such as those based on authorship (Hong and Davison, 2010; Zhao et al., 2011) and hashtags (Ramage et al., 2010; Mehrotra et al., 2013). Compared with these methods, we model messages in the context of their conversations, which has been demonstrated to be a more natural and effective text aggregation strategy for topic modeling (Alvarez-Melis and Saveski, 2016).

Conversation Discourse. Our work is also in the area of discourse analysis for conversations, ranging from the prediction of the shallow discourse roles on utterance level (Stolcke et al., 2000; Ji et al., 2016; Zhao et al., 2018) to the discourse parsing for a more complex conversation structure (Elsner and Charniak, 2008, 2010; Afantenos et al., 2015). In this area, most existing models heavily rely on the data annotated with discourse labels for learning (Zhao et al., 2017). Different from them, our model, in a fully unsupervised way, identifies distributional word clusters to represent latent discourse factors in conversations. Although such latent discourse variables have been studied in previous work (Ritter et al., 2010; Joty et al., 2011; Ji et al., 2016; Zhao et al., 2018), none of them explores the effects of latent discourse on the identification of conversation topic, which is a gap our work fills in.

3 Our Neural Model for Topics and Discourse in Conversations

This section introduces our neural model that jointly explores latent representations for topics and discourse in conversations. We first present an overview of our model in Section 3.1, followed by the model generative process and inference procedure in Section 3.2 and 3.3, respectively.

3.1 Model Overview

In general, our model aims to learn coherent word clusters that reflect the latent topics and discourse roles embedded in the microblog conversations. To this end, we distinguish two latent components in the given collection: *topics* and *discourse*, each represented by a certain type of word distribution (distributional word cluster). Specifically, at the corpus level, we assume that there are K topics, represented by ϕ^T_k $(k = 1, 2, \ldots, K)$, and D discourse roles, captured with ϕ^D_d $(d = 1, 2, \ldots, D)$. ϕ^T and ϕ^D are all multinomial word distributions over the vocabulary size V. Inspired by the neural topic models in Miao et al. (2017), our model encodes topic and discourse distributions (ϕ^T and ϕ^D) as latent variables in a neural network and learns the parameters via back propagation.

Before touching the details of our model, we first describe how we formulate the input. On microblogs, as a message might have multiple replies, messages in an entire conversation can be organized as a tree with replying relations (Li et al., 2016b, 2018). Though the recent progress in recursive models allows the representation learning from the tree-structured data, previous studies have pointed out that, in practice, sequence models serve as a more simple yet robust alternative (Li et al., 2015). In this work, we follow the common practice in most conversation modeling research (Ritter et al., 2010; Joty et al., 2011; Zhao et al., 2018) to take a conversation as a sequence of turns. To this end, each conversation tree is flattened into root-to-leaf paths. Each one of such paths is hence considered as a conversation instance, and a message on the path corresponds to a conversation turn (Zarisheva and Scheffler, 2015; Cerisara et al., 2018; Jiao et al., 2018).

The overall architecture of our model is shown in Figure 2. Formally, we formulate a conversation $\mathbf{c}$ as a sequence of messages $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_{M_c})$, where M_c denotes the number of messages in $\mathbf{c}$. In the conversation, each message $\mathbf{x}$, as the **target message**, is fed into our model sequentially. Here we process the target message $\mathbf{x}$ as the bag-of-words (BoW) term vector $\mathbf{x}_{BoW} \in \mathbb{R}^V$, following the bag-of-words assumption in most

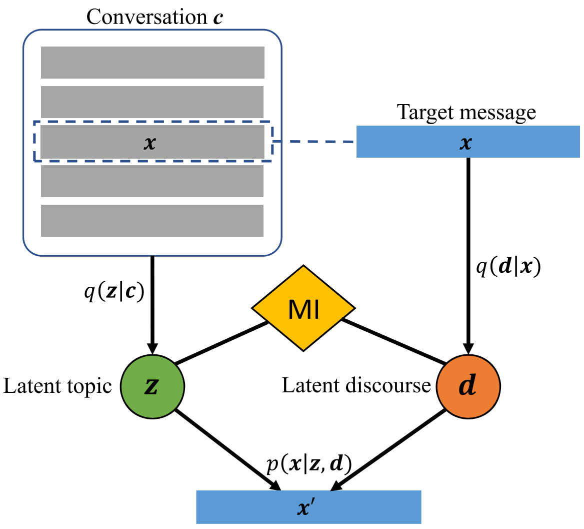


Figure 2: The architecture of our neural framework. The latent topics **z** and latent discourse **d** are jointly modeled from conversation **c** and target message **x**, respectively. Mutual information penalty (MI) is used to separate words clusters representing topics and discourse. Afterward, **z** and **d** are used to reconstruct the target message $\mathbf{x}'$.

topic models (Blei et al., 2003; Miao et al., 2017). The conversation, **c**, where the target message **x** is involved, is considered as the context of **x**. It is also encoded in the BoW form (denoted as $\mathbf{c}_{BoW} \in \mathbb{R}^V$) and fed into our model. In doing so, we ensure that the context of the target message is incorporated while learning its latent representations.

Following the previous practice in neural topic models (Miao et al., 2017; Srivastava and Sutton, 2017), we utilize the variational auto-encoder (VAE) (Kingma and Welling, 2013) to resemble the data generative process via two steps. First, given the target message **x** and its conversation **c**, our model converts them into two latent variables: topic variable **z** and discourse variable **d**. Then, using the intermediate representations captured by **z** and **d**, we reconstruct the target message, $\mathbf{x}'$.

3.2 Generative Process

In this section, we first describe the two latent variables in our model: the topic variable **z** and the discourse variable **d**. Then, we present our data generative process from the latent variables.

Latent Topics. For latent topic learning, we examine the main discussion points in the context of a conversation. Our assumption is that messages in the same conversation tend to focus on similar topics (Li et al., 2018; Zeng et al., 2018b). Concretely, we define the latent topic variable $\mathbf{z} \in \mathbb{R}^K$ at the *conversation* level and generate the topic mixture of **c**, denoted as a K-dimensional distribution θ, via a softmax construction conditioned on **z** (Miao et al., 2017).

