A Joint Model of Conversational Discourse and Latent Topics on Microblogs

Jing Li*
Tencent AI Lab
Shenzhen, China
ameliajli@tencent.com

Yan Song
Tencent AI Lab
Shenzhen, China
clksong@tencent.com

Zhongyu Wei
Fudan University
School of Data Science, Shanghai, China
zywei@fudan.edu.cn

Kam-Fai Wong
The Chinese University of Hong Kong
Department of Systems Engineering and
Engineering Management, HKSAR, China
kfwong@se.cuhk.edu.hk

Conventional topic models are ineffective for topic extraction from microblog messages, because the data sparseness exhibited in short messages lacking structure and contexts results in poor message-level word co-occurrence patterns. To address this issue, we organize microblog messages as conversation trees based on their reposting and replying relations, and propose an unsupervised model that jointly learns word distributions to represent: (1) different roles of conversational discourse, and (2) various latent topics in reflecting content information. By explicitly distinguishing the probabilities of messages with varying discourse roles in containing topical words, our model is able to discover clusters of discourse words that are indicative of topical content. In an automatic evaluation on large-scale microblog corpora, our joint model yields topics with better coherence scores than competitive topic models from previous studies.

* Jing Li is the corresponding author. This work was partially conducted when Jing Li was at Department of Systems Engineering and Engineering Management, The Chinese University of Hong Kong, HKSAR, China.

Submission received: 15 October 2017; revised version received: 4 May 2018; accepted for publication: 20 August 2018.

doi:10.1162/coli_a_00335

Qualitative analysis on model outputs indicates that our model induces meaningful representations for both discourse and topics. We further present an empirical study on microblog summarization based on the outputs of our joint model. The results show that the jointly modeled discourse and topic representations can effectively indicate summary-worthy content in microblog conversations.

1. Introduction

Over the past two decades, the Internet has been revolutionizing the way we communicate. Microblogging, a social networking channel over the Internet, further accelerates communication and information exchange. Popular microblog platforms, such as Twitter[1] and Sina Weibo,[2] have become important outlets for individuals to share information and voice opinions, which further benefit downstream applications such as instant detection of breaking events (Lin et al. 2010; Weng and Lee 2011; Peng et al. 2015), real-time and ad hoc search of microblog messages (Duan et al. 2010; Li et al. 2015b), public opinions and user behavior understanding on societal issues (Pak and Paroubek 2010; Popescu and Pennacchiotti 2010; Kouloumpis, Wilson, and Moore 2011), and so forth.

However, the explosive growth of microblog data far outpaces human beings' speed of reading and understanding. As a consequence, there is a pressing need for effective natural language processing (NLP) systems that can automatically identify gist information, and make sense of the unmanageable amount of user-generated social media content (Farzindar and Inkpen 2015). As one of the important and fundamental text analytic approaches, topic models extract key components embedded in microblog content by clustering words that describe similar semantic meanings to form latent "topics." The derived intermediate topic representations have proven beneficial to many NLP applications for social media, such as summarization (Harabagiu and Hickl 2011), classification (Phan, Nguyen, and Horiguchi 2008; Zeng et al. 2018a), and recommendation on microblogs (Zeng et al. 2018b).

Conventionally, probabilistic topic models (e.g., probabilistic latent semantic analysis [Hofmann 1999] and latent Dirichlet allocation [Blei et al. 2003]) have achieved huge success over the past decade, owing to their fully unsupervised manner and ease of extension. The semantic structure discovered by these topic models have facilitated the progress of many research fields, for example, information retrieval (Boyd-Graber, Hu, and Mimno 2017), data mining (Lin et al. 2015), and NLP (Newman et al. 2010). Nevertheless, ascribing to their reliance on document-level word co-occurrence patterns, the progress is still limited to formal conventional documents such as news reports (Blei, Ng, and Jordan 2003) and scientific articles (Rosen-Zvi et al. 2004). The aforementioned models work poorly when directly applied to short and colloquial texts (e.g., microblog posts) owing to severe sparsity exhibited in such text genre (Wang and McCallum 2006; Hong and Davison 2010).

Previous research has proposed several methods to deal with the sparsity issue in short texts. One common approach is to aggregate short messages into long pseudo-documents. Many studies heuristically aggregate messages based on authorship

1 twitter.com.
2 weibo.com.

(Hong and Davison 2010; Zhao et al. 2011), shared words (Weng et al. 2010), or hashtags (Ramage, Dumais, and Liebling 2010; Mehrotra et al. 2013). Quan et al. (2015) propose self-aggregation-based topic modeling (SATM) that aggregates texts jointly with topic inference. Another popular solution is to take into account word relations to alleviate document-level word sparseness. Biterm topic model (BTM) directly models the generation of word-pair co-occurrence patterns in each individual message (Yan et al. 2013; Cheng et al. 2014). More recently, word embeddings trained by large-scale external data are leveraged to capture word relations and improve topic models on short texts (Das, Zaheer, and Dyer 2015; Nguyen et al. 2015; Li et al. 2016a, 2017a; Shi et al. 2017; Xun et al. 2017).

To date, most efforts focus on content in messages, but ignore the rich discourse structure embedded in ubiquitous user interactions on microblog platforms. On microblogs, which were originally built for user communication and interaction, conversations are freely formed on issues of interests by reposting messages and replying to others. When joining a conversation, users generally post topically related content, which naturally provide effective contextual information for topic discovery. Alvarez-Melis and Saveski (2016) have shown that simply aggregating messages based on conversations can significantly boost the performance of conventional topic models and outperform models exploiting hashtag-based and user-based aggregations.

Another important issue ignored in most previous studies is the effective separation of topical words from non-topic ones (Li et al. 2016b). In microblog content, owing to its colloquial nature, non-topic words such as sentimental (e.g., "great" and "ToT"), functional (e.g., "doubt" and "why"), and other non-topic words (e.g., "oh" and "oops") are common and usually mixed with topical words. The occurrence of non-topic words may distract the models from recognizing topical content, which thus leads to the failure to produce coherent and meaningful topics. In this article, we propose a novel model that examines the entire context of a conversation and jointly explores word distributions representing varying types of topical content and *discourse roles* such as agreement, question-asking, argument, and other dialogue acts (Ritter, Cherry, and Dolan 2010).[3] Though Ritter, Cherry, and Dolan (2010) separate discourse, topic, and other words for modeling conversations, their model focuses on dialogue act modeling and only yields one distribution for topical content. Therefore, their model is unable to distinguish varying latent topics reflecting message content underlying the corpus. Li et al. (2016b) leverage conversational discourse structure to detect topical words from microblog posts, which explicitly explores the probabilities of different discourse roles that contain topical words. However, Li et al. (2016b) depend on a pre-trained discourse tagger and acquire a time-consuming and expensive manual annotation process for annotating conversational discourse roles on microblog messages, which does not scale for large data sets (Ritter, Cherry, and Dolan 2010; Joty, Carenini, and Lin 2011).

To exploit discourse structure of microblog conversations, we link microblog posts using reposting and replying relations to build conversation trees. Particularly, the root of a conversation tree refers to the original post and its edges represent the reposting or replying relations. To illustrate the interplay between topic and discourse, Figure 1 displays a snippet of Twitter conversation about "Trump administration's immigration ban." From the conversation, we can observe two major components: (1) *discourse*,

3 In this work, a discourse role refers to a certain type of dialogue act on message level, e.g., agreement or argument. The discourse structure of a conversation means some combination (or a probability distribution) of discourse roles.

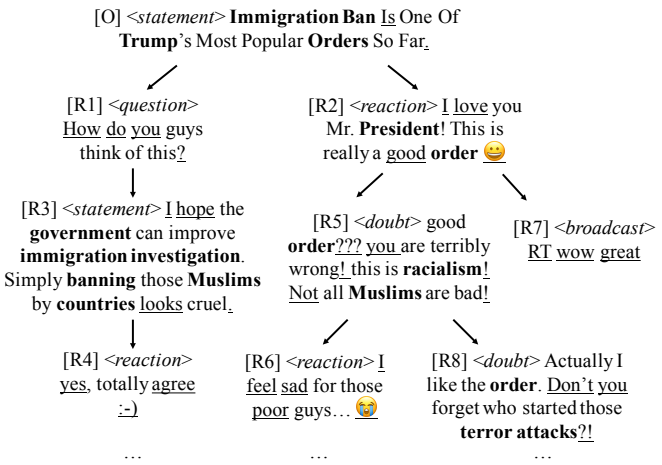

Figure 1
A sample Twitter conversation tree on "Trump administration's immigration ban." [O]: the original post; [Ri]: the i-th repost or reply; arrow lines: reposting or replying relations; *italic words* in ⟨⟩: discourse role of the message; underlined words: words indicating discourse roles; **bold words**: topical words representing the discussion focus.

indicated by the underlined words, describes the intention and pragmatic roles of messages in conversation structure, such as making a statement or asking a question; (2) *topic*, represented by the bold words, captures the topic and focus of the conversation, such as "racialism" and "Muslims." As we can see, different discourse roles vary in probabilities to contain key content reflecting the conversation focus. For example, in Figure 1, [R5] doubts the assertion of "immigration ban is good," and raises a new focus on "racialism." This in fact contains more topic-related words than [R6], which simply reacts to its parent. For this reason, in this article, we attempt to identify messages with "good" discourse roles that tend to describe key focuses and salient topics of a microblog conversation tree, which enables the discovery of "good" words reflecting coherent topics. Importantly, our joint model of conversational discourse and latent topics is fully unsupervised, therefore does not require any manual annotation.

For evaluation, we conduct quantitative and qualitative analysis on large-scale Twitter and Sina Weibo corpora. Experimental results show that topics induced by our model are more coherent than existing models. Qualitative analysis on discourse further shows that our model can yield meaningful clusters of words related to manually crafted discourse categories. In addition, we present an empirical study on downstream application of microblog conversation summarization. Empirical results on ROUGE (Lin 2004) show that summaries produced based on our joint model contain more salient information than state-of-the-art summarization systems. Human evaluation also indicates that our output summaries are competitive with existing unsupervised summarization systems in the aspects of informativeness, conciseness, and readability.

In summary, our contributions in this article are 3-fold:

- **Microblog posts organized as conversation trees for topic modeling.** We propose a novel concept of representing microblog posts as conversation trees by connecting microblog posts based on *reposting* and *replying*

relations for topic modeling. Conversation tree structure helps enrich context, alleviate data sparseness, and in turn improve topic modeling.

- **Exploiting discourse in conversations for topic modeling.** Our model differentiates the generative process of topical and non-topic words, according to the discourse role of the message where a word is drawn from being. This helps the model in identifying the topic-specific information from non-topic background.
- **Thorough empirical study on the inferred topic representations.** Our model shows better results than competitive topic models when evaluated on large-scale, real-world microblog corpora. We also present an effective method for using our induced results on microblog conversation summarization.

2. Related Work

This article builds upon diverse streams of previous work in lines of *topic modeling*, *discourse analysis*, and *microblog summarization*, which are briefly surveyed as follows.

2.1 Topic Models

Topic models aim to discover the latent semantic information, i.e., *topics*, from texts and have been extensively studied. This work is built upon the success of latent Dirichlet allocation (LDA) modeling (Blei et al. 2003; Blei, Ng, and Jordan 2003) and aims to learn topics in microblog messages. We first briefly introduce LDA in Section 2.1.1 and then review the related work on topic modeling for microblog content in Section 2.1.2.

2.1.1 LDA: Springboard of Topic Models. Latent Dirichlet allocation (Blei, Ng, and Jordan 2003) is one of the most popular and well-known topic models. It uses Dirichlet priors to generate document–topic and topic–word distributions, and has shown to be effective in extracting topics from conventional documents. LDA plays an important role in semantic representation learning and serves as the springboard of many famous topic models, e.g., hierarchical latent Dirichlet allocation (Blei et al. 2003), author–topic modeling (Rosen-Zvi et al. 2004), and so forth. In addition to "topic" modeling, it has also inspired *discourse* (Crook, Granell, and Pulman 2009; Ritter, Cherry, and Dolan 2010; Joty, Carenini, and Lin 2011) detection without supervision or with weak supervision. However, none of the aforementioned work jointly infers discourse and topics on microblog conversations, which is a gap the present article fills. Also, to the best of our knowledge, our work serves as the first attempt to exploit the joint effects of discourse and topic on unsupervised microblog conversation summarization.

2.1.2 Topic Models for Microblog Posts. Previous research has demonstrated that standard topic models, essentially focusing on document-level word co-occurrences, are not suitable for short and informal microblog messages because of the severe data sparsity exhibited in short texts (Wang and McCallum 2006; Hong and Davison 2010). As a result, one line of previous work focuses on enriching and exploiting contextual information. Weng et al. (2010), Hong and Davison (2010), and Zhao et al. (2011) first heuristically aggregate messages posted by the same user or sharing the same words before conventional topic models are applied. Their simple strategies, however, pose some problems. For example, it is common that a user has various interests and thus

posts messages covering a wide range of topics. Ramage, Dumais, and Liebling (2010) and Mehrotra et al. (2013) use hashtags as labels to train supervised topic models. Nevertheless, these models depend on large-scale hashtag-labeled data for model training. Moreover, their performance can be inevitably compromised when facing unseen topics that are irrelevant to any hashtag in training data. Such phenomena are common because of the rapid change and wide variety of topics on social media. BTM (Yan et al. 2013; Cheng et al. 2014) directly explores unordered word-pair co-occurrence patterns in each individual message, which is equivalent to extending short documents into a biterm set consisting of all combinations of any two distinct words appearing in the document. SATM (Quan et al. 2015) combines short text aggregation and topic induction into a unified model. However, in SATM, no prior knowledge is given to ensure the quality of text aggregation, which will further affect the performance of topic inference.

