Who Responded to Whom: The Joint Effects of Latent Topics and Discourse in Conversation Structure

Lu Ji[1*], Lei Chen[2*], Jing Li[3†], Zhongyu Wei[2,4], Qi Zhang[1], Xuanjing Huang[1]
[1] School of Computer Science, Fudan University, China
[2] School of Data Science, Fudan University, China
[3] Department of Computing, The Hong Kong Polytechnic University, Hong Kong, China
[4] Research Institute of Intelligent and Complex Systems, Fudan University, China
[1,2,4]{17210240034, chenl18, zywei, qi_zhang, xjhuang}@fudan.edu.cn
[3]jing-amelia.li@polyu.edu.hk;

Abstract

Vast amount of online conversations are produced on a daily basis, resulting in a pressing need to automatic conversation understanding. As a basis to structure a discussion, we identify the responding relations in the conversation discourse, which link response utterances to their initiations. To figure out who responded to whom, here we explore how the consistency of topic contents and dependency of discourse roles indicate such interactions, whereas most prior work ignore the effects of latent factors underlying word occurrences. We propose a neural model to learn latent topics and discourse in word distributions, and predict pairwise initiation-response links via exploiting topic consistency and discourse dependency. Experimental results on both English and Chinese conversations show that our model significantly outperforms the previous state of the arts.

1 Introduction

The growing popularity of online platforms have resulted in the revolution of interpersonal communications. Individuals now engage in diverse forms of online conversations to exchange viewpoints and share ideas. It allows users to access an abundance of fresh materials, whereas the explosive growth of online texts — essentially conversational and usually in multiple threads (Wang and Rosé, 2010) — has also hindered human capability to find the information needed. There consequently presents a pressing need to develop conversation understanding methods to automatically digest massive texts and complex interactions therein. To that end, it is crucial to capture the interactions of who responded to whom — the base to build and understand the conversation structure, as pointed out in many previous studies (Wang and Rosé, 2010; Zeng et al., 2019b). By reflecting how participants interact with each other, such structure has shown useful to predict users' online social activities (Zeng et al., 2019b), summarize key discussion topics (Qin et al., 2017; Li et al., 2018a), measure argument persuasiveness (Ji et al., 2018a), and so forth.

[C_1] I am aware that you can **thank** them in **private argument** but what does *that* matter?
[C_2] The most important part of my argument is that it **hurts** literally nobody.
[C_3] All they are doing is trying to be **polite**.
[C_4] Some people gild comments anonymously and do not respond to the **private messages**, so the gildee never knows who gave them gold.
[C_5] Note: for the purposes of my argument, assume I am talking about comments edited in such a way as to say **thanks** for the gold!

[R] We are all aware that you can do *that*, but sometimes people like to **express gratitude publicly**.

Figure 1: A Reddit conversation snippet. C_1 and R is an initiation-response pair while C_2 to C_5 are the other four candidates. **Topic words** reflecting the discussion points "public gratitude expression" are in bold. The blue and italic "*that*" occurring in both C_1 and R imply R's possible intention to answer C_1's question.

To date, despite of the extensive efforts on user interaction modeling, many of them employ user-annotated in-reply-to signals, such as *@-mention* on Twitter (Li et al., 2018a; Zeng et al., 2019b). Nonetheless, such labels are usually unavailable or unreliable (Du et al., 2017; He et al., 2019), especially for online conversations in informal styles. Other studies assume utterances only respond to their chronological neighbors (Jiao et al., 2018; Zhao et al., 2018), largely ignoring the long-distance interactions prominent in online conversations (Wang and Rosé, 2010). All these concerns lay down our objective to investigate who responded to whom in conversation contexts.

Following previous practice (Schegloff, 2007), we define our task to predict pairwise initiation utterances and their responses in an online con-

[*] Lu Ji and Lei Chen have equal contribution.
[†] Corresponding author.