Latent Discourse. For modeling the discourse structure of conversations, we capture the *message*-level discourse roles reflecting the dialogue acts of each message, as is done in Ritter et al. (2010). Concretely, given the target message **x**, we use a D-dimensional one-hot vector to represent the latent discourse variable **d**, where the high bit indicates the index of a discourse word distribution that can best express **x**'s discourse role. In the generative process, the latent discourse **d** is drawn from a multinomial distribution with parameters estimated from the input data.

Data Generative Process As mentioned previously, our entire framework is based on VAE, which consists of an encoder and a decoder. The encoder maps a given input into latent topic and discourse representations and the decoder reconstructs the original input from the latent representations. In the following, we first describe the decoder followed by the encoder.

In general, our *decoder* is learned to reconstruct the words in the target message **x** (in the BoW form) from the latent topic **z** and latent discourse **d**. We show the generative story that reflects the reconstruction process below:

- Draw the latent topic $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma}^2)$
- **c**'s topic mixture $\theta = \mathrm{softmax}(f_\theta(\mathbf{z}))$
- Draw the latent discourse $\mathbf{d} \sim Multi(\boldsymbol{\pi})$
- For the n-th word in **x**
  - $\beta_n = \mathrm{softmax}(f_{\phi^T}(\boldsymbol{\theta}) + f_{\phi^D}(\mathbf{d}))$
  - Draw the word $w_n \sim Multi(\beta_n)$

where $f_*(\cdot)$ is a neural perceptron, with a linear transformation of inputs activated by a non-linear transformation. Here we use rectified linear units (Nair and Hinton, 2010) as the activate functions. In particular, the weight matrix of $f_{\phi^T}(\cdot)$ (after the softmax normalization) is considered as the topic-word distributions ϕ^T. The discourse-word distributions ϕ^D are similarly obtained from $f_{\phi^D}(\cdot)$.

For the *encoder*, we learn the parameters $\boldsymbol{\mu}$, $\boldsymbol{\sigma}$, and $\boldsymbol{\pi}$ from the input $\mathbf{x}_{BoW}$ and $\mathbf{c}_{BoW}$ (the BoW form of the target message and its conversation), following the following formula:

$$\begin{gathered}\boldsymbol{\mu} = f_\mu(f_e(\mathbf{c}_{BoW})), \log \boldsymbol{\sigma} = f_\sigma(f_e(\mathbf{c}_{BoW})) \\ \boldsymbol{\pi} = \text{softmax}(f_\pi(\mathbf{x}_{BoW}))\end{gathered} \tag{1}$$

3.3 Model Inference

For the objective function of our entire framework, we take three aspects into account: the learning of latent topics and discourse, the reconstruction of the target messages, and the separation of topic-associated words and discourse-related words.

Learning Latent Topics and Discourse. For learning the latent topics/discourse in our model, we utilize the variational inference (Blei et al., 2016) to approximate posterior distribution over the latent topic $\mathbf{z}$ and the latent discourse $\mathbf{d}$ given all the training data. To this end, we maximize the variational lower bound $\mathcal{L}_z$ for $\mathbf{z}$ and $\mathcal{L}_d$ for $\mathbf{d}$, each defined as following:

$$\begin{aligned}\mathcal{L}_z &= \mathbb{E}_{q(\mathbf{z}\,|\,\mathbf{c})}[p(\mathbf{c}\,|\,\mathbf{z})] - D_{KL}(q(\mathbf{z}\,|\,\mathbf{c})\,||\,p(\mathbf{z})) \\ \mathcal{L}_d &= \mathbb{E}_{q(\mathbf{d}\,|\,\mathbf{x})}[p(\mathbf{x}\,|\,\mathbf{d})] - D_{KL}(q(\mathbf{d}\,|\,\mathbf{x})\,||\,p(\mathbf{d}))\end{aligned} \tag{2}$$

$q(\mathbf{z}\,|\,\mathbf{c})$ and $q(\mathbf{d}\,|\,\mathbf{x})$ are approximated posterior probabilities describing how the latent topic $\mathbf{z}$ and the latent discourse $\mathbf{d}$ are generated from the data. $p(\mathbf{c}\,|\,\mathbf{z})$ and $p(\mathbf{x}\,|\,\mathbf{d})$ represent the corpus likelihoods conditioned on the latent variables. Here, to facilitate coherent topic production, in $p(\mathbf{c}\,|\,\mathbf{z})$, we penalize the likelihood of stopwords to be generated from latent topics following Li et al. (2018). $p(\mathbf{z})$ follows the standard normal prior $\mathcal{N}(\mathbf{0}, \mathbf{I})$ and $p(\mathbf{d})$ is the uniform distribution $Unif(0, 1)$. D_{KL} refers to the Kullback-Leibler divergence that ensures the approximated posteriors to be close to the true ones. For more derivation details, we refer readers to Miao et al. (2017).

Reconstructing target messages. From the latent variables $\mathbf{z}$ and $\mathbf{d}$, the goal of our model is to reconstruct the target message $\mathbf{x}$. The corresponding learning objective is to maximize $\mathcal{L}_x$ defined as:

$$\mathcal{L}_x = \mathbb{E}_{q(z\,|\,\mathbf{x})q(d\,|\,\mathbf{c})}[\log p(\mathbf{x}\,|\,\mathbf{z}, \mathbf{d})] \tag{3}$$

Here we design $\mathcal{L}_x$ to ensure that the learned latent topics and discourse can reconstruct $\mathbf{x}$.