Different from the aforementioned work, we organize microblog messages as conversation trees based on reposting and replying relations. It allows us to explore word co-occurrence patterns in richer context, as messages in one conversation generally focus on relevant topics. Even though researchers have started to take the contexts provided by conversations into account when discovering topics on microblogs (Alvarez-Melis and Saveski 2016; Li et al. 2016b), there is much less work that jointly predicts the topical words along with the discourse structure in conversations. Ritter, Cherry, and Dolan (2010) model dialogue acts in conversations via separating discourse words from topical words and others. Whereas their model produces only one word distribution to represent the topical content, our model is capable of generating varying discourse and topic word distributions. Another main difference is that our model explicitly explores the probabilities of messages with different discourse roles in containing topical words for topic representation, whereas their model generates topical words from a conversation-specific distribution over word types regardless of the different discourse roles of messages. The work by Li et al. (2016b) serves as another prior effort to leverage conversation structure, captured by a supervised discourse tagger, on topic induction. Different from them, our model learns discourse structure for conversations in a fully unsupervised manner, which does not require annotated data.

Another line of research tackles data sparseness by modeling word relations rather than word occurrences in documents. For example, recent research work has shown that distributional similarities of words captured by word embeddings (Mikolov et al. 2013; Mikolov, Yih, and Zweig 2013) are useful in recognizing interpretable topic word clusters from short texts (Das, Zaheer, and Dyer 2015; Nguyen et al. 2015; Li et al. 2016a, 2017a; Shi et al. 2017; Xun et al. 2017). These topic models heavily rely on meaningful word embeddings needed to be trained on a large-scale, high-quality external corpus, which should be both in the same domain and the same language as the data for topic modeling (Bollegala, Maehara, and Kawarabayashi 2015). However, such external resource is not always available. For example, to the best of our knowledge, there currently exists no high-quality word embedding corpus for Chinese social media. In contrast to these prior methods, our model does not have the prerequisite to an external resource, whose general applicability in cold-start scenarios is therefore ensured.

2.1.3 Topic Modeling and Summarization. Previous studies have shown that the topic representation captured by topic models is useful for summarization (Nenkova and McKeown 2012). Specifically, there are two different purposes of using topic models in existing summarization systems: (1) to separate summary-worthy content and non-content background (general information) (Daumé and Marcu 2006; Haghighi and Vanderwende 2009; Çelikyilmaz and Hakkani-Tür 2010), and (2) to cluster sentences

or documents into topics, with summaries then generated from each topic cluster for minimizing content redundancy (Salton et al. 1997; McKeown et al. 1999; Siddharthan, Nenkova, and McKeown 2004). Similar techniques have also been applied to summarize events or opinions on microblogs (Chakrabarti and Punera 2011; Long et al. 2011; Rosa et al. 2011; Duan et al. 2012; Shen et al. 2013; Meng et al. 2012).

Our downstream application on microblog summarization lies in the research line of point (1), whereas we integrate the effects of discourse on key content identification, which has not been studied in any prior work. Also it is worth noting that, following point (2) to cluster messages before summarization is beyond the scope of this work because we are focusing on summarizing a single conversation tree, on which there are limited topics. We leave the potential of using our model to segment topics for multi-conversation summarization to future work.

2.2 Discourse Analysis

Discourse reflects the architecture of textual structure, where the semantic or pragmatic relations among text units (e.g., clauses, sentences, paragraphs) are defined. Here we review prior work on single document discourse analysis in Section 2.2.1, followed by a description on discourse extension to represent conversation structures in Section 2.2.2.

2.2.1 Traditional View of Discourse. It has been long pointed out that a coherent document, which gives readers continuity of senses (De Beaugrande and Dressler 1981), is not simply a collection of independent sentences. Linguists have striven to the study of discourse analysis ever since ancient Greece (Bakker and Wakker 2009). Early work shapes the modern concept of discourse (Hovy and Maier 1995) via depicting connections between text units, which reveals the structural art behind a coherent document.

Rhetorical structure theory (RST) (Mann and Thompson 1988) is one of the most influential discourse theories. According to its assumption, a coherent document can be represented by text units at different levels (e.g., clauses, sentences, paragraphs) in a hierarchical tree structure. In particular, the minimal units in RST (i.e., leaves of the tree structure) are defined as sub-sentential clauses, namely, elementary discourse units. Adjacent units are linked by rhetorical relations—condition, comparison, elaboration, and so forth. Based on RST, early work uses hand-coded rules for automatic discourse analysis (Marcu 2000; Thanh, Abeysinghe, and Huyck 2004). Later, thanks to the development of large-scale discourse corpora—RST corpus (Carlson, Marcu, and Okurovsky 2001), Graph Bank corpus (Wolf and Gibson 2005), and Penn Discourse Treebank (Prasad et al. 2008)—data-driven and learning-based discourse parsers that exploit various manually designed features (Soricut and Marcu 2003; Baldridge and Lascarides 2005; Fisher and Roark 2007; Lin, Kan, and Ng 2009; Subba and Eugenio 2009; Joty, Carenini, and Ng 2012; Feng and Hirst 2014) and representative learning (Ji and Eisenstein 2014; Li, Li, and Hovy 2014) became popular.

2.2.2 Discourse Analysis on Conversations. Stolcke et al. (2000) provide one of the first studies focusing on this problem, and it provides a general schema of understanding conversations with discourse analysis. Because of the complex structure and informal language style, discourse parsing on conversations is still a challenging problem (Perret et al. 2016). Most research focuses on the detection of dialogue acts (DAs),[4] which are

4 *Dialogue act* can be used interchangeably with *speech act* (Stolcke et al. 2000).

defined in Stolcke et al. (2000) as the first-level conversational discourse structure. It is worth noting that a DA represents the shallow discourse role that captures illocutionary meanings of an utterance ("statement," "question," "agreement," etc.).

Automatic dialogue act taggers have been conventionally trained in a supervised way with predefined tag inventories and annotated data (Stolcke et al. 2000; Cohen, Carvalho, and Mitchell 2004; Bangalore, Fabbrizio, and Stent 2006). However, DA definition is generally domain-specific and usually involves the manual designs from experts. Also, the data annotation process is slow and expensive, resulting in the limitation of data available for training DA classifiers (Jurafsky, Shriberg, and Biasca 1997; Dhillon et al. 2004; Ritter, Cherry and Dolan 2010; Joty, Carenini, and Lin 2011). These issues are pressing with the arrival of the Internet era in which new domains of conversations and even new types of dialogue act tags have boomed (Ritter, Cherry, and Dolan 2010; Joty, Carenini, and Lin 2011).

For this reason, researchers have proposed unsupervised or weakly supervised dialogue act taggers that identify indicative discourse word clusters based on probabilistic graphical models (Crook, Granell, and Pulman 2009; Ritter, Cherry, and Dolan 2010; Joty, Carenini, and Lin 2011). In our work, the discourse detection module is inspired by these previous models, where discourse roles are represented by word distributions and recognized in an unsupervised manner. Different from the previous work that focuses on discourse analysis, we explore the effects of discourse structure of conversations on distinguishing varying latent topics underlying the given collection, which has not been studied before. In addition, most existing unsupervised approaches for conversation modeling follows hidden Markov model convention and induces discourse representations in conversation threads. Considering that most social media conversations are in tree structure because one post is likely to spark multiple replying or reposting messages, our model allows the modeling of discourse roles in tree structure, which enables richer contexts to be captured. More details will be provided in Section 3.1.

2.3 Microblog Summarization

Microblog summarization can be considered as a special case of text summarization, which is conventionally defined as discovering essential content from given document(s), and producing concise and informative summaries covering important information (Radev, Hovy, and McKeown 2002). Summarization techniques can be generally categorized as extractive or abstractive methods (Das and Martins 2007). **Extractive summarization** captures and distills salient content, which are usually sentences, to form summaries. **Abstractive summarization** focuses on identifying key text units (e.g., words and phrases) and then generates grammatical summaries based on these units. Our summarization application falls into the category of extractive summarization.

Early work on microblog summarization attempts to apply conventional extractive summarization models directly—LexRank (Erkan and Radev 2004), the University of Michigan's summarization system MEAD (Radev et al. 2004), TF-IDF (Inouye and Kalita 2011), integer linear programming (Liu, Liu, and Weng 2011; Takamura, Yokono, and Okumura 2011), graph learning (Sharifi, Hutton, and Kalita 2010), and so on. Later, researchers found that standard summarization models are not suitable for microblog posts because of the severe redundancy, noise, and sparsity problems exhibited in short and colloquial messages (Chang et al. 2013; Li et al. 2015a). To solve these problems, one common solution is to use social signals such as the user influence and retweet counts to help summarization (Duan et al. 2012; Liu et al. 2012; Chang et al. 2013). Different from the aforementioned studies, we do not include external features such

as the social network structure, which ensures the general applicability of our approach when applied to domains without such information.

Discourse has been reported useful to microblog summarization. Zhang et al. (2013) and Li et al. (2015a) leverage dialogue acts to indicate summary-worthy messages. In the field of conversation summarization from other domains (e.g., meetings, forums, and e-mails), it is also popular to leverage the pre-detected discourse structure for summarization (Murray et al. 2006; McKeown, Shrestha, and Rambow 2007; Wang and Cardie 2013; Bhatia, Biyani, and Mitra 2014; Bokaei, Sameti, and Liu 2016). Oya and Carenini (2014) and Qin, Wang, and Kim (2017) address discourse tagging together with salient content discovery on e-mails and meetings, and show the usefulness of their relations in summarization. For all the systems mentioned here, manually crafted tags and annotated data are required for discourse modeling. Instead, the discourse structure is discovered in a fully unsupervised manner in our model, which is represented by word distributions and can be different from any human designed discourse inventory. The effects of such discourse representations on salient content identification have not been explored in previous work.

3. The Joint Model of Conversational Discourse and Latent Topics

We assume that the given corpus of microblog posts is organized as C conversation trees, based on reposting and replying relations. Each tree c contains M_c microblog messages and each message m has $N_{c,m}$ words in vocabulary. The vocabulary size is V. We separate three components, *discourse*, *topic*, and *background*, underlying the given conversations, and use three types of word distributions to represent them.

At the corpus level, there are K topics represented by word distribution $\phi_k^T \sim Dir(\beta)$ $(k = 1, 2, ..., K)$. $\phi_d^D \sim Dir(\beta)$ $(d = 1, 2, ..., D)$ represents the D discourse roles embedded in the corpus. In addition, we add a background word distribution $\phi^B \sim Dir(\beta)$ to capture general information (e.g., common words) that cannot indicate either discourse or topic. ϕ_k^T, ϕ_d^D, and ϕ^B are all V-dimensional multinomial word distributions over the vocabulary. For each conversation tree c, $\theta_c \sim Dir(\alpha)$ models the mixture of topics and any message m on c is assumed to contain a single topic $z_{c,m} \in \{1, 2, ..., K\}$.

3.1 Message-Level Modeling

For each message m on conversation tree c, our model assigns two message-level multinomial variables to it: $d_{c,m}$ representing its discourse role and $z_{c,m}$ reflecting its topic assignment, whose definitions are given in turn in the following.

Discourse Roles. Our discourse detection is inspired by Ritter, Cherry, and Dolan (2010), which exploits the discourse dependencies derived from reposting and replying relations for assigning discourse roles. For example, a "doubt" message is likely to start controversy thus triggering another "doubt," (e.g., [R5] and [R8] in Figure 1). Assuming that the index of m's parent is $pa(m)$, we use transition probabilities $\pi_d \sim Dir(\gamma)$ $(d = 1, 2, ..., D)$ to explicitly model discourse dependency of m to $pa(m)$. π_d is a distribution over the D discourse roles and $\pi_{d,d'}$ denotes the probability of m assigned discourse d' given the discourse of $pa(m)$ being d. Specifically, $d_{c,m}$ (discourse role of message m) is generated from discourse transition distribution $\pi_{d_{t,pa(m)}}$ where $d_{t,pa(m)}$ is the discourse assignment on $pa(m)$. In particular, to create a unified generation story, we place a pseudo message emitting no word before the root of each conversation tree

and assign dummy discourse indexing $D+1$ to it. π_{D+1}, the discourse transition from pseudo messages to roots, in fact models the probabilities of different discourse roles as conversation starter.

Topic Assignments. Messages on one conversation tree focus on related topics. To exploit such intuition in topic assignments, the topic of each message m on conversation tree c (i.e., $z_{c,m}$) is sampled from the topic mixture θ_c of conversation tree c.

3.2 Word-Level Modeling

To distinguish varying types of word distributions to separately capture *discourse*, *topic*, and *background* representations, we follow the solutions from previous work to assign each word as a discrete and exact source that reflects one particular type of word representation (Daumé and Marcu 2006; Haghighi and Vanderwende 2009; Ritter, Cherry, and Dolan 2010). To this end, for each word n in message m and tree c, a ternary variable $x_{c,m,n} \in \{\text{DISC}, \text{TOPIC}, \text{BACK}\}$ controls word n to fall into one of the three types: *discourse*, *topic*, and *background* word. In doing so, words in the given collection are explicitly separated into three types, based on which the word distributions representing discourse, topic, and background components are separated accordingly.