Proceedings of the 10th SIGHAN Workshop on Chinese Language Processing (SIGHAN-10), pages 58–68
August 16, 2024 ©2024 Association for Computational Linguistics

versation (henceforth **initiation-response pairs**), where an initiation sets up an expectation earlier and its response later react to it in process of a discussion. To illustrate our task, Figure 1 shows an example response R and the other five utterances C_1 to C_5 from R's previous post in a Reddit conversation. Our goal is to identify which utterance from C_1 to C_5 is R's initiation. As can be seen, R is most likely to respond to C_1 for two possible reasons: First, they both focus on the topic of *public gratitude expression* (as topic words "*thank*", "*public*", "*gratitude*" are mentioned); Second, C_1 raises a question (signaled by "*what*" and the question mark "*?*") that can be well answered by R (via echoing the pronoun "*that*").

Here, we examine two latent factors that implicitly link an initiation and its response — the consistency of the topics they center around (henceforth **topic consistency**) and the dependency of their discourse roles (henceforth **discourse dependency**). Our intuition is that responses tend to follow the points pushed forward in their initiations (such as *public gratitude expression* in Figure 1) and their discourse roles are likely to exhibit some dependency in interactions, such as an answer responding to an initiated question (like R answering C_1 in Figure 1) and an argument followed by another argument in a back-and-forth debate. To the best of our knowledge, *we are the first to analyze the effects of topics and discourse in conversational responding behavior*, while previous work predict initiation-response pairs without modeling such latent factors embedded in the relations (Du et al., 2017; He et al., 2019).

To learn topics and discourse, we separate two word distributions for representing each of them. The latent variables are inferred with a neural architecture in an unsupervised manner (Zeng et al., 2019a), which enables topic and discourse inference without either manually annotated data (Zhao et al., 2017) or expertise involvements to customize model inference (Li et al., 2018b). Afterwards, two neural modules are employed, one to capture topic consistency and the other discourse dependency, both aim to explore the implicit links of a response and a candidate initiation. The learned representations are hence coupled to predict how likely the two utterances form an initiation-response pair.

In an empirical study, we carry out extensive experiments on two conversation datasets, one contains English argumentative discussions on Reddit (from the ChangeMyView subreddit), and the other Chinese customer service dialogues from e-commerce platform *Wangwang*. Both of them will be released upon publication as part of our work. The experimental results show that our model significantly outperforms state-of-the-art methods from previous work. For example, we achieve 79.02 MRR on the Wangwang dialogues compared with 72.69 produced by He et al. (2019). In extensive analyses on latent topics and discourse, we find that meaningful representations can be learned by our model and both topics and discourse may contribute to indicate initiation-response pairs. Lastly, we show that our learned representation to indicate initiation-response relations can further benefit to identify persuasive arguments in social media debates.

2 Study Design

2.1 Task Formulation

We define initiation-response pairs following Schegloff (2007) and refer both initiations and responses to conversation utterances from different participants. In a discussion flow, responses appear and react to the points raised earlier in their initiations and hence hold responding relations with them.

In previous practice, an initiation-response pair is defined to cover a wide range of user interactions, such as questions and answers, quotations and replies, blames and denials, all existing in diverse genres of conversations (Wang and Rosé, 2010). In empirical study, we will experiment on quotation-reply pairs in forum discussions (Wang and Rosé, 2010) and question-answer pairs in customer service dialogues (He et al., 2019). We thus describe these two types of initiation-response relations in the following.

Quotations and Replies. Many popular online forums, such as Reddit and Usenet, allow users to quote utterances from previous messages to indicate what they are commenting on. Such quoting behaviors provide us with abundant user-annotated data to extensively study initiation-response relations in forum conversations.

Here we are interested in a specific type of online conversations — argumentative dialogues from the *ChangeMyView* subreddit (henceforth *CMV*) (Tan et al., 2016), exhibiting rich user interactions in back-and-forth social media debates. In *CMV*, an opinion holder (*OH*) first initiates a debate with their viewpoints and challengers then engage in, raising their arguments in comments and attempting to change *OH*'s mind. As challengers carry on the persuasion process, they usually quote *OH*'s utterances to explicitly point out what they are arguing against, followed by their own responsive arguments (replies). An example quotation in a *CMV* comment is shown in Figure 2, where the reply utterance questions the *positive aspects of early Americans* — a point initialized by *OH*.