Distinguishing Topics and Discourse. Our model aims to distinguish word distributions for topics (ϕ^T) and discourse (ϕ^D), which enables topics and discourse to capture different information in conversations. Concretely, we use the mutual information, given below, to measure the mutual dependency between the latent topics $\mathbf{z}$ and the latent discourse $\mathbf{d}$.[3]

$$\mathbb{E}_{q(\mathbf{z})q(\mathbf{d})}\left[\log \frac{p(\mathbf{z}, \mathbf{d})}{p(\mathbf{z})p(\mathbf{d})}\right] \tag{4}$$

Equation 4 can be further derived as the Kullback-Leibler divergence of the conditional distribution, $p(\mathbf{d}\,|\,\mathbf{z})$, and marginal distribution, $p(\mathbf{d})$. The derived formula, defined as the mutual information (MI) loss ($\mathcal{L}_{MI}$) and shown in Equation 5, is used to map $\mathbf{z}$ and $\mathbf{d}$ into the separated semantic space.

$$\mathcal{L}_{MI} = \mathbb{E}_{q(\mathbf{z})}[D_{KL}(p(\mathbf{d}\,|\,\mathbf{z})||p(\mathbf{d}))] \tag{5}$$

We can hence minimize $\mathcal{L}_{MI}$ for guiding our model to separate word distributions that represent topics and discourse.

The Final Objective. To capture the joint effects of the learning objectives described above ($\mathcal{L}_z$, $\mathcal{L}_d$, $\mathcal{L}_x$, and $\mathcal{L}_{MI}$), we design the final objective function for our entire framework as the following:

$$\mathcal{L} = \mathcal{L}_z + \mathcal{L}_d + \mathcal{L}_x - \lambda \mathcal{L}_{MI} \tag{6}$$

where the hyperparameter λ is the trade-off parameter for balancing between the MI loss ($\mathcal{L}_{MI}$) and the other learning objectives. By maximizing the final objective $\mathcal{L}$ via back propagation, the word distributions of topics and discourse can be jointly learned from microblog conversations.[4]

4 Experimental Setup

Data Collection. For our experiments, we collected two microblog conversation data sets from Twitter. One is released by the TREC 2011 microblog track (henceforth **TREC**), containing conversations concerning a wide range of topics.[5]

[3] The distributions in Equation 4 are all conditional probability distributions given the target message $\mathbf{x}$ and its conversation $\mathbf{c}$. We omit the conditions for simplicity.

[4] To smooth the gradients in implementation, for $\mathbf{z} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\sigma})$, we apply the reparameterization on $\mathbf{z}$ (Kingma and Welling, 2013; Rezende et al., 2014), and for $\mathbf{d} \sim Multi(\boldsymbol{\pi})$, we adopt the Gumbel-Softmax trick (Maddison et al., 2016; Jang et al., 2016).

[5] `http://trec.nist.gov/data/tweets/`.

Data sets	# of convs	Avg msgs per conv	Avg words per msg	|Vocab|
TREC	116,612	3.95	11.38	9,463
TWT16	29,502	8.67	14.70	7,544

Table 1: Statistics of the two data sets containing Twitter conversations.

The other is crawled from January to June 2016 with Twitter streaming API[6] (henceforth **TWT16**, short for Twitter 2016), following the way of building the TREC data set. During this period, there are a large volume of discussions centered around the U.S. presidential election. In addition, for both data sets, we apply Twitter search API[7] to retrieve the missing tweets in the conversation history, as the Twitter streaming API (used to collect both data sets) only returns sampled tweets from the entire pool.

The statistics of the two experiment data sets are shown in Table 1. For model training and evaluation, we randomly sampled 80%, 10%, and 10% of the data to form the training, development, and test set, respectively.

Data Preprocessing. We preprocessed the data with the following steps. First, non-English tweets were filtered out. Then, hashtags, mentions (@username), and links were replaced with generic tags ‘‘HASH’’, ‘‘MENT’’, and ‘‘URL’’, respectively. Next, the natural languge toolkit was applied for tweet tokenization.[8] After that, all letters were normalized to lower cases. Finally, words that occurred fewer than 20 times were filtered out from the data.

Parameter Setting. To ensure comparable results with Li et al. (2018) (the prior work focusing on the same task as ours), in the topic coherence evaluation, we follow their setup to report the results under two sets of K (the number of topics): $K = 50$ and $K = 100$, and with the number of discourse roles (D) set to 10. The analysis for the effects of K and D will be further presented in Section 5.5. For all the other hyper-parameters, we tuned them on development set by grid search. The trade-off parameter λ (defined in Equation 6), balancing the MI loss and the other objective functions, is set to 0.01. In model training, we use the Adam optimizer (Kingma and Ba, 2014) and run 100 epochs with early stop strategy adopted.

Baselines. In topic modeling experiments, we consider the five topic model baselines treating each tweet as a document: LDA (Blei et al., 2003), BTM (Yan et al., 2013), LF-LDA, LF-DMM (Nguyen et al., 2015), and NTM (Miao et al., 2017). In particular, BTM and LF-DMM are the state-of-the-art topic models for short texts. BTM explores the topics of all word pairs (biterms) in each message to alleviate data sparsity in short texts. LF-DMM incorporates word embeddings pre-trained on external data to expand semantic meanings of words, so does LF-LDA. In Nguyen et al. (2015), LF-DMM, based on one-topic-per-document Dirichlet Multinomial Mixture (DMM) (Nigam et al., 2000), was reported to perform better than LF-LDA, based on LDA. For LF-LDA and LF-DMM, we use GloVe Twitter embeddings (Pennington et al., 2014) as the pre-trained word embeddings.[9]

For the discourse modeling experiments, we compare our results with LAED (Zhao et al., 2018), a VAE-based representation learning model for conversation discourse. In addition, for both topic and discourse evaluation, we compare with Li et al. (2018), a recently proposed model for microblog conversations, where topics and discourse are jointly explored with a *non-neural* framework. Besides the existing models from previous studies, we also compare with the variants of our model that only models topics (henceforth TOPIC ONLY) or discourse (henceforth DISC ONLY).[10] Our joint model of topics and discourse is referred to as TOPIC+DISC.

In the preprocessing procedure for the baselines, we removed stop words and punctuation for topic models unable to learn discourse representations following the common practice in previous work (Yan et al., 2013; Miao et al., 2017). For the other models, stop words and punctuation were retained in the vocabulary, considering their usefulness as discourse indicators (Li et al., 2018).

[6] https://developer.twitter.com/en/docs/tweets/filter-realtime/api-reference/post-statuses-filter.html.

[7] https://developer.twitter.com/en/docs/tweets/search/api-reference/get-savedsearches-show-id.