Discourse words. (DISC) indicate the discourse role of a message; for example, in Figure 1, "How" and the question mark "?" reflect that [R1] should be assigned the discourse role of "question." If $x_{c,m,n} = \text{DISC}$ (i.e., n is assigned as a discourse word), then word $w_{c,m,n}$ is generated from discourse word distribution $\phi^D_{d_{c,m}}$, where $d_{c,m}$ is m's discourse role.

Topic words. (TOPIC) are the core topical words that describe topics being discussed in a conversation tree, such as "Muslim," "order," and "Trump" in Figure 1. When $x_{c,m,n} = \text{TOPIC}$, i.e., n is assigned as a topic word, word $w_{c,m,n}$ is hence generated from the word distribution of the topic assigned to message m, i.e., $\phi^T_{z_{c,m}}$

Background words. (BACK) capture the general words irrelevant to either discourse or topic, such as "those" and "of" in Figure 1. When word n is assigned as a background word ($x_{c,m,n} = \text{BACK}$), word $w_{c,m,n}$ is then drawn from background distribution ϕ^B.

Switching Among Topic, Discourse, and Background. We assume that messages of different discourse roles may show different distributions of the word types as discourse, topic, and background. The ternary word type switcher $x_{c,m,n}$ is hence controlled by the discourse role of message m. In specific, $x_{c,m,n}$ is drawn from the three-dimensional distribution $\tau_{d_{c,m,n}} \sim Dir(\delta)$ that captures the appearing probabilities of three types of words (DISC, TOPIC, BACK), when the discourse assignment to m is $d_{c,m,n}$, that is, $x_{c,m,n} \sim Multi(\tau_{d_{c,m}})$. For instance, a statement message (e.g., [R3] in Figure 1) may contain more content words for topic representation than a question to other users (e.g., [R1] in Figure 1). In particular, stop words and punctuation are forced to be labeled as discourse or background words. By explicitly distinguishing different types of words with switcher $x_{c,m,n}$, we can thus separate the three types of word distributions that reflect discourse, topic, and background information.

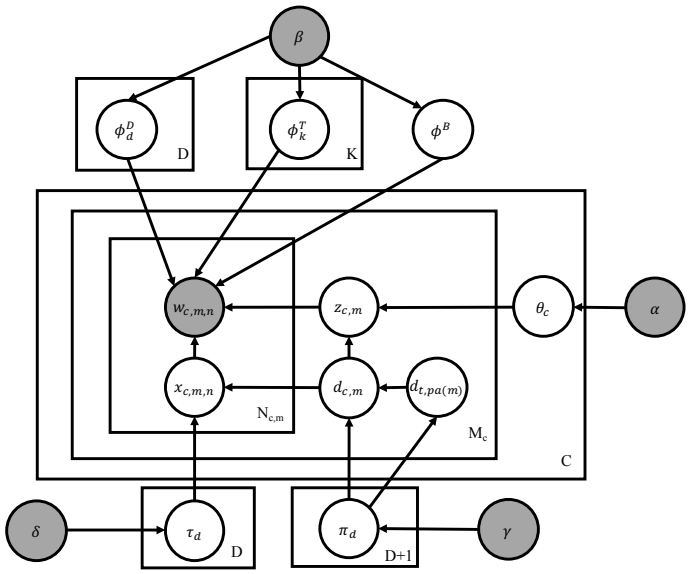

Figure 2
The graphical model illustrating the generative process of our joint model of conversational discourse and latent topics.

3.3 Generative Process and Parameter Estimation

In summary, Figure 2 illustrates the graphical model of our generative process that jointly explores conversational discourse and latent topics. The following shows the detailed generative process of the conversation tree c:

- Draw topic mixture of conversation tree $\theta_c \sim Dir(\alpha)$
- For message $m = 1$ to M_c
  - Draw discourse role $d_{c,m} \sim Multi(\pi_{d_{c,pa(m)}})$
  - Draw topic assignment $z_{c,m} \sim Multi(\theta_c)$
  - For word $n = 1$ to $N_{c,m}$
    Draw ternary word type switcher $x_{c,m,n} \sim Multi(\tau_{d_{c,m}})$
    If $x_{c,m,n} ==$ DISC
      Draw $w_{t,s,n} \sim Multi(\phi^D_{d_{c,m}})$
    If $x_{c,m,n} ==$ TOPIC
      Draw $w_{c,m,n} \sim Multi(\phi^T_{z_{c,m}})$
    If $x_{c,m,n} ==$ BACK
      Draw $w_{c,m,n} \sim Multi(\phi^B)$

For parameter estimation, we use collapsed Gibbs sampling (Griffiths and Steyvers 2004) to carry out posterior inference for parameter learning. The hidden multinomial variables (i.e., message-level variables [d and z] and word-level variable [x]) are sampled in turn, conditioned on a complete assignment of all other hidden variables and hyper-parameters $\Theta = (\alpha, \beta, \gamma, \delta)$. For more details, we refer the readers to Appendix A.

Table 1
Statistics of our five data sets on Twitter and Sina Weibo for evaluating topic coherence.

Data Set	# of trees	# of messages	Vocab size
Twitter			
SemEval	8,652	13,582	3,882
PHEME	7,961	92,883	10,288
US Election	4,396	33,960	5,113
Weibo			
Weibo-1	9,959	91,268	11,849
Weibo-2	21,923	277,931	19,843

4. Experiments on Topic Coherence

This section presents an experiment on the coherence of topics yielded by our joint model of conversational discourse and latent topics.

4.1 Data Collection and Experiment Set-up

Data Sets. To examine the coherence of topics on diverse microblog data sets, we conduct experiments on data sets collected from two popular microblog Web sites: Twitter and Weibo,[5] where the messages are mostly in English and Chinese, respectively. Table 1 shows the statistics of our five data sets used to evaluate topic coherence. In the following, we give the details of their collection processes in turn.

For Twitter data, we evaluate the coherence of topics on three data sets: *SemEval*, *PHEME*, and *US Election*, and tune all models in our experiments on a large-scale development data set from TREC2011 microblog track.[6]

- *SemEval*. We combine the data released for topic-oriented sentiment analysis task in SemEval 2015[7] and 2016.[8] To recover the missing ancestors in conversation trees, we use Tweet Search API to retrieve messages with the "in-reply-to" relations, and collect tweets in a recursive way until all the ancestors in a conversation are recovered.[9]
- *PHEME*. This data set was released by Zubiaga, Liakata, and Procter (2016), and contains conversations around rumors and non-rumors posted during five breaking events: *Charlie Hebdo*, *Ferguson*, *Germanwings Crash*, *Ottawa Shooting*, and *Sydney Siege*.

5 Weibo, short for Sina Weibo, is the biggest microblog platform in China and shares the similar market penetration as Twitter (Rapoza 2011). Similar to Twitter, it has length limitation of 140 Chinese characters.
6 http://trec.nist.gov/data/tweets/.
7 http://alt.qcri.org/semeval2015/task10/.
8 http://alt.qcri.org/semeval2016/task4/.
9 Twitter search API: https://developer.twitter.com/en/docs/tweets/search/api-reference/get-saved_searches-show-id. Twitter has allowed users to add comments in retweets (reposting messages on Twitter) since 2015, which enables retweets to become part of a conversation. In our data set, the parents of 91.4% of such retweets can be recovered from the "in reply to status id" field returned by Twitter search API.

- *US Election*. Considering that the *SemEval* and *PHEME* data sets cover a relatively wide range of topics, we are interested in studying a more challenging problem: whether topic models can differentiate latent topics in a narrow scope. To this end, we take political tweets as an example and conduct experiments on a data set with Twitter discussions about the U.S. presidential election 2016. The data set is extended from the one released by Zeng et al. (2018b) following three steps. First, some raw tweets that are likely to be in a conversation are collected by searching conversation-type keywords via Twitter Streaming API,[10] which samples and returns tweets matching the given keywords.[11] Second, conversations are recovered via "in-reply-to" relations as is done to build the *SemEval* data set. Third, the relevant conversations are selected where there exist at least one tweet containing election-related keywords.[12]

For Weibo data, we track the real-time trending hashtags[13] on Sina Weibo and use the hashtag-search API[14] to crawl the posts matching the given hashtag queries. In the end, we build a large-scale corpus containing messages posted from 2 January 2014 to 31 July 2014. To examine the performance of models on varying topic distributions, we split the corpus into seven subsets, each containing messages posted in one month. We report the topic coherence on two randomly selected subsets, *Weibo-1* and *Weibo-2*. The remaining five data sets are used as development sets.

Comparisons. Our model jointly identifies word clusters of discourse and topics, and explicitly explores their relations, namely, the probabilities of different discourse roles in containing topical words (see Section 3.2), which is called the TOPIC+DISC+REL model in the rest of the article. In comparison, we consider the following established models: (1) LDA: In this model, we consider each message as a document and directly apply LDA (Blei et al. 2003; Blei, Ng, and Jordan 2003) on the collection. The implementation of LDA is based on the public toolkit GibbsLDA++.[15] (2) BTM: BTM[16] (Yan et al. 2013; Cheng et al. 2014) is a state-of-the-art topic model for short texts. It directly models the topics of all word pairs (biterms) in each message, which has proven more effective on social media texts than LDA (Blei et al. 2003; Blei, Ng, and Jordan 2003), one-topic-per-post Dirichlet multinomial mixture (DMM) (Nigam et al. 2000), and Zhao et al. (2011) (a DMM version on posts aggregated by authorship). According to the empirical study by Li et al. (2016b), BTM has in general better performance than a newer SATM model (Quan et al. 2015) on microblog data.

In particular, this article attempts to induce topics with little external resource. Therefore, we do not compare with either Li et al. (2016b), which depends on human annotation to train a discourse tagger, or topic models that exploit word embeddings

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

11 Conversation-type keywords are used to obtain tweets reflecting *agreement*, *disagreement*, and *response*, which are likely to appear in Twitter conversations. Keyword list: *agreement* – "agreed," "great point," "agree," "good point"; *disagreement* – "wrong," "bad idea," "stupid idea," "disagree"; *response* – "understand," "interesting," "i see."

12 The full list of election-related keywords: "trump," "clinton," "hillary," "election," "president," "politics."

13 http://open.weibo.com/wiki/Trends/hourly?sudaref=www.google.com.hk&retcode=6102.

14 http://open.weibo.com/wiki/2/search/topics.

15 http://gibbslda.sourceforge.net/.

16 https://github.com/xiaohuiyan/BTM.

(Das, Zaheer, and Dyer 2015; Nguyen et al. 2015; Li et al. 2016a, 2017a; Shi et al. 2017; Xun et al. 2017) pre-trained on large-scale external data. The external data in training embeddings should be in both the same domain and the same language of the given collection used for topic models, which limits the applicability of topic models in the scenarios without such data. Also, Li et al. (2016b) have shown that topic models combining word embeddings trained on internal data result in worse coherence scores than BTM, which is considered in our comparison.

In addition to the existing models from previous work, we consider the following variants that explore topics by organizing messages as conversation trees:

- The TOPIC ONLY model aggregates messages from one conversation tree as a pseudo-document, on which Chemudugunta, Smyth, and Steyvers (2006) (a model proven better than LDA in topic coherence) is used to induce topics on conversation aggregations, without modeling discourse structure. It involves a background word distribution to capture non-topic words, like our TOPIC+DISC+REL model. However, different from our TOPIC+DISC+REL model, the background word distribution is controlled by a general Beta prior without differentiating discourse roles of messages.
- The TOPIC+DISC model is an extension to Ritter, Cherry, and Dolan (2010), following which the switcher indicating a word as a discourse, topic, or background word are drawn from a conversation-level distribution over word types. Instead, in TOPIC+DISC+REL, word-type switcher depends on message-level discourse roles (shown in Section 3.2). In terms of topic generation of TOPIC+DISC, as the model of Ritter, Cherry, and Dolan (2010) is incapable of differentiating various latent topics, we follow the same procedure of TOPIC ONLY and TOPIC+DISC+REL models to draw topics from conversation-level topic mixture. Another difference between TOPIC+DISC and Ritter, Cherry, and Dolan (2010) is that the discourse roles of TOPIC+DISC are explored in tree-structured conversations whereas those in Ritter, Cherry, and Dolan (2010) are captured in context of conversation treads (paths of the conversation tree).

Hyper-parameters. For the hyper-parameters of our joint TOPIC+DISC+REL model, we fix $\alpha = 50/K$, $\beta = 0.01$, following the common practice in previous work (Yan et al. 2013; Cheng et al. 2014). For Twitter corpora, we set the count of discourse roles as $D = 10$, according to previous setting in Ritter, Cherry, and Dolan (2010). Because there is no analog of γ (controlling the prior for discourse role dependencies of children messages to their parents), δ (controlling the prior of distributions over topic, discourse, and background words given varying discourse roles), and discourse count D in Chinese Weibo corpora, we tune them by grid search on development sets and obtain $\gamma = 0.5$, $\delta = 0.25$, and $D = 6$ on Weibo data.

The hyper-parameters of LDA and BTM are set according to the best hyper-parameters reported in their original papers. For TOPIC ONLY and TOPIC+DISC models, the parameter settings are kept the same as TOPIC+DISC +REL, because they are its variants. And the background switchers are parameterized by symmetric Beta prior on 0.5, following the original setting from Chemudugunta, Smyth, and Steyvers (2006). We run Gibbs samplings of all models with 1,000 iterations to ensure convergence, following Zhao et al. (2011), Yan et al. (2013), and Cheng et al. (2014).