Original Post from an Opinion Holder:

... Strong family values in society lead to great results. *I want society to take positive aspects of the early Americans and implement that into society.* This would be a huge improvement than what we have now. ...

Comment from a Challenger:

&gt; I want society to take positive aspects of the early Americans and implement that into society. What do you believe those aspects to be? ...

Figure 2: An original post and its comment from the *CMV* subreddit. The comment quotes an utterance from the original post (in italic), followed by its reply utterance.

Questions and Answers. We also examine questions and answers in customer service dialogues on Chinese e-commerce platform *Wangwang* (henceforth *CS*) (He et al., 2019). In a dialogue thread, customers may raise multiple questions in a sequence of utterances and the seller's answers may appear in the following turns. Our goal is to pair a question from the customer's utterances and an answer from the seller's. Figure 3 shows a customer service dialogue excerpt centered around *a dress in winter style*. We observe two question-answer pairs therein focusing on the *product quality* and *dress style*, respectively.

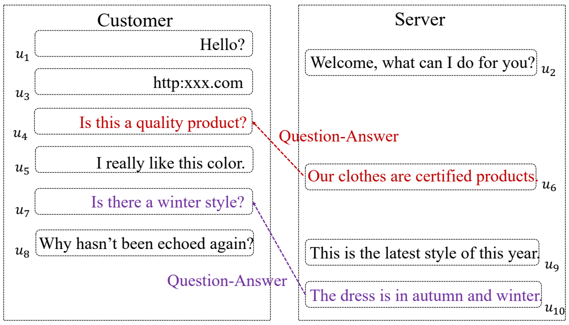

Customer
Server
u_1
Hello?
Welcome, what can I do for you?
u_2
u_3
http:xxx.com
u_4
Is this a quality product?
Question-Answer
u_5
I really like this color.
Our clothes are certified products.
u_6
u_7
Is there a winter style?
u_8
Why hasn't been echoed again?
This is the latest style of this year.
u_9
Question-Answer
The dress is in autumn and winter.
u_{10}

Figure 3: A Wangwang dialogue between a customer (on the left) and a seller (from the server team on the right) from He et al. (2019). Pairwise questions and answers are linked and displayed with the same color.

Pairwise Ranking. To explore how responses and initiations interact with each other, here we follow previous settings to formulate our task into a pairwise ranking problem (Wang and Rosé, 2010). It is shown that the determination of who responds to whom largely relies on subjective judgements (without explicit indicators); thus we view the pairing of responses to their initiations from a comparison perspective (instead of answering "yes or no" in binary classification fashion).

Specifically, given a response utterance r, we rank a set of candidate utterances with one positive initiation q^+ and u negative ones $q_1^- \sim q_u^-$. In practice, we measure a matching score $S(q, r)$ to indicate the likelihood of q as r's initiation and the one with the highest score will be considered as r's predicted initiation. In Section 2.2, we will describe how we form the candidate initiations.

2.2 Data Collection and Analysis

Data Collection. The *CMV* dataset gathers social media arguments, whose raw data is released by Tan et al. (2016). For each discussion, we only examine the context of an *OH*'s post and a challenger's comment to focus on the quotation-reply relations therein. In challenger's comments, we form a quotation and the utterance right after it to be an initiation-response pair. The rest utterances in the quoted post (from *OH*) are used as the negative instances, and the samples are randomly selected with a cap at 4 to avoid unbalanced labels. In addition, the quotation of the *OH*'s post is removed from the challenger's comment when forming an instance.

The *CS* dataset is annotated and released with He et al. (2019). The newest 4 consecutive customer's utterances (skipping the positive initiation) before a seller's response serve as the negative instances. Here the candidate number is also capped at 4 for comparable results with *CMV*.