[8] https://www.nltk.org/.

[9] https://nlp.stanford.edu/projects/glove/.

[10] In our ablation without mutual information loss ($\mathcal{L}_{MI}$ defined in Equation 4), topics and discourse are learned independently. Thus, its topic representation can be used for the output of TOPIC ONLY, so does its discourse one for DISC ONLY.

Models	$K = 50$		$K = 100$	
	TREC	TWT16	TREC	TWT16
Baselines				
LDA	0.467	0.454	0.467	0.454
BTM	0.460	0.461	0.466	0.463
LF-DMM	0.456	0.448	0.463	0.466
LF-LDA	0.470	0.456	0.467	0.453
NTM	0.478	0.479	0.482	0.443
Li et al. (2018)	0.463	0.433	0.464	0.435
Our models				
Topic only	0.478	0.482	0.481	0.471
Topic+Disc	**0.485**	**0.487**	**0.496**	**0.480**

Table 2: C_v coherence scores for latent topics produced by different models. The best result in each column is highlighted in **bold**. Our joint model Topic+Disc achieves significantly better coherence scores than all the baselines ($p < 0.01$, paired test).

5 Experimental Results

In this section, we first report the topic coherence results in Section 5.1, followed by a discussion in Section 5.2 comparing the latent discourse roles discovered by our model with the manually annotated dialogue acts. Then, we study whether we can capture useful representations for microblog messages in a tweet classification task (in Section 5.3). A qualitative analysis, showing some example topics and discourse roles, is further provided in Section 5.4. Finally, in Section 5.5, we provide more discussions on our model.

5.1 Topic Coherence

For the topic coherence, we adopt the C_v scores measured via the open-source Palmetto toolkit as our evaluation metric.[11] C_v scores assume that the top N words in a coherent topics (ranked by likelihood) tend to co-occur in the same document and have shown comparable evaluation results to human judgments (Röder et al., 2015). Table 2 shows the average C_v scores over the produced topics given $N = 5$ and $N = 10$. The values range from 0.0 to 1.0, and higher scores indicate better topic coherence. We can observe that:

• ***Models assuming a single topic for each message do not work well.*** It has long been pointed out that the one-topic-per-message assumption (each message contains only one topic) helps topic models alleviate the data sparsity issue in short texts on microblogs (Zhao et al., 2011; Quan et al., 2015; Nguyen et al., 2015; Li et al., 2018). However, we observe contradictory results because both LF-DMM and Li et al. (2018), following this assumption, achieve generally worse performance than the other models. This might be attributed to the large-scale data used in our experiments (each data set has over 250K messages as shown in Table 1), which potentially provide richer word co-occurrence patterns and thus partially alleviate the data sparsity issue.

• ***Pre-trained word embeddings do not bring benefits.*** Comparing LF-LDA with LDA, we found that they result in similar coherence scores. This shows that with sufficiently large training data, with or without using the pre-trained word embeddings do not make any difference in the topic coherence results.

• ***Neural models perform better than non-neural baselines.*** When comparing the results of neural models (NTM and our models) with the other baselines, we find the former yield topics with better coherence scores in most cases.

• ***Modeling topics in conversations is effective.*** Among neural models, we found our models outperform NTM (without exploiting conversation contexts). This shows that the conversations provide useful context and enables more coherent topics to be extracted from the entire conversation thread instead of a single short message.

• ***Modeling topics together with discourse helps produce more coherent topics.*** We can observe better results with the joint model Topic+Disc in comparison with the variant considering topics only. This shows that Topic+Disc, via the joint modeling of topic- and discourse-word distributions (reflecting non-topic information), can better separate topical words from non-topical ones, hence resulting in more coherent topics.

5.2 Discourse Interpretability

In this section, we evaluate whether our model can discover meaningful discourse representations. To this end, we train the comparison models for discourse modeling on the TREC data set and test the learned latent discourse on a benchmark data set released by Cerisara et al. (2018). The benchmark data set consists of 2,217 microblog messages forming 505 conversations collected

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

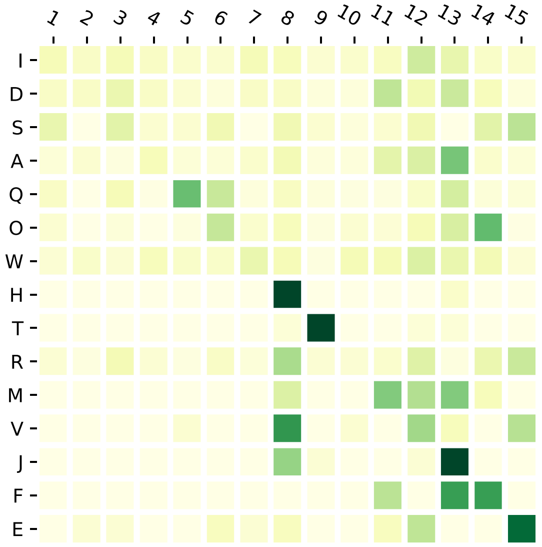
Models	Purity	Homogeneity	VI
Baselines			
LAED	0.505	0.022	6.418
Li et al. (2018)	0.511	0.096	5.540
Our models			
DISC ONLY	0.510	0.112	5.532
TOPIC+DISC	**0.521**	**0.142**	**5.097**

Table 3: The purity, homogeneity, and variation of information (VI) scores for the latent discourse roles measured against the human-annotated dialogue acts. For purity and homogeneity, higher scores indicate better performance, while for VI scores, lower is better. In each column, the best results are in **boldface**. Our joint model TOPIC+DISC significantly outperforms all the baselines ($p < 0.01$, paired t-test).

from Mastodon,[12] a microblog platform exhibiting Twitter-like user behavior (Cerisara et al., 2018). For each message, there is a human-assigned discourse label, selected from one of the 15 dialogue acts, such as *question*, *answer*, *disagreement*, and so forth.