Preprocessing. Before training topic models, we preprocess the data sets as follows. For Twitter corpora, we (1) filter non-English messages; (2) replace links, mentions (i.e., @username), and hashtags with generic tags of "URL," "MENTION," and "HASHTAG"; (3) tokenize messages and annotate part-of-speech (POS) tags to each word using Tweet NLP toolkit (Gimpel et al. 2011; Owoputi et al. 2013)[17]; and (4) normalize all letters to lowercases. For Weibo corpora, we (1) filter non-Chinese messages; and (2) use FudanNLP toolkit (Qiu, Zhang, and Huang 2013) for word segmentation. Then, for each data set from Twitter or Sina Weibo, we generate a vocabulary and remove low-frequency words (i.e., words occurring fewer than five times).

For our TOPIC+DISC+REL model and its variants TOPIC ONLY and TOPIC+DISC considering the conversation structure, we only remove digits but retain stop words and punctuation in the data because: (1) stop words and punctuation can be useful discourse indicators, such as the question mark "?" and "what" in indicating "question" messages; (2) these models are equipped with a background distribution ϕ^B to separate general information useless to indicate either discourse or topic, e.g., "do" and "it"; and (3) we forbid stop words and punctuation to be sampled as topical words by forcing their word type switcher $x \neq$TOPIC in word generation (shown in Section 3.2). For LDA and BTM that cannot separate non-topic information, we filter out stop words and short messages with fewer than two words in preprocessing, which retains the same common settings to ensure comparable performance.[18]

Evaluation Metrics. Topic model evaluation is inherently difficult. Although in many previous studies perplexity is a popular metric to evaluate the predictive abilities of topic models given held-out data set with unseen words (Blei, Ng, and Jordan 2003), we do not consider perplexity here because high perplexity does not necessarily indicate semantically coherent topics in human perception (Chang et al. 2009).

The quality of topics is commonly measured by UCI (Newman et al. 2010) and UMass coherence scores (Mimno et al. 2011), assuming that words representing a coherent topic are likely to co-occur within the same document. We only consider UMass coherence here as UMass and UCI generally agree with each other, according to Stevens et al. (2012). We also consider a newer evaluation metric, the CV coherence measure (Röder, Both, and Hinneburg 2015), as it has been proven to provide the scores closest to human evaluation compared with other widely used topic coherence metrics, including UCI and UMass scores.[19] For the CV coherence measure, in brief, given a word list for topic representations (i.e., the top N words by topic–word distribution), some known topic coherence measures are combined, which estimates how similar their co-occurrence patterns with other words are in the context of a sliding window from Wikipedia.

4.2 Main Comparison Results

We evaluate topic models with two sets of K (i.e., the number of topics, $K = 50$ and $K = 100$) following previous settings (Li et al. 2016b). Tables 2 and 3 show the UMass and CV scores for topics produced on the evaluation corpora, respectively. For UMass

17 http://www.cs.cmu.edu/~ark/TweetNLP/.

18 We also conducted evaluations on the LDA and BTM versions without this preprocessing step, and we obtained worse coherence scores.

19 http://aksw.org/Projects/Palmetto.html.

Table 2
UMass coherence scores for topics produced by various models. Higher is better. K50 = 50 topics; K100 = 100 topics; N = the number of top words ranked by topic-word probabilities; w/o conversation = the models consider each message as a document and infers topics without using conversational information; w/ conversation = messages first are organized as conversation trees and then the model induces topics in the context of conversation trees. Higher scores indicate better coherence. Best result for each setting is in **bold**.

N	Model	*Weibo-1*		*Weibo-2*		*SemEval*		*PHEME*		*US Election*	
		K50	K100	K50	K100	K50	K100	K50	K100	K50	K100
	W/o conversation										
	LDA	−11.77	−11.57	−10.56	−12.08	−12.20	−12.02	−13.27	−13.98	**−10.07**	−10.89
	BTM	−9.56	−8.74	−8.65	−9.88	−9.93	−9.61	−10.22	−10.44	−12.15	−12.15
5	**W/ conversation**										
	TOPIC ONLY	**−8.00**	−8.78	−9.45	−10.06	−8.93	−8.88	−10.82	−10.63	−10.75	−10.98
	TOPIC+DISC	−9.47	−8.87	−9.85	**−9.60**	**−8.42**	−8.26	−10.54	**−10.40**	−10.36	−11.17
	TOPIC+DISC+REL	−8.53	**−8.66**	**−8.00**	−9.84	−8.47	**−8.19**	**−10.21**	−10.41	−11.14	**−10.75**
	W/o conversation										
	LDA	−120.06	−123.74	−117.00	−123.98	−128.15	−132.96	−138.99	−145.44	−105.21	−110.82
	BTM	−89.98	−86.96	−87.97	**−93.03**	−105.76	−105.98	−108.70	−111.32	−114.85	−123.42
10	**W/ conversation**										
	TOPIC ONLY	−90.53	−89.89	−108.51	−101.20	−89.02	−90.53	−105.62	−108.25	**−104.29**	−108.51
	TOPIC+DISC	−91.96	−88.75	−100.77	−100.58	−87.89	−91.82	−106.58	**−107.14**	−105.21	−108.31
	TOPIC+DISC+REL	**−86.91**	**−87.05**	**−83.59**	−98.19	**−86.48**	**−90.02**	**−105.27**	−107.91	−104.99	**−107.03**
	W/o conversation										
	LDA	−367.68	−357.88	−366.31	−373.98	−383.05	−391.67	−418.58	−424.66	−429.89	−436.87
	BTM	−265.80	−262.62	−281.06	**−281.46**	−307.23	−323.37	−328.36	−339.94	−344.99	−360.95
15	**W/ conversation**										
	TOPIC ONLY	−261.98	−260.62	−298.77	−294.51	−257.92	−266.86	−313.96	**−315.78**	−313.07	−319.99
	TOPIC+DISC	−261.30	−259.23	−301.99	−293.21	−261.25	−265.88	−313.22	−320.05	−317.14	−317.82
	TOPIC+DISC+REL	**−254.94**	**−256.47**	**−249.32**	−287.82	**−256.83**	**−265.71**	**−312.49**	−319.01	**−312.90**	**−315.59**
	W/o conversation										
	LDA	−771.34	−736.55	−718.48	−741.77	−777.00	−782.51	−856.37	−859.59	−898.77	−892.84
	BTM	−559.69	−553.62	−526.01	−586.65	−636.15	−669.16	−682.39	−709.81	−713.05	−739.35
20	**W/ conversation**										
	TOPIC ONLY	−528.13	−527.71	−602.16	−597.80	**−529.39**	−541.31	−643.91	**−647.74**	−634.97	−638.10
	TOPIC+DISC	−530.23	−524.15	−607.84	−585.99	−535.22	−541.18	−641.82	−656.16	−641.77	−639.35
	TOPIC+DISC+REL	**−518.97**	**−519.11**	**−509.79**	**−578.80**	−530.56	**−538.31**	**−637.18**	−650.70	**−629.42**	**−634.22**

Table 3
CV coherence scores for topics produced by various models on Twitter. Higher is better. K50 = 50 topics; K100 = 100 topics; N = the number of top words ranked by topic-word probabilities. Higher scores indicate better coherence. Best result for each setting is in **bold**.

N	Model	*SemEval*		*PHEME*		*US Election*	
		K50	K100	K50	K100	K50	K100
	W/o conversation						
	LDA	0.514	0.498	0.474	0.470	0.473	0.470
	BTM	0.528	0.518	0.486	0.477	0.481	0.480
5	**W/ conversation**						
	TOPIC ONLY	0.526	0.521	**0.492**	0.485	0.477	0.475
	TOPIC+DISC	0.526	0.523	0.481	0.483	0.475	0.478
	TOPIC+DISC+REL	**0.535**	**0.524**	0.491	**0.493**	**0.482**	**0.483**
	W/o conversation						
	LDA	0.404	0.401	0.375	0.378	0.351	0.359
	BTM	0.412	0.406	0.386	0.385	0.354	0.363
10	**W/ conversation**						
	TOPIC ONLY	0.399	**0.410**	0.388	0.385	0.359	0.360
	TOPIC+DISC	0.408	**0.410**	0.388	**0.386**	0.356	0.364
	TOPIC+DISC+REL	**0.414**	**0.410**	**0.398**	**0.386**	**0.366**	**0.366**

coherence, the top 5, 10, 15, and 20 words of each topic are selected for evaluation. For CV coherence, the top 5 and 10 words are selected.[20] Note that we cannot report CV scores on Chinese Weibo corpora because CV coherence is calculated based on a Wikipedia data set, which does not yet have a Chinese version. From the results, we identify the following observations:

- ***Conventional topic models cannot perform well on microblog messages.*** From all the comparison results, the topic coherence given by LDA is the worst, which may be because of the sparseness of document-level word concurrence patterns in short posts.
- ***Considering conversation structure is useful to topic inference.*** Using the contextual information provided by conversations, TOPIC ONLY produced competitive results compared to the state-of-the-art BTM model on short text topic modeling. This observation indicates the effectiveness of using the conversation structure to enrich context and thus results in latent topics of reasonably good quality.
- ***Jointly learning discourse information helps produce coherent topics.*** TOPIC+DISC and TOPIC+DISC+REL models yield generally better coherence scores than TOPIC ONLY, which explores topics without considering discourse. The reason may be that additionally exploring discourse in non-topic information helps recognize non-topic words, which further facilitates the separation of topical words from non-topic ones.
- ***Considering discourse roles of messages in topical word generation is useful.*** The results of TOPIC+DISC+REL are the best in most settings. One

20 Palmetto toolkit only allows at most 10 words as input for CV score calculation.

important reason is that TOPIC+DISC+REL explicitly explores the different probabilities of messages with varying discourse roles in containing topical or non-topic words, whereas the other models separate topical content from non-topic information regardless of the different discourse roles of messages. This observation demonstrates that messages with different discourse roles do vary in tendencies to cover topical words, which provides useful clues for key content words to be identified for topic representation.

4.3 Case Study

To further evaluate the interpretability of the latent topics and discourse roles learned by our TOPIC+DISC+REL model, we present a qualitative analysis on the output samples.

Sample Latent Topics. We first present a qualitative study on the sample-produced topics. Table 4 displays the top 15 words of topic "Trump is a racist" induced by different models on the US election data set given $K = 100$.[21] We have the following observations:

- It is challenging to extract coherent and meaningful topics from short and informal microblog messages. Without using an effective strategy to alleviate the data sparsity problem, LDA mixes the generated topic with *non-topic words*,[22] such as "direct," "describe," and "opinion," which are also likely to appear in messages whose topics are very different from "Trump is a racist."
- By aggregating messages based on conversations, TOPIC ONLY yields the topic competitive to the one produced by a state-of-the-art BTM model. The reason behind this observation could be that the conversation context provides rich word co-occurrence patterns in topic induction, which is beneficial to alleviate the data sparsity.
- The topics produced by TOPIC+DISC and TOPIC+DISC+REL contain fewer non-topic words than TOPIC ONLY, which does not consider discourse information when generating topics, and thus contains many general words, such as "thing" and "work," which cannot clearly indicate "Trump is a racist."
- The topic generated by TOPIC+DISC+REL best describes the topic "Trump is a racist" except for a non-topic word "call" at the end of the list. This is because it successfully discovers messages with discourse roles that are more likely to cover words describing the key focus in the conversations centering on "Trump is a racist." Without capturing such information, the topic produced by TOPIC+DISC contains some non-topic words like "yeah" and "agree."

21 If there are multiple latent topics related to "Trump is a racist," we pick up the most relevant one and display its representative words.

22 *Non-topic words* cannot clearly indicate the corresponding topic. Such words can occur in messages covering very different topics. For example, in Table 4, the word "opinion" is a non-topic word for "Trump is a racist," because an "opinion" can be voiced on diverse people, events, entities, and so on.

Table 4
The extracted topics describing "Trump is a racist." Top 15 words are selected by likelihood. Words are listed in decreasing order of probability given the topic. The words in red and **boldface** indicate non-topic words.

W/o conversation
LDA:
call racist democracy **opinion** race racism ignorant **definition** bigot **direct** mexican entitle **describe** card bigotry
BTM:
racist people hate trump white racism muslims **agree** race **call** group **make** fact problem immigrant

W/ conversation
TOPIC ONLY:
people black white **understand** racist **vote** agree wrong **work** president **thing** trump privilege disagree **system**
TOPIC+DISC:
white racist **yeah** trump privilege black race **agree** people bias **true** state issue **understand** muslims
TOPIC+DISC+REL:
white people black racist hate race wrong privilege america muslims trump kill racism illegal **call**

Table 5
Top 30 representative terms for sample discourse roles discovered by our TOPIC+DISC+REL model in PHEME data set given $K = 100$. Names of the discourse roles are our interpretations, according to the generated word distributions.

Statement	MENTION . the they are HASHTAG we , to and of in them all their ! be will these our who & should do this for if us need have
Reaction	MENTION ! . you URL HASHTAG for your thank this , on my i thanks so ... and a !! the are me please oh all very !!! - is
Question	MENTION ? you the what are is do HASHTAG why they that how this a to did about who in he so or u was it know can does on
Doubt	MENTION . you i , your a are to don't it but that if know u not me i'm and do have my think ? you're just about was it's
Reference	: URL MENTION HASHTAG in " " . at , the of on a - has from is to after and are " been have as more for least 2

Sample Discourse Roles. To show the discourse representation exploited by our TOPIC+DISC+REL model, we then present the sample discourse roles learned from the PHEME data set in Table 5. Although this is merely a qualitative human judgment, there appear to be interesting word clusters that reflect varying discourse roles found by our model without the guidance from any manual annotation on discourse. In the first column of Table 5, we intuitively name the sample generated discourse roles, which are based on our interpretations of the word cluster, and are provided to benefit the reader. We discuss each displayed discourse role in turn:

- *Statement* presents arguments and judgments, where words like "should," and "need" are widely used in suggestions and "if" occurs when conditions are given.

- *Reaction* expresses non-argumentative opinions. Compared with "statement" messages, "reaction" messages are straightforward and generally do not contain detailed explanations (e.g., conditions). Examples include simple feeling expressions, indicated by "oh" and "!!!," and acknowledgements, indicated by "thank" or "thanks."
- *Question* represents users asking questions to other users, implied by the question mark "?," "what," "why," and so forth.
- *Doubt* expresses strong opinions against something. Examples of indicative words include "but," "don't," "just," the question mark "?," and so on.
- *Reference* is for quoting external resource, which is implied by words like "from," colon, and quotation marks. The use of hashtags[23] and URLs are also prominent.

5. Downstream Application: Conversation Summarization on Microblogs

Section 4 has shown that conversational discourse is helpful to recognizing key topical information from short and informal microblog messages. We are interested in whether the induced topic and discourse representations can also benefit downstream applications. Here we take microblog summarization as an example that suffers from the data sparsity problem (Chang et al. 2013; Li et al. 2015a), similar to topic modeling on short texts. In this article, we focus on a subtask of microblog summarization, namely, **microblog conversation summarization**, and present an empirical study to show how our output can be used to predict critical content in conversations.

We first present the task description. Given a conversation tree, succinct summaries should be produced by extracting salient content from the massive reposting and replying messages in the conversation. It helps users understand the key focus of a microblog conversation. It is also called microblog context summarization in some previous work (Chang et al. 2013; Li et al. 2015a), because the produced summaries capture informative content in the lengthy conversations and provide valuable contexts to a short post, such as the background information and public opinions. In this task, the *input* is a microblog conversation tree, such as the one shown in Figure 1, and the *output* is a subset of replying or reposting messages covering salient content of the input post.

5.1 Data Collection and Experiment Set-up

We conduct an empirical study on the outputs of our joint model on microblog conversation summarization, whose data preparation and set-up processes are presented in this section.

Data Sets. Our experiments are conducted on a large-scale corpus containing ten large conversation trees collected from Sina Weibo, which is released by our prior work (Li et al. 2015a) and constructed following the settings described in Chang et al. (2013). The conversation trees discuss hot events taking place during 2 January to 28 July 2014,

23 On Twitter, a hashtag serves as a special URL, which can link other messages sharing the same hashtag.

Table 6
Description of the ten conversation trees for summarization. Each line describes the statistic information of one conversation.

# of messages	Height	Description
21,353	16	HKU dropping out student wins the college entrance exam again.
9,616	11	German boy complains hard schoolwork in Chinese High School.
13,087	8	Movie Tiny Times 1.0 wins high grossing in criticism.
12,865	8	"I am A Singer" states that singer G.E.M asking for resinging conforms to rules.
10,666	8	Crystal Huang clarified the rumor of her derailment.
21,127	11	Germany routs Brazil 7:1 in World-Cup semi-final.
18,974	13	The pretty girl pregnant with a second baby graduated with her master's degree.
2,021	18	Girls appealed for equality between men and women in college admission.
9,230	14	Violent terrorist attack in Kunming railway station.
10,052	25	MH17 crash killed many top HIV researchers.

and are crawled using the PKUVIS toolkit (Ren et al. 2014). The detailed descriptions of the ten conversation trees are shown in Table 6. As can be observed, more than 12K messages on average and covers discussions about social issues, breaking news, jokes, celebrity scandals, love, and fashion, which matches the official list of typical categories for microblog posts released by Sina Weibo.[24] For each conversation tree, three experienced editors are invited to write summaries. Based on the manual summaries written by them, we conduct ROUGE evaluation, shown in Section 5.2.

Though compared with many other tasks in the NLP and information retrieval community, the corpus looks relatively small. However, to the best of our knowledge, it is currently the only publicly available data set for conversation summarization.[25] It is difficult and time-consuming for human editors to write summaries for conversation trees because of the substantial number of nodes and complex structure involved (Chang et al. 2013); in fact, it would be impossible for human editors to reconstruct conversation trees in a reasonable amount of time. In the evaluation for each tree, we compute the average ROUGE F1 score between the model-generated summary and the three human-generated summaries.

Summary Extraction. Here we describe how summaries are produced given the outputs of topics models. For each conversation tree c, given the latent topics produced by topic models, we use a content word distribution γ_c to describe its core focus and topic. Equation (1) shows the formula to compute γ_c.

$$\gamma_{c,v} = Pr(v \text{ is a TOPIC word in } c) = \sum_{k=1}^{K} \theta_{c,k} \cdot \phi_{k,v}^{T} \tag{1}$$

We further plug in γ_c to the criterion proposed by Haghighi and Vanderwende (2009). The goal is to extract L messages to form a summary set E_c^* that closely matches

24 d.weibo.com/.
25 The corpus of Chang et al. (2013) is not publicly available.

γ_c. In our joint model, salient content of tree c is captured without including background noise (modeled with ϕ^B) or discourse indicative words (modeled with δ_d^D). Following Haghighi and Vanderwende (2009), conversation summarization is cast into the following Integer Programming problem:

$$E_c^* = \arg\min_{|E_c|=L} KL(\gamma_c||U(E_c)) \quad (2)$$

where $U(E_c)$ denotes the empirical unigram distribution of the candidate summary set E_c and $KL(P||Q)$ is the Kullback-Lieber (KL) divergence defined as $\sum_w P(w)\log\frac{P(w)}{Q(w)}$.[26] In implementation, as globally optimizing Equation (2) is exponential in the total number of messages in a conversation, which is a non-deterministic polynomial-time (NP) problem, we use the greedy approximation adopted in Haghighi and Vanderwende (2009) for local optimization. Specifically, messages are greedily added to a summary so long as they minimize the KL-divergence in the current step.

Comparisons. We consider baselines that rank and select messages by (1) LENGTH; (2) POPULARITY (# of reposts and replies); (3) USER influence (# of authors' followers); and (4) message–message text similarities using LEXRANK (Erkan and Radev 2004). We also consider two state-of-the-art summarizers in comparison: (1) CHANG ET AL. (2013), a fully *supervised* summarizer with manually crafted features; and (2) LI ET AL. (2015A), a random walk variant summarizer incorporating outputs of *supervised* discourse tagger. In addition, we compare the summaries extracted based on the topics yielded by our TOPIC+DISC+REL model with those based on the outputs of its variants (i.e., TOPIC ONLY and TOPIC+DISC).

Preprocessing. For baselines and the two state-of-the-art summarizers, we filter out non-Chinese characters in a preprocessing step following their common settings.[27] For summarization systems based on our topic model variants (i.e., TOPIC ONLY, TOPIC+DISC, and TOPIC+DISC+REL), the hyper-parameters and preprocessing steps are kept the same as in Section 4.1.

5.2 ROUGE Comparison

We quantitatively evaluate the performance of summarizers using ROUGE scores (Lin 2004) as a benchmark, a widely used standard for automatic summarization evaluation based on the overlapping units between a produced summary and a gold-standard reference. Specifically, Table 7 reports ROUGE-1, ROUGE-2, ROUGE-L, and ROUGE-SU4 output by ROUGE 1.5.5.[28] From the results, we can observe that:

26 To ensure the value of KL-divergence to be finite, we smooth $U(E_c)$ with β, which also serves as the smoothing parameter of ϕ_k^T (Section 3).

27 We have also conducted evaluations on the versions without this preprocessing step, and they gave worse ROUGE scores.

28 `github.com/summanlp/evaluation/tree/master/ROUGE-RELEASE-1.5.5`. Note that the absolute scores of comparison models here are different from those reported in Li et al. (2015a). Because the ROUGE scores reported here are given by ROUGE 1.5.5, whereas Li et al. (2015a) uses the Dragon toolkit (Zhou, Zhang, and Hu 2007) for ROUGE calculation. Despite the difference in absolute scores, the trends reported here remain similar to Li et al. (2015a).

- ***Simple features are not effective for summarization.*** The poor performance of all baselines demonstrates that microblog summarization is a challenging task. It is not possible to trivially rely on simple features such as length, message popularity, user influence, or text similarities to identify summary-worthy messages because of the colloquialness, noise, and redundancy exhibited in microblog texts.
- ***Discourse can indicate summary-worthy content.*** The summarization system based on the TOPIC+DISC+REL model has generally better ROUGE scores than the TOPIC+DISC based system. It also yields competitive and even slightly better results than Li et al. (2015a), which relies on a

Table 7
Average ROUGE score for model-produced summaries against the three human-generated references. Len = count of Chinese characters in the extracted summary; Prec, Rec, and F1 = average precision, recall, and F1 ROUGE measures over the ten conversation trees (%). The best scores in each setting is highlighted in **bold**. Higher scores indicate better results.

Models	Len	ROUGE-1			ROUGE-2		
		Prec	Rec	F1	Prec	Rec	F1
Baselines							
LENGTH	95.4	19.6	**53.2**	28.1	5.1	**14.3**	7.3
POPULARITY	27.2	33.8	25.3	27.9	8.6	6.1	6.8
USER	37.6	32.2	34.2	32.5	8.0	8.9	8.2
LEXRANK	25.7	**35.3**	22.2	25.8	11.7	6.9	8.3
State-of-the-art							
CHANG ET AL. (2013)	68.6	25.4	48.3	32.8	7.0	13.4	9.1
LI ET AL. (2015A)	58.6	27.3	45.4	33.7	7.6	12.6	9.3
Our models							
TOPIC ONLY	48.6	30.4	40.4	33.6	9.2	12.0	10.0
TOPIC+DISC	37.8	**38.1**	35.5	33.1	**13.2**	11.5	**10.8**
TOPIC+DISC+REL	48.9	32.3	41.3	**34.0**	10.3	12.5	10.5

Models	Len	ROUGE-L			ROUGE-SU4		
		Prec	Rec	F1	Prec	Rec	F1
Baselines							
LENGTH	95.4	16.4	**44.4**	23.4	6.2	**17.2**	8.9
POPULARITY	27.2	28.6	21.3	23.6	10.4	7.6	8.4
USER	37.6	28.0	29.6	28.2	9.8	10.6	10.0
LEXRANK	25.7	**30.6**	18.8	22.1	**12.3**	7.5	8.8
State-of-the-art							
CHANG ET AL. (2013)	68.6	21.6	41.1	27.9	8.3	16.0	10.8
LI ET AL. (2015A)	58.6	23.3	38.6	28.7	8.8	14.7	10.9
Our models							
TOPIC ONLY	48.6	26.3	34.9	29.0	10.2	13.8	11.3
TOPIC+DISC	37.8	**33.3**	30.7	28.6	**13.3**	12.2	11.3
TOPIC+DISC+REL	48.9	28.0	35.4	**29.3**	10.9	14.0	**11.5**

supervised discourse tagger. These observations demonstrate that TOPIC+DISC+REL, without requiring gold-standard discourse annotation, is able to discover the discourse roles that are likely to convey topical words, which further reflect salient content for conversation summarization.

- ***Directly applying the outputs of our joint model of discourse and topics to summarization might not be perfect.*** In general, the TOPIC+DISC+REL based system achieves the best F1 scores in ROUGE comparison, which implies that the yielded discourse and topic representations can somehow indicate summary-worthy content, although large margin improvements are not observed. In Section 5.4, we will analyze the errors and present a potential solution for further improving summarization results.

5.3 Human Evaluation Results

To further evaluate the generated summaries, we conduct human evaluations on informativeness (Info), conciseness (Conc), and readability (Read) of the extracted summaries. Two native Chinese speakers are invited to read the output summaries and subjectively rate on a 1–5 Likert scale and in 0.5 units, where a higher rating indicates better quality. Their overall inter-rater agreement achieves Krippendorff's α of 0.73, which indicates reliable results (Krippendorff 2004). Table 8 shows the average ratings by the two raters and over ten conversation trees.

As can be seen, despite of the closing results produced by supervised and well-performed unsupervised systems in automatic ROUGE evaluation (shown in Section 5.2), when the outputs are judged by humans, supervised systems CHANG ET AL.

Table 8
Overall human ratings on summaries produced by varying summarization systems. Info, Conc, Read are short forms of informativeness, conciseness, and readability, respectively. The ratings are given in a 1–5 Likert scale. Higher scores indicate better ratings. The best score in each setting is highlighted in **bold**.

Models	Info	Conc	Read
Baselines			
LENGTH	2.33	2.93	2.28
POPULARITY	2.38	2.35	3.05
USER	3.13	3.10	3.75
LEXRANK	3.05	2.70	3.03
State-of-the-art			
CHANG ET AL. (2013)	3.43	3.50	3.70
LI ET AL. (2015A)	**3.70**	**3.90**	**4.15**
Our models			
TOPIC ONLY	3.33	3.03	3.35
TOPIC+DISC	3.25	3.15	3.55
TOPIC+DISC+REL	3.35	3.28	3.73

(2013) and LI ET AL. (2015A), with supervision on summarization and discourse respectively, achieve much higher ratings than unsupervised systems on all three criteria. This observation demonstrates that microblog conversation summarization is essentially challenging, where manual annotations (although with high cost in time and efforts involved) can provide useful clues in guiding systems to produce summaries that will be liked by humans. Particularly, the ratings given to LI ET AL. (2015A) are higher than all systems in comparison by large margins, which indicates that the human-annotated discourse can well indicate summary-worthy content and also confirms the usefulness of considering discourse in microblog conversation summarization.