For discourse evaluation, we measure whether the model-produced discourse assignments are consistent with the human-annotated dialogue acts. Hence, following Zhao et al. (2018), we assume that an interpretable latent discourse role should cluster messages labeled with the same dialogue act. Therefore, we adopt purity (Manning et al., 2008), homogeneity (Rosenberg and Hirschberg, 2007), and variation of informa tion (VI) (Meila, 2003; Goldwater and Griffiths, 2007) as our automatic evaluation metrics. Here, we set $D = 15$ to ensure the number of latent discourse roles to be the same as the number of manually labeled dialogue acts. Table 3 shows the comparison results of the average scores over the 15 latent discourse roles. Higher values indicate better performance for purity and homogeneity, while for VI, lower is better.

It can be observed that our models exhibit generally better performance, showing the effectiveness of our framework in inducing interpretable discourse roles. Particularly, we observe the best results achieved by our joint model TOPIC+DISC, which is learned to distinguish topic- and discourse-words, important in recognizing indicative words to reflect latent discourse.



I: statement, D: disagreement, S: suggest, A: agreement, Q: yes/no question, O: wh*/open question, W: open+choice answer, H: initial greetings, T: thanking, R: request, M: sympathy, V: explicit performance, J: exclamation, F: acknowledge, and E: offer.

Figure 3: A heatmap showing the alignments of the latent discourse roles and human-annotated dialogue act labels. Each line visualizes the distribution of messages with the corresponding dialogue act label over varying discourse roles (indexed from 1 to 15), where darker colors indicate higher values.

To further analyze the consistency of varying latent discourse roles (produced by our TOPIC+DISC model) with the human-labeled dialogue acts, Figure 3 displays a heatmap, where each line visualizes how the messages with a dialogue act distribute over varying discourse roles. It is seen that among all dialogue acts, our model discovers more interpretable latent discourse for ‘‘*greetings*’’, ‘‘*thanking*’’, ‘‘*exclamation*’’, and ‘‘*offer*’’, where most messages are clustered into one or two dominant discourse roles. It may be because these dialogue acts can be relatively easier to detect based on their associated indicative words, such as the word ‘‘*thanks*’’ for ‘‘*thanking*’’, and the word ‘‘*wow*’’ for ‘‘*exclamation*’’.

5.3 Message Representations

To further evaluate our ability to capture effective representations for microblog messages, we take tweet classification as an example and test the classification performance with the topic and discourse representations as features. Here, the user-generated hashtags capturing the topics of online messages are used as the proxy class labels (Li et al., 2016b; Zeng et al., 2018a). We construct

[12] https://mastodon.social.

Models	TREC		TWT16	
	Acc	Avg F1	Acc	Avg F1
Baselines				
BoW	0.120	0.026	0.132	0.030
TF-IDF	0.116	0.024	0.153	0.041
LDA	0.128	0.041	0.146	0.046
BTM	0.123	0.035	0.167	0.054
LF-DMM	0.158	0.072	0.162	0.052
NTM	0.138	0.042	0.186	0.068
Our model	**0.259**	**0.180**	**0.341**	**0.269**

Table 4: Evaluation of tweet classification results in accuracy (Acc) and average F1 (Avg F1). Representations learned by various models serve as the classification features. For our model, both the topic and discourse representations are fed into the classifier.

the classification data set from TREC and TWT16 with the following steps. First, we removed the tweets without hashtags. Second, we ranked hashtags by their frequencies. Third, we manually removed the hashtags that are not topic-related (e.g. ‘‘*#fb*’’ for indicating the source of tweets from Facebook), and combined the hashtags referring to the same topic (e.g., ‘‘*#DonaldTrump*’’ and ‘‘*#Trump*’’). Finally, we selected the top 50 frequent hashtags, and all tweets containing these hashtags as our classification data set. Here, we simply use the support vector machines as the classifier, since our focus is to compare the representations learned by various models. Li et al. (2018) are unable to produce vector representation on tweet level, hence not considered here.

Table 4 shows the classification results of accuracy and average F1 on the two data sets with the representations learned by various models serving as the classification features. We observe that our model outperforms other models with a large margin. The possible reasons are twofold. First, our model derives topics from conversation threads and thus potentially yields better message representations. Second, the discourse representations (only produced by our model) are indicative features for hashtags, because users will exhibit various discourse behaviors in discussing diverse topics (hashtags). For instance, we observe prominent ‘‘*argument*’’ discourse from tweets with ‘‘*#Trump*’’ and ‘‘*#Hillary*’’, attributed to the controversial opinions to the two candidates in the 2016 U.S. presidential election.

5.4 Example Topics and Discourse Roles

We have shown that joint modeling of topics and discourse presents superior performance on a quantitative measure. In this section, we qualitatively analyze the interpretability of our outputs via analyzing the word distributions of some example topics and discourse roles.

Example Topics. Table 5 lists the top 10 words of some example latent topics discovered by various models from the TWT16 data set. According to the words shown, we can interpret the extracted topics as ‘‘gun control’’ — discussion about gun law and the failure of gun control in Chicago. We observe that LDA wrongly includes off-topic word ‘‘*flag*’’. From the outputs of BTM, LF-DMM, Li et al. (2018), and our TOPIC ONLY variant, though we do not find off-topic words, there are some non-topic words, such as ‘‘*said*’’ and ‘‘*understand*’’.[13] The output of our TOPIC+DISC model appears to be the most coherent, with words such as ‘‘*firearm*’’ and ‘‘*criminals*’’ included, which are clearly relevant to ‘‘gun control’’. Such results indicate the benefit of examining the conversation contexts and jointly exploring topics and discourse in them.

Example Discourse Roles. To qualitatively analyze whether our TOPIC+DISC model can discover interpretable discourse roles, we select the top 10 words from the distributions of some example discourse roles and list them in Table 6. It can be observed that there are some meaningful word clusters reflecting varying discourse roles found without any supervision. Interestingly, we observe that the latent discourse roles from TREC and TWT16, though learned separately, exhibit some notable overlap in their associated top 10 words, particularly for ‘‘*question*’’ and ‘‘*statement*’’. We also note that ‘‘*argument*’’ is represented by very different words. The reason is that TWT16 contains a large volume of arguments centered around candidates Clinton and Trump, resulting in the frequent appearance of words like ‘‘*he*’’ and ‘‘*she*’’.