Among unsupervised methods, the summarization results based on our TOPIC+DISC+REL model achieves generally better ratings than other comparison methods. The possible reasons are: (1) When separating topic words from discourse and background words, it also filters out irrelevant noise and distills important content. (2) It can exploit the tendencies of messages with varying discourse roles in containing core content, thus is able to identify "bad" discourse roles that bring redundancy or irrelevant words, which disturbs reading experience.

To further analyze the generated summaries, we conduct a case study and display a sample summary, generated based on the TOPIC+DISC+REL model in Table 9. In this case, the input conversation is about the sexism issue in Chinese college entrance. As we can see, the produced summary covers salient comments that are helpful in understanding public opinions toward the gender discrimination problem. However, taking a closer look at the produced summaries, we observe that the system selects messages that contain sentiment-only information, such as "Good point! I have to repost this!," and therefore affects the quality of the generated summary. The observation from this summary case suggests that, in addition to discourse and background, a sentiment component should be effectively captured and well separated for further improving the summarization results. The potential extension of the current summarization system to additionally incorporate sentiment will be discussed in Section 5.4.

5.4 Error Analysis and Further Discussions

Taking a closer look at our produced summaries, one major source of incorrect selection of summary-worthy messages is based on the fact that sentiment is prevalent on microblog conversations, such as "love" in [R5] and "poor" in [R6] of Figure 1. Without an additional separation of sentiment-specific information, the yielded topic representations might be mixed with sentiment components. For example, in Table 4, the topic generated by the TOPIC+DISC+REL model contains sentiment words like "wrong" and "hate." Therefore, a direct use of the topic representations to extract summaries will unavoidably select messages that mostly reflect sentimental component, which is also illustrated by the case study in Section 5.3.

Therefore, we argue that reliable estimation of the summary-worthy content in microblog conversations requires additional consideration of sentiment, because sentiment can also be represented by word distributions and captured via topic models in an unsupervised or weakly supervised manner (Lin and He 2009; Jo and Oh 2011; Lim and Buntine 2014). In future work, we can propose another model based on our joint model TOPIC+DISC+REL that can additionally separate sentiment word representations from discourse and topics. Lazaridou, Titov, and Sporleder (2013) have demonstrated that sentiment shifts can indicate sentence-level discourse functions in product reviews. We can then hypothesize that modeling discourse roles of messages can also benefit

Table 9
The sample summary produced by the TOPIC+DISC+REL-based summarization system. For each message, we display the original Chinese version followed by its English translation.

Root message of the conversation:

近日，各高校招生录取分数纷纷出炉，国际关系学院等院校分数线设置女高男低，引起了广州一位女大学生的关注。她认为这种做法对女考生很不公平，于是写信给国家主席习近平，希望能关注高考录取中的性别歧视现象，重视女性在科技国防军事中的力量。
Recently, the admission criteria for colleges are coming out. Women should get better grades in College Entrance Exam to go to colleges like University of International Relations. A female undergraduate student in Guangzhou was concerned about the unfair treatment. She wrote a letter to President Xi Jinping for reducing gender discrimination in college admission and emphasized the important role female plays in technology and military.

The produced summary:

以保护之名提高女性受教育门槛，实质上是一种"把女性视为弱者"的社会刻板印象作祟，这违背了联合国《消除对妇女一切形式歧视公约》中"保护性歧视"的规定。再者，"本专业需要多熬夜女生吃不消"这一理由并不正当，难道分数线以上考进去的女生的生理健康就不需要保护了吗？分数高的学生更能熬夜？ Raising the bar for women to get education in order to protect them is ascribed to a stereotype of "*women are weaker sex*." This is "*special protections for women*" in "*The Convention on the Elimination of all Forms of Discrimination Against Women*" released by UN. Besides, "*students in our department should stay up to learn*" is not an appropriate reason. What about their female students? Don't they have to take care of their physical health? Or students achieving higher grades don't need much sleep?
嗯其实…要是没有分数差不限制男女比例的话…学校里男生又会特别少抱怨的还会是我们妹子自己啦…所以…多方面看吧 In fact, we need to use different admission criteria to avoid gender imbalance. If a college has too few boys, girls will complain. Every coin has two sides.
因为大学都承认男人就是不如女人啊，呵呵 Because colleges admit that women are better than men, hehe.
以前看到的一则新闻说的是为了调整语言类专业的男女比例，下调男考生的录取分数线。如果像这样以女生体质为借口，那同样寒窗苦读十二载，女生的成绩分量不应该更重才对么？希望习大大能看到吧 An earlier news reported that men could be admitted with worse grades than women for encouraging men to study language. If women do have worse physical condition, then it is more difficult for women to get the same grades as men. Women should have lower bar in college admission. I hope President Xi can see this.
说怕体力吃不消严格要求体育分也就罢了，文化分数高低能作为一个人适不适合一个工作辛苦的岗位的理由么？ If they are concerned about physical conditions of women, then they should require a test in PE. Why use paper based exam to test physical conditions?
说得好！必须转～ Good point! I have to repost this.
哈哈哈，国关课业繁重经常熬夜～ Hahaha, the workload in International Relations is so heavy that students should stay up to learn.
女性普遍比男生更努力却换来不同的的结果，要我说男女平等，男性角色弱化无可厚非，抱着几千年的传统观念看今天的男生是不行的，男孩危机是个伪命题，况且真有本事不至于就在学校受益 Generally, women work harder than men but have worse endings. For gender equality, it is alright to weaken the role of men. We should have a different view on the boys today. "*The boy crisis*" is nonsense. Besides, you can still be a great guy without education.
这难道就是女性为什么越来越优秀，而男性越来越丝的部分原因？呵呵。男同胞要感谢性别歧视，让他们越来越弱了。 Isn't this part of the reason why women become more and more excellent while men go to the opposite direction? Interesting. Men should appreciate for sexism, which makes them weaker and weaker.

from exploring sentiment shifts in conversations. As it might be out of the scope of this article to thoroughly explore the joint effects of topic, discourse, and sentiment on microblog conversation summarization, we hence leave the study on such extended model to future work.

6. Conclusion and Future Work

In this article, we have presented a novel topic model for microblog messages that allows the joint induction of conversational discourse and latent topics in a fully unsupervised manner. By comparing our joint model with a number of competitive topic models on real-world microblog data sets, we have demonstrated the effectiveness of using conversational discourse structure to help in identifying topical content embedded in short and colloquial microblog messages. Moreover, our empirical study on microblog conversation summarization has shown that the produced discourse and topical representations can also predict summary-worthy content. Both ROUGE evaluation and human assessment have demonstrated that the summaries generated based on the outputs of our joint model are informative, concise, and easy-to-read. Error analysis on the produced summaries has shown that sentiment should be effectively captured and separated to further advance our current summarization system forward. As a result, the joint effects of discourse, topic, and sentiment on microblog conversation summarization is worth exploring in future study.

For other lines of future work, one potential is to extend our joint model to identify topic hierarchies from microblog conversation trees. In doing so, one could learn how topics change in a microblog conversation along a certain hierarchical path. Another potential line is to combine our work with representation learning on social media. Although some previous studies have provided intriguing approaches to learning representations at the level of words (Mikolov et al. 2013; Mikolov, Yih, and Zweig 2013), sentences (Le and Mikolov 2014), and paragraphs (Kiros et al. 2015), they are limited in modeling social media content with colloquial relations. Following similar ideas in this work, where discourse and topics are jointly explored, we can conduct other types of representation learning, embeddings for words (Li et al. 2017b), messages (Dhingra et al. 2016), or users (Ding, Bickel, and Pan 2017), in the context of conversations, which should complement social media representation learning and vice versa.

Appendix A

In this section, we present the key steps for inferring our joint model of conversational discourse and latent topics. Its generation process has been described in Section 3. As described in Section 3, we use collapsed Gibbs sampling (Griffiths et al. 2004) for model inference. Before providing the formula of sampling steps, we first define the notations of all variables used in the formulations of Gibbs sampling, described in Table A.1. In particular, the various $\mathbb{C}$ variables refer to counts excluding the message m on conversation tree c.

Table A.1
The notations of symbols in the sampling formulas, Equations (A.1) and (A.2) (c, m): message m on conversation tree c.

x	word-level word type switcher. $x = 1$: discourse word (DISC); $x = 2$: topic word (TOPIC); $x = 3$: background word (BACK).
$\mathbb{C}^{DX}_{d,(x)}$	# of words with word type as x and occurring in messages with discourse d.
$\mathbb{C}^{DX}_{d,(\cdot)}$	# of words that occur in messages whose discourse assignments are d, i.e., $\sum_{x=1}^{3} \mathbb{C}^{DX}_{d,(x)}$.
$\mathbb{N}^{DX}_{(x)}$	# of words occurring in message (c, m) and with word type assignment as x.
$\mathbb{N}^{DX}_{(\cdot)}$	# of words in message (c, m), i.e., $\mathbb{N}^{DX}_{(\cdot)} = \sum_{x=1}^{3} \mathbb{N}^{DX}_{(x)}$.
$\mathbb{C}^{DW}_{d,(v)}$	# of words indexing v in vocabulary, assigned as discourse word, and occurring in messages assigned discourse d.
$\mathbb{C}^{DW}_{d,(\cdot)}$	# of words assigned as discourse words (DISC) and occurring in messages assigned as discourse d, i.e., $\mathbb{C}^{DW}_{d,(\cdot)} = \sum_{v=1}^{V} \mathbb{C}^{DW}_{d,(v)}$.
$\mathbb{N}^{TW}_{(v)}$	# of words indexing v in vocabulary that occur in messages (c, m) and are assigned as topic words (TOPIC).
$\mathbb{N}^{TW}_{(\cdot)}$	# of words assigned as topic words (TOPIC) and occurring in message (c, m), i.e., $\mathbb{N}^{TW}_{(\cdot)} = \sum_{v=1}^{V} \mathbb{N}^{TW}_{(v)}$.
$\mathbb{N}^{DW}_{(v)}$	# of words indexing v in vocabulary that occur in messages (c, m) and are assigned as discourse words (DISC).
$\mathbb{N}^{DW}_{(\cdot)}$	# of words assigned as discourse words (DISC) and occurring in message (c, m), i.e., $\mathbb{N}^{DW}_{(\cdot)} = \sum_{v=1}^{V} \mathbb{N}^{DW}_{(v)}$.
$\mathbb{C}^{DD}_{d,(d')}$	# of messages assigned discourse d' whose parent is assigned discourse d.
$\mathbb{C}^{DD}_{d,(\cdot)}$	# of messages whose parents are assigned discourse d, i.e., $\mathbb{C}^{DD}_{d,(\cdot)} = \sum_{d'=1}^{D} \mathbb{C}^{DD}_{d,(d')}$.
$I(\cdot)$	An indicator function, whose value is 1 when its argument inside () is true, and 0 otherwise.
$\mathbb{N}^{DD}_{(d)}$	# of messages whose parent is (c, m) and assigned discourse d.
$\mathbb{N}^{DD}_{(\cdot)}$	# of messages whose parent is (c, m), i.e., $\mathbb{N}^{DD}_{(\cdot)} = \sum_{d=1}^{D} \mathbb{N}^{DD}_{(d)}$
$\mathbb{C}^{BW}_{(v)}$	# of words indexing v in vocabulary and assigned as background words (BACK)
$\mathbb{C}^{BW}_{(\cdot)}$	# of words assigned as background words (BACK), i.e., $\mathbb{C}^{BW}_{(\cdot)} = \sum_{v=1}^{V} \mathbb{C}^{BW}_{(v)}$
$\mathbb{C}^{CT}_{c,(k)}$	# of messages on conversation tree c and assigned topic k.
$\mathbb{C}^{CT}_{c,(\cdot)}$	# of messages on conversation tree c, i.e., $\mathbb{C}^{CT}_{c,(\cdot)} = \sum_{k=1}^{K} \mathbb{C}^{CT}_{c,(k)}$
$\mathbb{C}^{TW}_{k,(v)}$	# of words indexing v in vocabulary, sampled as topic words (TOPIC), and occurring in messages assigned topic k.
$\mathbb{C}^{TW}_{k,(\cdot)}$	# of words assigned as topic word and occurring in messages assigned topics k (TOPIC), i.e., $\mathbb{C}^{TW}_{k,(\cdot)} = \sum_{v=1}^{V} \mathbb{C}^{TW}_{k,(v)}$.

For each message m on conversation tree c, we sample its discourse role $d_{c,m}$ and topic assignment $z_{c,m}$, according to the following conditional probability distribution:

$$
\begin{aligned}
&p(d_{c,m} = d, z_{c,m} = k | \mathbf{d}_{\neg(c,m)}, \mathbf{z}_{\neg(c,m)}, \mathbf{w}, \mathbf{x}, \Theta) \\
\propto &\frac{\Gamma(\mathbb{C}^{DD}_{d_{c,pa(m)},(\cdot)} + D \cdot \gamma)}{\Gamma(\mathbb{C}^{DD}_{d_{c,pa(m)},(\cdot)} + I(d_{c,pa(m)} \neq d) + D \cdot \gamma)} \cdot \frac{\Gamma(\mathbb{C}^{DD}_{d_{c,pa(m)},(d)} + I(d_{c,pa(m)} \neq d) + \gamma)}{\Gamma(\mathbb{C}^{DD}_{d_{c,pa(m)},(d)} + \gamma)} \\
&\cdot \frac{\Gamma(\mathbb{C}^{DD}_{d,(\cdot)} + D \cdot \gamma)}{\Gamma(\mathbb{C}^{DD}_{d,(\cdot)} + I(d_{c,pa(m)} = d) + \mathbb{N}^{DD}_{(\cdot)} + D \cdot \gamma)} \cdot \prod_{d'=1}^{D} \frac{\Gamma(\mathbb{C}^{DD}_{d,(d')} + \mathbb{N}^{DD}_{(d')} + I(d_{c,pa(m)} = d = d') + \gamma)}{\Gamma(\mathbb{C}^{DD}_{d,(d')} + \gamma)} \\
&\cdot \frac{\mathbb{C}^{CT}_{c,(k)} + \alpha}{\mathbb{C}^{CT}_{c,(\cdot)} + K \cdot \alpha} \cdot \frac{\Gamma(\mathbb{C}^{TW}_{k,(\cdot)} + V \cdot \beta)}{\Gamma(\mathbb{C}^{TW}_{k,(\cdot)} + \mathbb{N}^{TW}_{(\cdot)} + V \cdot \beta)} \cdot \prod_{v=1}^{V} \frac{\Gamma(\mathbb{C}^{TW}_{k,(v)} + \mathbb{N}^{TW}_{(v)} + \beta)}{\Gamma(\mathbb{C}^{TW}_{k,(v)} + \beta)} \\
&\cdot \frac{\Gamma(\mathbb{C}^{DX}_{d,(\cdot)} + 3 \cdot \delta)}{\Gamma(\mathbb{C}^{DX}_{d,(\cdot)} + \mathbb{N}^{DX}_{(\cdot)} + 3 \cdot \delta)} \cdot \prod_{x=1}^{3} \frac{\Gamma(\mathbb{C}^{DX}_{d,(x)} + \mathbb{N}^{DX}_{(x)} + \delta)}{\Gamma(\mathbb{C}^{DX}_{d,(x)} + \delta)} \\
&\cdot \frac{\Gamma(\mathbb{C}^{DW}_{d,(\cdot)} + V \cdot \beta)}{\Gamma(\mathbb{C}^{DW}_{d,(\cdot)} + \mathbb{N}^{DW}_{(\cdot)} + V \cdot \beta)} \cdot \prod_{v=1}^{V} \frac{\Gamma(\mathbb{C}^{DW}_{d,(v)} + \mathbb{N}^{DW}_{(v)} + \beta)}{\Gamma(\mathbb{C}^{DW}_{d,(v)} + \beta)}
\end{aligned}
\tag{A.1}
$$

where the discourse role and topic assignments of message m on conversation c are determined by: (1) the discourse role assignments of the parent and all the children of message m on conversation c (shown in the first four factors); (2) the topic mixture of conversation tree c (shown in the fifth factor); (3) the topic assignments of other messages sharing TOPIC words with m (shown in the sixth and the seventh factor); (4) the distribution of words in m as DISC, TOPIC, and BACK words (shown in the eigth and the ninth factor); and (5) the discourse role assignments of other messages sharing DISC words with m (shown in the last two factors).

For each word n in m on c, the sampling formula of its word type $x_{c,m,n}$ (as discourse [DISC], topic [TOPIC], and background [BACK]) is given as the following:

$$
\begin{aligned}
&p(x_{c,m,n} = x | \mathbf{x}_{\neg(c,m,n)}, \mathbf{d}, \mathbf{z}, \mathbf{w}, \Theta) \\
\propto &\frac{\mathbb{C}^{DX}_{d_{c,m},(x)} + \delta}{\mathbb{C}^{DX}_{d_{c,m},(\cdot)} + 3 \cdot \delta} \cdot g(x, c, m, n)
\end{aligned}
\tag{A.2}
$$

where

$$
g(x, c, m, n) = \begin{cases} \frac{\mathbb{C}^{DW}_{d_{c,m},(w_{c,m,n})} + \beta}{\mathbb{C}^{DW}_{d_{c,m},(\cdot)} + V \cdot \beta} & \text{if } x == \text{DISC} \\ \frac{\mathbb{C}^{TW}_{z_{c,m},(w_{c,m,n})} + \beta}{\mathbb{C}^{TW}_{z_{c,m},(\cdot)} + V \cdot \beta} & \text{if } x == \text{TOPIC} \\ \frac{\mathbb{C}^{BW}_{(w_{c,m,n})} + \beta}{\mathbb{C}^{BW}_{(\cdot)} + V \cdot \beta} & \text{if } x == \text{BACK} \end{cases}
\tag{A.3}
$$

Here the word type switcher of word n in message m on conversation c is determined by: (1) the distribution of word types in messages sharing the same discourse role as m (shown in the first factor); and (2) the word types of word $w_{c,m,n}$ appearing elsewhere (shown in the second factor $g(x, c, m, n)$).

Acknowledgments

This work is partially supported by Innovation and Technology Fund (ITF) project no. 6904333, General Research Fund (GRF) project no. 14232816 (12183516), National Natural Science Foundation of China (grant no. 61702106), and Shanghai Science and Technology Commission (grant no. 17JC1420200 and grant no. 17YF1427600). We are grateful for the contributions of Yulan He, Lu Wang, and Wei Gao in shaping part of our ideas, and the efforts of Nicholas Beautramp, Sarah Shugars, Ming Liao, Xingshan Zeng, Shichao Dong, and Dingmin Wang in preparing some of the experiment data. Also, we thank Shuming Shi, Dong Yu, Tong Zhang, and the three anonymous reviewers for the insightful suggestions on various aspects of this work.

References

Alvarez-Melis, David 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.

Bakker, Stéphanie J. and Gerrigje Catharina Wakker. 2009. *Discourse Cohesion in Ancient Greek*. Brill.

Baldridge, Jason and Alex Lascarides. 2005. Probabilistic head-driven parsing for discourse structure. In *Proceedings of the Ninth Conference on Computational Natural Language Learning*, pages 96–103, Ann Arbor, MI.

Bangalore, Srinivas, Giuseppe Di Fabbrizio, and Amanda Stent. 2006. Learning the structure of task-driven human-human dialogs. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 201–208, Sydney.

Bhatia, Sumit, Prakhar Biyani, and Prasenjit Mitra. 2014. Summarizing online forum discussions—Can dialog acts of individual messages help? In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014*, pages 2127–2131, Doha.

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

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

Bokaei, Mohammad Hadi, Hossein Sameti, and Yang Liu. 2016. Extractive summarization of multi-party meetings through discourse segmentation. *Natural Language Engineering*, 22(1):41–72.

Bollegala, Danushka, Takanori Maehara, and Ken-ichi Kawarabayashi. 2015. Unsupervised cross-domain word representation learning. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the Seventh International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 730–740, Beijing.

Boyd-Graber, Jordan L., Yuening Hu, and David M. Mimno. 2017. Applications of topic models. *Foundations and Trends in Information Retrieval*, 11(2-3):143–296.

Carlson, Lynn, Daniel Marcu, and Mary Ellen Okurovsky. 2001. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In *Proceedings of the Second SIGdial Workshop on Discourse and Dialogue*, pages 1–10, Aalborg.

Çelikyilmaz, Asli and Dilek Hakkani-Tür. 2010. A hybrid hierarchical model for multi-document summarization. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 815–824, Uppsala.

Chakrabarti, Deepayan and Kunal Punera. 2011. Event summarization using tweets. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, pages 66–73, Barcelona.

Chang, Jonathan, Jordan L. Boyd-Graber, Sean Gerrish, Chong Wang, and David M. Blei. 2009. Reading tea leaves: How humans interpret topic models. In *Advances in the 23rd Annual Conference on Neural Information Processing Systems*, pages 288–296, Vancouver.

Chang, Yi, Xuanhui Wang, Qiaozhu Mei, and Yan Liu. 2013. Towards Twitter context summarization with user influence models. In *Proceedings of the Sixth ACM International Conference on Web Search and Data Mining*, pages 527–536, Rome.

Chemudugunta, Chaitanya, Padhraic Smyth, and Mark Steyvers. 2006. Modeling general and specific aspects of documents with a probabilistic topic model. In *Advances in the 20th Annual Conference on Neural Information Processing Systems*, pages 241–248, Vancouver.

Cheng, Xueqi, Xiaohui Yan, Yanyan Lan, and Jiafeng Guo. 2014. BTM: topic modeling over short texts. *IEEE Transactions on Knowledge and Data Engineering*, 26(12):2928–2941.

Cohen, William W., Vitor R. Carvalho, and Tom M. Mitchell. 2004. Learning to classify email into "speech act." In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pages 309–316, Barcelona.

Crook, Nigel, Ramón Granell, and Stephen G. Pulman. 2009. Unsupervised classification of dialogue acts using a Dirichlet process mixture model. In *Proceedings of the Tenth Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 341–348, London.

Das, Dipanjan and André FT Martins. 2007. A survey on automatic text summarization. *Literature Survey for the Language and Statistics II Course at CMU*, 4:192–195.

Das, Rajarshi, Manzil Zaheer, and Chris Dyer. 2015. Gaussian LDA for topic models with word embeddings. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing*, pages 795–804, Beijing.

Daumé III, Hal and Daniel Marcu. 2006. Bayesian query-focused summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 305–312, Sydney.

De Beaugrande, Robert and Wolfgang Dressler. 1981. *Textlinguistics*. Longman, New York.

Dhillon, Rajdip, Sonali Bhagat, Hannah Carvey, and Elizabeth Shriberg. 2004. Meeting recorder project: Dialog act labeling guide. Technical report, DTIC Document.

Dhingra, Bhuwan, Zhong Zhou, Dylan Fitzpatrick, Michael Muehl, and William W. Cohen. 2016. Tweet2vec: Character-based distributed representations for social media. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 269–274, Berlin.

Ding, Tao, Warren K. Bickel, and Shimei Pan. 2017. Multi-view unsupervised user feature embedding for social media-based substance use prediction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2275–2284, Copenhagen.

Duan, Yajuan, Zhimin Chen, Furu Wei, Ming Zhou, and Heung-Yeung Shum. 2012. Twitter topic summarization by ranking tweets using social influence and content quality. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 763–780, Mumbai.

Duan, Yajuan, Long Jiang, Tao Qin, Ming Zhou, and Heung-Yeung Shum. 2010. An empirical study on learning to rank of tweets. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 295–303, Mumbai.

Erkan, Günes and Dragomir R. Radev. 2004. LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22:457–479.

Farzindar, Atefeh and Diana Inkpen. 2015. *Natural Language Processing for Social Media*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers.

Feng, Vanessa Wei and Graeme Hirst. 2014. A linear-time bottom-up discourse parser with constraints and post-editing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 511–521, Baltimore, MD.

Fisher, Seeger and Brian Roark. 2007. The utility of parse-derived features for automatic discourse segmentation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*, Prague.

Gimpel, Kevin, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: Annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 42–47, Portland, OR.

Griffiths, Thomas L. and Mark Steyvers. 2004. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101:5228–5235.

Griffiths, Thomas L., Mark Steyvers, David M. Blei, and Joshua B. Tenenbaum. 2004. Integrating topics and syntax. In *Advances in the 18th Annual Conference on Neural Information Processing Systems*, pages 537–544, Vancouver.

Haghighi, Aria and Lucy Vanderwende. 2009. Exploring content models for multi-document summarization. In *Proceedings of Human Language*

Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, pages 362–370, Boulder, CO.
Harabagiu, Sanda M. and Andrew Hickl. 2011. Relevance modeling for microblog summarization. *Proceedings of the Fifth International Conference on Web and Social Media*, pages 514–517, Barcelona.
Hofmann, Thomas. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM Conference on Research and Development in Information Retrieval*, pages 50–57, Berkeley, CA.
Hong, Liangjie and Brian D. Davison. 2010. Empirical study of topic modeling in Twitter. In *Proceedings of the First Workshop on Social Media Analytics*, pages 80–88, Washington, DC.
Hovy, Eduard H. and Elisabeth Maier. 1995. Parsimonious or profligate: How many and which discourse structure relations. `https://www.cs.cmu.edu/~hovy/`
Inouye, David and Jugal K. Kalita. 2011. Comparing Twitter summarization algorithms for multiple post summaries. In *Proceedings of the 2011 IEEE Third International Conference on Privacy, Security, Risk and Trust and 2011 IEEE Third International Conference on Social Computing*, pages 298–306, Boston, MA.
Ji, Yangfeng and Jacob Eisenstein. 2014. Representation learning for text-level discourse parsing. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 13–24, Baltimore, MD.
Jo, Yohan and Alice H. Oh. 2011. Aspect and sentiment unification model for online review analysis. In *Proceedings of the Fourth International Conference on Web Search and Web Data Mining*, pages 815–824, Hong Kong.
Joty, Shafiq R., 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*, pages 1807–1813, Barcelona.
Joty, Shafiq R., Giuseppe Carenini, and Raymond T. Ng. 2012. A novel discriminative framework for sentence-level discourse analysis. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 904–915, Jeju Island.
Jurafsky, Dan, Elizabeth Shriberg, and Debra Biasca. 1997. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual. *Institute of Cognitive Science Technical Report*, pages 97–102.
Kiros, Ryan, Yukun Zhu, Ruslan Salakhutdinov, Richard S. Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in the 28th Annual Conference on Neural Information Processing Systems*, pages 3294–3302, Montréal.
Kouloumpis, Efthymios, Theresa Wilson, and Johanna D. Moore. 2011. Twitter sentiment analysis: The good the bad and the OMG! In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, pages 538–541, Barcelona.
Krippendorff, Klaus. 2004. *Content analysis: An introduction to its methodology*. SAGE.
Lazaridou, Angeliki, Ivan Titov, and Caroline Sporleder. 2013. A Bayesian model for joint unsupervised induction of sentiment, aspect and discourse representations. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1630–1639, Sofia.
Le, Quoc V. and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of the 31th International Conference on Machine Learning*, pages 1188–1196, Beijing.
Li, Chenliang, Yu Duan, Haoran Wang, Zhiqian Zhang, Aixin Sun, and Zongyang Ma. 2017a. Enhancing topic modeling for short texts with auxiliary word embeddings. *ACM Transactions on Information Systems*, 36(2):11:1–11:30.
Li, Chenliang, 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*, pages 165–174, Pisa.
Li, Jing, Wei Gao, Zhongyu Wei, Baolin Peng, and Kam-Fai Wong. 2015a. Using content-level structures for summarizing microblog repost trees. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2168–2178, Lisbon.
Li, Jing, 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*, pages 2112–2123, Berlin.
Li, Jing, Zhongyu Wei, Hao Wei, Kangfei Zhao, Junwen Chen, and Kam-Fai Wong. 2015b. Learning to rank microblog posts for real-time ad-hoc search. In *Proceedings*

of the 4th CCF Conference on Natural Language Processing and Chinese Computing, pages 436–443, Nanchang.