5.5 Further Discussions

In this section, we further present more discussions on our joint model: TOPIC+DISC.

Parameter Analysis. Here, we study the two important hyper-parameters in our model, the

[13] Non-topic words do not clearly indicate the corresponding topic, whereas off-topic words are more likely to appear in other topics.

LDA	people trump police violence gun death protest guns flag shot
BTM	gun guns people police wrong right think law agree black
LF-DMM	gun police black said people guns killing ppl amendment laws
Li et al. (2018)	wrong don trump gun understand laws agree guns doesn make
NTM	gun understand yes guns world dead real discrimination trump silence
TOPIC ONLY	shootings gun guns cops charges control mass commit know agreed
TOPIC+DISC	guns gun shootings chicago shooting cops firearm criminals commit laws

Table 5: Top 10 representative words of example latent topics discovered from the TWT16 data set. We interpret the topics as ‘‘gun control’’ by the displayed words. Non-topic words are wave-underlined and in blue, and off-topic words are underlined and in red.

Discourse Roles	**TREC**	**TWT16**
Question	was what why is how that like ? ?? you	? why what MENT do does it the to did
Response	! love ha !! you saw lmao lol awesome !!!	doin uhhh ❗ awards yay joseph 😖 👋 muted
Agreement	okaay thankss wateva okayy txtd twitcam entertained havee goooood darlin	! you are agree re to they we with their
Quotation	& ' < > (... feat “ ” ")	» « (< ┆ ┆ < MENT .< ,- - ?< "
Statement	to will ! the be rt my in on and	will have if do be can want vote should ?
Argument	fuck damn rt lmfao hair girl thing lmao ass bitch	😂 he said him she her but wrong did never

Table 6: Top 10 representative words of example discourse roles learned from TREC and TWT16. The discourse roles of the word clusters are manually assigned according to their associated words.

number of topics (K) and the number of discourse roles (D). In Figure 4, we show the C_v topic coherence given varying K in (a) and the homogeneity measure given varying D in (b). As can be seen, the curves corresponding to the performance on topics and discourse are not monotonic. In particular, better topic coherence scores are achieved given relatively larger topic numbers for TREC with the best result observed at $K = 80$. On the contrary, the optimum topic number for TWT16 is $K = 20$, and increasing the number of topics results in worse C_v scores in general. This may be attributed to the relatively centralized topic concerning U.S. election in the TWT16 corpus. For discourse homogeneity, the best result is achieved given $D = 15$, with same the number of manually annotated dialogue acts in the benchmark.

Case Study. To further understand why our model learns meaningful representations for topics and discourse, we present a case study based on the example conversation shown in Figure 1. Specifically, we visualize the topic words (with $p(w \mid \mathbf{z}) > p(w \mid \mathbf{d})$) in red and the rest of the words in blue to indicate discourse. Darker red indicates the higher topic likelihood ($p(w \mid \mathbf{z})$) and darker blue shows the higher discourse likelihood ($p(w \mid \mathbf{d})$). The results are shown in Figure 5. We can observe that topic and discourse words are well separated by our model, which explains why it can generate high-quality representations for both topics and discourse.

Model Extensibility. Recall that in the Introduction, we mentioned that our neural-based model has an advantage to be easily combined with other neural network architectures and allows for the joint training of both models. Here, we take message classification (with the setup in Section 5.3) as an example, and study whether

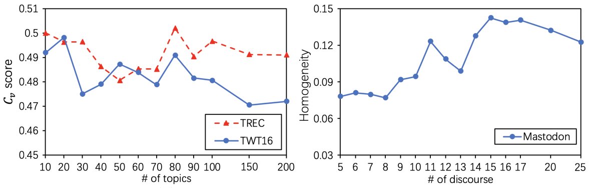
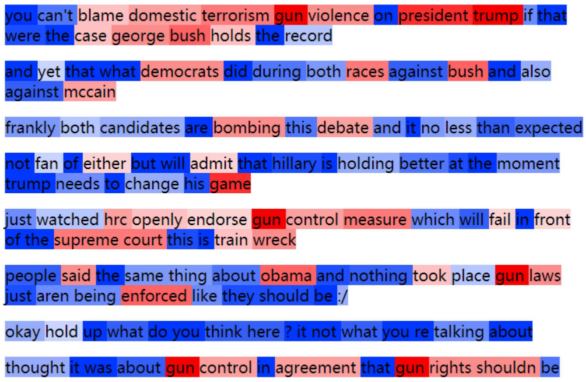


Figure 4: (a) The impact of topic numbers. The horizontal axis shows the number of topics; the vertical axis shows the C_v topic coherence. (b) The impact of discourse numbers. The horizontal axis represents the number of discourse; the vertical axis represents the homogeneity measure.


you can't blame domestic terrorism gun violence on president trump if that were the case george bush holds the record

and yet that what democrats did during both races against bush and also against mccain

frankly both candidates are bombing this debate and it no less than expected

not fan of either but will admit that hillary is holding better at the moment trump needs to change his game

just watched hrc openly endorse gun control measure which will fail in front of the supreme court this is train wreck

people said the same thing about obama and nothing took place gun laws just aren being enforced like they should be :/

okay hold up what do you think here ? it not what you re talking about

thought it was about gun control in agreement that gun rights shouldn be

Figure 5: Visualization of the topic-discourse assignment of a twitter conversion from TWT16. The annotated blue words are prone to be discourse words, and the red are topic words. The shade indicates the confidence of the current assignment.

joint training our model with convolutional neural network (CNN) (Kim, 2014), the widely used model on short text classification, can bring benefits to the classification performance. We set the embedding dimension to 200, with random initialization. The results are shown in Table 7, where we observe that joint training our model and the classifier can successfully boost the classification performance.

Error Analysis. We further analyze the errors in our outputs. For topics, taking a closer look at their word distributions, we found that our model sometimes mixes sentiment words with topic words. For example, among the top 10 words of a topic ‘‘*win people illegal americans hate lt racism social tax wrong*’’, there are words ‘‘*hate*’’ and ‘‘*wrong*’’, expressing sentiment rather than conveying topic-related information. This is due to the prominent co-occurrences of topic words and sentiment words in our data, which results in the similar distributions for topics and sentiment. Future work could focus on the further separation of sentiment and topic words.

For discourse, we found that our model can induce some discourse roles beyond the 15 manually defined dialogue acts in the Mastodon data set (Cerisara et al., 2018). For example, as shown in Table 6, our model discovers the ‘‘*quotation*’’ discourse from both TREC and TWT16, which is, however, not defined in the Mastodon data set. This perhaps should not be considered as an error. We argue that it is not sensible to pre-define a fixed set of dialogue acts for diverse microblog conversations due to the rapid change and a wide variety of user behaviors in social media. Therefore, future work should involve a better alternative to evaluate the latent discourse without relying on manually defined dialogue acts. We also notice that our model sometimes fails to identify discourse behaviors requiring more in-depth semantic understanding, such as sarcasm, irony, and humor. This is because our model detects latent discourse purely based on the observed words, whereas the detection of sarcasm, irony, or humor requires deeper language understanding, which is beyond the capacity of our model.

6 Conclusion and Future Work

We have presented a neural framework that jointly explores topic and discourse from microblog conversations. Our model, in an unsupervised manner, examines the conversation contexts and discovers word distributions that reflect latent topics and discourse roles. Results from extensive experiments show that our model can generate coherent topics and meaningful discourse roles. In addition, our model can be easily combined with other neural network architectures (such as CNN) and allows for joint training, which has presented better message classification results compared with the pipeline approach without joint training.

Our model captures topic and discourse representations embedded in conversations. They are potentially useful for a broad range of downstream applications, worthy to be explored in future research. For example, our model is useful for developing social chatbots (Zhou et al., 2018). By explicitly modeling ‘‘*what you say*’’ and ‘‘*how you say*’’, our model can be adapted to track the

Models	TREC		TWT16	
	Acc	Avg F1	Acc	Avg F1
CNN only	0.199	0.167	0.334	0.311
Separate-Train	0.284	0.270	0.391	0.390
Joint-Train	**0.297**	**0.286**	**0.428**	**0.413**

Table 7: Accuracy (Acc) and average F1 (Avg F1) on tweet classification (hashtags as labels). CNN only: CNN without using our representations. Seperate-Train: CNN fed with our pre-trained representations. Joint-Train: Joint training CNN and our model.

change of topics in conversation context, helpful to determine ‘‘*what to say and how to say*’’ in the next turn. Also, it would be interesting to study how our learned latent topics and discourse affect recommendation (Zeng et al., 2018b) and summarization of microblog conversations (Li et al., 2018).

Acknowledgments

This work is partially supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (No. CUHK 14208815 and No. CUHK 14210717 of the General Research Fund), Innovate UK (grant no. 103652), and Microsoft Research Asia (2018 Microsoft Research Asia Collaborative Research Award). We thank Shuming Shi, Dong Yu, and TACL reviewers and editors for the insightful suggestions on various aspects of this work.

References

Stergos D. Afantenos, Eric Kow, Nicholas Asher, and Jérémy Perret. 2015. Discourse parsing for multi-party chat dialogues. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 928–937. Lisbon.

David Alvarez-Melis and Martin Saveski. 2016. Topic modeling in twitter: Aggregating tweets by conversations. In *Proceedings of the Tenth International Conference on Web and Social Media*, pages 519–522. Cologne.

Eytan Bakshy, Solomon Messing, and Lada A. Adamic. 2015. Exposure to ideologically diverse news and opinion on facebook. *Science*, 348(6239):1130–1132.

David M. Blei, Thomas L. Griffiths, Michael I. Jordan, and Joshua B. Tenenbaum. 2003. Hierarchical topic models and the nested chinese restaurant process. In *Advances in Neural Information Processing Systems, NIPS 2003*, pages 17–24. Vancouver and Whistler.

David M. Blei, Alp Kucukelbir, and Jon D. McAuliffe. 2016. Variational inference: A review for statisticians. *CoRR*, abs/1601.00670.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2001. Latent Dirichlet allocation. In *Advances in Neural Information Processing Systems 14*, pages 601–608. Vancouver.

Christophe Cerisara, Somayeh Jafaritazehjani, Adedayo Oluokun, and Hoa T. Le. 2018. Multi-task dialog act and sentiment recognition on mastodon. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018*, pages 745–754. Santa Fe, NM.

Micha Elsner and Eugene Charniak. 2008. You talking to me? A corpus and algorithm for conversation disentanglement. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, ACL 2008*, pages 834–842. Columbus, OH.

Micha Elsner and Eugene Charniak. 2010. Disentangling chat. *Computational Linguistics*, 36(3): 389–409.

Sharon Goldwater and Thomas L. Griffiths. 2007. A fully Bayesian approach to unsupervised part-of-speech tagging. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*. Prague.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 50–57. Berkeley, CA.

Liangjie Hong and Brian D. Davison. 2010. Empirical study of topic modeling in Twitter. In *Proceedings of the 3rd Workshop on Social Network Mining and Analysis, SNAKDD 2009*, pages 80–88. Paris.

Zhiting Hu, Gang Luo, Mrinmaya Sachan, Eric P. Xing, and Zaiqing Nie. 2016. Grounding topic models with knowledge bases. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016*, pages 1578–1584. New York, NY.

Eric Jang, Shixiang Gu, and Ben Poole. 2016. Categorical reparameterization with gumbel-softmax. *CoRR*, abs/1611.01144.

Yangfeng Ji, Gholamreza Haffari, and Jacob Eisenstein. 2016. A latent variable recurrent neural network for discourse-driven language models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL HLT 2016*, pages 332–342. San Diego, CA.

Yunhao Jiao, Cheng Li, Fei Wu, and Qiaozhu Mei. 2018. Find the conversation killers: A predictive study of thread-ending posts. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018*, pages 1145–1154. Lyon.

Shafiq R. Joty, Giuseppe Carenini, and Chin-Yew Lin. 2011. Unsupervised modeling of dialog acts in asynchronous conversations. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence, IJCAI 2011*, pages 1807–1813. Barcelona.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751. Doha.

Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Diederik P. Kingma and Max Welling. 2013. Auto-encoding variational Bayes. *CoRR*, abs/1312.6114.

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

Jing Li, Ming Liao, Wei Gao, Yulan He, and Kam-Fai Wong. 2016b. Topic extraction from microblog posts using conversation structures. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, Volume 1: Long Papers*. Berlin.

Jing Li, Yan Song, Zhongyu Wei, and Kam-Fai Wong. 2018. A joint model of conversational discourse and latent topics on microblogs. *Computational Linguistics*, 44(4):719–754.

Jiwei Li, Thang Luong, Dan Jurafsky, and Eduard H. Hovy. 2015. When are tree structures necessary for deep learning of representations? In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 2304–2314. Lisbon.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. 2016. The concrete distribution: A continuous relaxation of discrete random variables. *CoRR*, abs/1611.00712.

Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze. 2008. *Introduction to Information Retrieval*. Cambridge University Press.

Rishabh Mehrotra, Scott Sanner, Wray L. Buntine, and Lexing Xie. 2013. Improving LDA topic models for microblogs via tweet pooling and automatic labeling. In *The 36th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '13*, pages 889–892. Dublin.

Marina Meila. 2003. Comparing clusterings by the variation of information. In *Computational Learning Theory and Kernel Machines, 16th Annual Conference on Computational Learning Theory and 7th Kernel Workshop, COLT/Kernel, Proceedings*, pages 173–187. Washington, DC.

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

Vinod Nair and Geoffrey E. Hinton. 2010. Rectified linear units improve restricted Boltzmann machines. In *Proceedings of the 27th International*

Conference on Machine Learning (ICML 2010), pages 807–814. Haifa.

Dat Quoc Nguyen, Richard Billingsley, Lan Du, and Mark Johnson. 2015. Improving topic models with latent feature word representations. *Transactions of the Association for Computational Linguistics, TACL*, 3:299–313.

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

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 1532–1543. Doha.

Kechen Qin, Lu Wang, and Joseph Kim. 2017. Joint modeling of content and discourse relations in dialogues. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017*, Volume 1: Long Papers, pages 974–984. Vancouver.

Xiaojun Quan, Chunyu Kit, Yong Ge, and Sinno Jialin Pan. 2015. Short and sparse text topic modeling via self-aggregation. In *Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence, IJCAI 2015*, pages 2270–2276. Buenos Aires.

Daniel Ramage, Susan T. Dumais, and Daniel J. Liebling. 2010. Characterizing microblogs with topic models. In *Proceedings of the Fourth International Conference on Weblogs and Social Media, ICWSM 2010*, Washington, DC.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31th International Conference on Machine Learning, ICML 2014*, pages 1278–1286. Beijing.

Alan Ritter, Colin Cherry, and Bill Dolan. 2010. Unsupervised modeling of twitter conversations. In *Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings*, pages 172–180. Los Angeles, CA.

Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the Space of Topic Coherence Measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining, WSDM 2015*, pages 399–408. Shanghai.

Michal Rosen-Zvi, Thomas L. Griffiths, Mark Steyvers, and Padhraic Smyth. 2004. The author-topic model for authors and documents. In *UAI '04, Proceedings of the 20th Conference Uncertainty in Artificial Intelligence*, pages 487–494. Banff.

Andrew Rosenberg and Julia Hirschberg. 2007. V-measure: A conditional entropy-based external cluster evaluation measure. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, EMNLP-CoNLL 2007*, pages 410–420. Prague.

Bei Shi, Wai Lam, Shoaib Jameel, Steven Schockaert, and Kwun Ping Lai. 2017. Jointly learning word embeddings and latent topics. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 375–384. Tokyo.

Yangqiu Song, Haixun Wang, Zhongyuan Wang, Hongsong Li, and Weizhu Chen. 2011. Short text conceptualization using a probabilistic knowledgebase. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence, IJCAI 2011*, pages 2330–2336. Barcelona.

Akash Srivastava and Charles Sutton. 2017. Autoencoding variational inference for topic models. In *Proceedings of the Fifth International Conference on Learning Representations, ICLR 2017*. Toulon.

Andreas Stolcke, Noah Coccaro, Rebecca Bates, Paul Taylor, Carol Van Ess-Dykema, Klaus Ries, Elizabeth Shriberg, Daniel Jurafsky, Rachel Martin, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3).

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *22nd International World Wide Web Conference, WWW '13*, pages 1445–1456. Rio de Janeiro.

Yi Yang, Doug Downey, and Jordan L. Boyd-Graber. 2015. Efficient methods for incorporating knowledge into topic models. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*, pages 308–317. Lisbon.

Elina Zarisheva and Tatjana Scheffler. 2015. Dialog act annotation for twitter conversations. In *Proceedings of the SIGDIAL 2015 Conference*, pages 114–123. Prague.

Jichuan Zeng, Jing Li, Yan Song, Cuiyun Gao, Michael R. Lyu, and Irwin King. 2018a. Topic memory networks for short text classification. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018*. Brussels.

Xingshan Zeng, Jing Li, Lu Wang, Nicholas Beauchamp, Sarah Shugars, and Kam-Fai Wong. 2018b. Microblog conversation recommendation via joint modeling of topics and discourse. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, Volume 1 (Long Papers)*, pages 375–385. New Orleans, LA.

Tiancheng Zhao, Kyusong Lee, and Maxine Eskénazi. 2018. Unsupervised discrete sentence representation learning for interpretable neural dialog generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Volume 1: Long Papers*, pages 1098–1107. Melbourne.

Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Volume 1: Long Papers*, pages 654–664. Vancouver.

Wayne Xin Zhao, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing Twitter and traditional media using topic models. In *Proceedings of Advances in Information Retrieval - 33rd European Conference on IR Research, ECIR 2011*, pages 338–349. Dublin.

Li Zhou, Jianfeng Gao, Di Li, and Heung-Yeung Shum. 2018. The design and implementation of Xiaoice, an empathetic social chatbot. *CoRR*, abs/1812.08989.