Li, Jiwei, Rumeng Li, and Eduard H. Hovy. 2014. Recursive deep models for discourse parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 2061–2069, Doha.

Li, Quanzhi, Sameena Shah, Xiaomo Liu, and Armineh Nourbakhsh. 2017b. Data sets: Word embeddings learned from tweets and general data. In *Proceedings of the 11th International Conference on Web and Social Media*, pages 428–436, Montréal.

Lim, Kar Wai and Wray L. Buntine. 2014. Twitter opinion topic model: Extracting product opinions from tweets by leveraging hashtags and sentiment lexicon. In *Proceedings of the 23rd ACM International Conference on Information and Knowledge Management*, pages 1319–1328, Shanghai.

Lin, Chenghua and Yulan He. 2009. Joint sentiment/topic model for sentiment analysis. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management*, pages 375–384, Hong Kong.

Lin, Chenghua, Ebuka Ibeke, Adam Z. Wyner, and Frank Guerin. 2015. Sentiment-topic modeling in text mining. *Wiley Interdisc. Rew.: Data Mining and Knowledge Discovery*, 5(5):246–254.

Lin, Chin-Yew. 2004. ROUGE: A package for automatic evaluation of summaries. In *Proceedings of the ACL-04 Workshop: Text Summarization Branches Out*, pages 74–81, Barcelona.

Lin, Cindy Xide, Bo Zhao, Qiaozhu Mei, and Jiawei Han. 2010. PET: a statistical model for popular events tracking in social communities. In *Proceedings of the 16th International Conference on Knowledge Discovery and Data Mining*, pages 929–938, Washington, DC.

Lin, Ziheng, Min-Yen Kan, and Hwee Tou Ng. 2009. Recognizing implicit discourse relations in the Penn discourse treebank. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 343–351, Singapore.

Liu, Fei, Yang Liu, and Fuliang Weng. 2011. Why is SXSW trending?: Exploring multiple text sources for Twitter topic summarization. In *Proceedings of the Workshop on Language in Social Media*, pages 66–75, Portland, OR.

Liu, Xiaohua, Yitong Li, Furu Wei, and Ming Zhou. 2012. Graph-based multi-tweet summarization using social signals. In *Proceedings of the 24th International Conference on Computational Linguistics*, pages 1699–1714, Mumbai.

Long, Rui, Haofen Wang, Yuqiang Chen, Ou Jin, and Yong Yu. 2011. Towards effective event detection, tracking and summarization on microblog data. In *Proceedings of the 12th International Conference on Web-Age Information Management*, pages 652–663, Wuhan.

Mann, William and Sandra Thompson. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text*, 8(3):243–281.

Marcu, Daniel. 2000. The rhetorical parsing of unrestricted texts: A surface-based approach. *Computational Linguistics*, 26(3):395–448.

McKeown, Kathleen, Judith Klavans, Vasileios Hatzivassiloglou, Regina Barzilay, and Eleazar Eskin. 1999. Towards multidocument summarization by reformulation: Progress and prospects. In *Proceedings of the 16th National Conference on Artificial Intelligence and the 11th Conference on Innovative Applications of Artificial Intelligence*, pages 453–460, Orlando, FL.

McKeown, Kathleen, Lokesh Shrestha, and Owen Rambow. 2007. Using question-answer pairs in extractive summarization of email conversations. In *Proceedings of the Eighth International Conference on Computational Linguistics and Intelligent Text Processing*, pages 542–550, Mexico City.

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

Meng, Xinfan, Furu Wei, Xiaohua Liu, Ming Zhou, Sujian Li, and Houfeng Wang. 2012. Entity-centric topic-oriented opinion summarization in Twitter. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 379–387, Beijing.

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Gregory S. Corrado, and Jeffrey Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in the 27th Annual Conference on Neural Information Processing Systems*, pages 3111–3119, Lake Tahoe, NV.

Mikolov, Tomas, Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in

continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 746–751, Atlanta, GA.

Mimno, David M., Hanna M. Wallach, Edmund M. Talley, Miriam Leenders, and Andrew McCallum. 2011. Optimizing semantic coherence in topic models. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 262–272, Edinburgh.

Murray, Gabriel, Steve Renals, Jean Carletta, and Johanna D. Moore. 2006. Incorporating speaker and discourse features into speech summarization. In *Proceedings of the 2006 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 367–374, New York, NY.

Nenkova, Ani and Kathleen McKeown. 2012. A survey of text summarization techniques. In *Mining Text Data*. Springer, pages 43–76.

Newman, David, Jey Han Lau, Karl Grieser, and Timothy Baldwin. 2010. Automatic evaluation of topic coherence. In *Proceedings of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 100–108, Los Angeles, CA.

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

Nigam, Kamal, 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.

Owoputi, Olutobi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A. Smith. 2013. Improved part-of-speech tagging for online conversational text with word clusters. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 380–390, Atlanta, GA.

Oya, Tatsuro and Giuseppe Carenini. 2014. Extractive summarization and dialogue act modeling on email threads: An integrated probabilistic approach. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 133–140, Philadelphia, PA.

Pak, Alexander and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh Conference on International Language Resources and Evaluation*, pages 1320–1326, Valletta.

Peng, Baolin, Jing Li, Junwen Chen, Xu Han, Ruifeng Xu, and Kam-Fai Wong. 2015. Trending sentiment-topic detection on Twitter. In *Proceedings of the 16th International Conference on Computational Linguistics and Intelligent Text*, pages 66–77, Cairo.

Perret, Jérémy, Stergos D. Afantenos, Nicholas Asher, and Mathieu Morey. 2016. Integer linear programming for discourse parsing. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 99–109, San Diego, CA.

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

Popescu, Ana-Maria and Marco Pennacchiotti. 2010. Detecting controversial events from Twitter. In *Proceedings of the 19th ACM Conference on Information and Knowledge Management*, pages 1873–1876, Toronto.

Prasad, Rashmi, Nikhil Dinesh, Alan Lee, Eleni Miltsakaki, Livio Robaldo, Aravind K. Joshi, and Bonnie L. Webber. 2008. The Penn discourse treebank 2.0. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation*, pages 2961–2968, Marrakech.

Qin, Kechen, 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*, pages 974–984, Vancouver.

Qiu, Xipeng, Qi Zhang, and Xuanjing Huang. 2013. FudanNLP: A toolkit for Chinese natural language processing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 49–54, Sofia.

Quan, Xiaojun, Chunyu Kit, Yong Ge, and Sinno Jialin Pan. 2015. Short and sparse text topic modeling via self-aggregation. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence*, pages 2270–2276, Buenos Aires.

Radev, Dragomir R., Timothy Allison, Sasha Blair-Goldensohn, John Blitzer, Arda Çelebi, Stanko Dimitrov, Elliott Drábek, Ali Hakim, Wai Lam, Danyu Liu, Jahna Otterbacher, Hong Qi, Horacio Saggion, Simone Teufel, Michael Topper, Adam Winkel, and Zhu Zhang. 2004. MEAD—A platform for multidocument multilingual text summarization. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation*, pages 699–702, Lisbon.

Radev, Dragomir R., Eduard H. Hovy, and Kathleen McKeown. 2002. Introduction to the special issue on summarization. *Computational Linguistics*, 28(4):399–408.

Ramage, Daniel, Susan T. Dumais, and Daniel J. Liebling. 2010. Characterizing microblogs with topic models. In *Proceedings of the Fourth International Conference on Web and Social Media*, pages 130–137, Washington, DC.

Rapoza, Kenneth. 2011. China's Weibos vs US's Twitter: And the winner is? *Forbes* 17 May.

Ren, Donghao, Xin Zhang, Zhenhuang Wang, Jing Li, and Xiaoru Yuan. 2014. WeiboEvents: A crowd sourcing Weibo visual analytic system. In *Proceedings of the Seventh IEEE Pacific Visualization Symposium*, pages 330–334, Yokohama.

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

Röder, Michael, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pages 399–408, Shanghai.

Rosa, Kevin Dela, Rushin Shah, Bo Lin, Anatole Gershman, and Robert Frederking. 2011. Topical clustering of tweets. In *Proceedings of the ACM SIGIR 3rd Workshop on Social Web Search and Mining*, Beijing. `https://www.cs.cmu.edu/~/sigir-swsm-2011.pdf`

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

Salton, Gerard, Amit Singhal, Mandar Mitra, and Chris Buckley. 1997. Automatic text structuring and summarization. *Information Processing and Management*, 33(2):193–207.

Sharifi, Beaux, Mark-Anthony Hutton, and Jugal Kalita. 2010. Automatic summarization of Twitter topics. Paper presented at the National Workshop on Design and Analysis of Algorithm, Tezpur, Assam.

Shen, Chao, Fei Liu, Fuliang Weng, and Tao Li. 2013. A participant-based approach for event summarization using Twitter streams. In *Proceedings of the 2013 Conference of the North American Chapter of the Association of Computational Linguistics: Human Language Technologies*, pages 1152–1162, Atlanta, GA.

Shi, Bei, 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.

Siddharthan, Advaith, Ani Nenkova, and Kathleen McKeown. 2004. Syntactic simplification for improving content selection in multi-document summarization. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 896–902, Geneva.

Soricut, Radu and Daniel Marcu. 2003. Sentence level discourse parsing using syntactic and lexical information. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 149–156, Edmonton.

Stevens, Keith, W. Philip Kegelmeyer, David Andrzejewski, and David Buttler. 2012. Exploring topic coherence over many models and many topics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 952–961, Jeju Island.

Stolcke, Andreas, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca A. Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational Linguistics*, 26(3):339–373.

Subba, Rajen and Barbara Di Eugenio. 2009. An effective discourse parser that uses rich linguistic information. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American*

Chapter of the Association for Computational Linguistics, pages 566–574, Boulder, CO.
Takamura, Hiroya, Hikaru Yokono, and Manabu Okumura. 2011. Summarizing a document stream. In *Advances in the 33rd European Conference on Information Retrieval*, pages 177–188, Dublin.
Thanh, Huong Lê, Geetha Abeysinghe, and Christian R. Huyck. 2004. Generating discourse structures for written text. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 329–335, Geneva.
Wang, Lu and Claire Cardie. 2013. Domain-independent abstract generation for focused meeting summarization. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics*, pages 1395–1405, Sofia.
Wang, Xuerui and Andrew McCallum. 2006. Topics over time: A non-Markov continuous-time model of topical trends. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 424–433, Philadelphia, PA.
Weng, Jianshu and Bu-Sung Lee. 2011. Event detection in Twitter. In *Proceedings of the Fifth International Conference on Weblogs and Social Media*, pages 401–408, Barcelona.
Weng, Jianshu, Ee-Peng Lim, Jing Jiang, and Qi He. 2010. Twitterrank: Finding topic-sensitive influential Twitterers. In *Proceedings of the 3rd International Conference on Web Search and Web Data Mining*, pages 261–270, New York, NY.
Wolf, Florian and Edward Gibson. 2005. Representing discourse coherence: A corpus-based study. *Computational Linguistics*, 32(2):249–287.
Xun, Guangxu, Yaliang Li, Wayne Xin Zhao, Jing Gao, and Aidong Zhang. 2017. A correlated topic model using word embeddings. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pages 4207–4213, Melbourne.
Yan, Xiaohui, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd International World Wide Web Conference*, pages 1445–1456, Rio de Janeiro.
Zeng, Jichuan, 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*, Brussels. arxiv: 1809.03664.
Zeng, Xingshan, 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*, pages 375–385, New Orleans, LA.
Zhang, Renxian, Wenjie Li, Dehong Gao, and Ouyang You. 2013. Automatic Twitter topic summarization with speech acts. *IEEE Transactions on Audio, Speech & Language Processing*, 21(3):649–658.
Zhao, Wayne Xin, Jing Jiang, Jianshu Weng, Jing He, Ee-Peng Lim, Hongfei Yan, and Xiaoming Li. 2011. Comparing Twitter and traditional media using topic models. In *Advances in the 33rd European Conference on Information Retrieval*, pages 338–349, Dublin.
Zhou, Xiaohua, Xiaodan Zhang, and Xiaohua Hu. 2007. Dragon toolkit: Incorporating auto-learned semantic knowledge into large-scale text retrieval and mining. In *Proceedings of the 19th IEEE International Conference on Tools with Artificial Intelligence*, pages 197–201, Patras.
Zubiaga, Arkaitz, Maria Liakata, and Rob Procter. 2016. Learning reporting dynamics during breaking news for rumour detection in social media. *ArXiv preprint, arXiv:1610.07363.*