	CMV Dataset	CS Dataset
# of utt. per conv	21.2±15.6	9.6±2.8
# of words per conv	403.1±292.5	130.8±73.1
# of convs	7,937	4,277
# of words per r	19.7±6.0	15.0±20.8
# of words per q^+	20.6±6.2	6.5±4.3
# of words per q^-	16.5±5.0	11.2±18.7
max # of pairs	14	7
avg. # of pairs	1.1±0.3	1.7±1.1

Table 1: Data statistics. Means and standard deviations appear before and after ±. utt. and r refers to utterance and response, while q^+ and q^- for positive and negative initiation. # of pairs represents the number of initiation-response pairs per conversation.

Data Analysis. Table 1 shows the data statistics, where the two datasets exhibit different characteristics. *CMV* arguments contain more utterances and richer contexts (with more words) compared with *CS*. For initiation-response pairs, *CMV* challengers only quote once on average while the maximum number is 14 (to extensively criticize *OH*'s weak points); whereas the number of question-answer pairs are diverse in *CS* dataset, ranging from 1 to 7 with 1.1 standard deviation.

3 Learning Topics and Discourse Effects for Initiation-Response Prediction

The overall architecture of our model is shown in Figure 4 (a). It takes an initiation candidate q, a response r, and their corresponding contexts c_q and c_r as inputs. The outputs are matching scores indicating how likely r responds to q.

3.1 Latent Topics and Discourse Modeling

Inspired by previous efforts in neural topic models (Miao et al., 2017; Zeng et al., 2019a), we adopt variational autoencoder (VAE) (Kingma and Welling, 2013; Rezende et al., 2014) to learn latent topics and discourse. It allows their associated word distributions to be learned in neural architecture and end-to-end training with other components in a deep learning framework. The corresponding networks are illustrated in Figure 4 (b). In below, we first describe how we model the topics, followed by the process to learn discourse.

Latent Topics. We first assume there are K latent topics in the corpus, each represented by a word distribution $\Phi_k^T (k = 1, 2, ..., K)$ over the vocabulary V. The latent topics of each utterance is defined as z and generated from the topic composition of its context c. Here we learn utterance-level topics in its conversation context assuming that utterances in a discussion excerpt tend to focus on similar topics. It allows the modeling of rich patterns of word statistics for topic inference.

The following process presents how to generate an utterance x in context of c. Here, we adopt the bag-of-words assumption of most latent topic models (Blei et al., 2002; Miao et al., 2017) and generate x in its bag of words (BoW) form x^{BoW}.

- Draw the latent topic $z \sim N(\mu, \sigma^2)$
- c's topic mixture $\theta = softmax(f_\theta(z))$
- For the n-th word in x:
  - $\beta_n = softmax(f_{\Phi^T}(\theta))$
  - Draw the word $w_n \sim Multi(\beta_n)$

where $f_*(\cdot)$ is a neural perceptron (fully connected layer). The weight matrix of $f_{\Phi^T}(\cdot)$ (after the softmax normalization) is viewed as the topic-word distributions Φ^T.

The prior parameters μ and σ are estimated from conversation c's bag of words c^{BoW}:

$$\mu = f_\mu(f_e(c^{BoW})), \log \sigma = f_\sigma(f_e(c^{BoW})) \qquad (1)$$

f_μ, f_e and f_σ are neural perceptron defined above.

As can be seen, the entire topic modeling process follows a VAE fashion — for each utterance x, we first encode its latent topic z from the conversation context c (in BoW form c^{BoW}) and then reconstruct its BoW (x^{BoW}) via decoding.

Latent Discourse. Similar to latent topics, we represent latent discourse with word distributions Φ_d^D $(d = 1, 2, ..., D)$ and D denotes the number of discourse roles observed from the corpus.

Following Ritter et al. (2010), we assume each utterance x reflects only one discourse role d (to signal its dialogue act). It is hence represented by a D-dimensional one-hot vector over the discourse inventory (the high bit indicates x's discourse role). To learn latent discourse, we adopt the similar VAE-based process as topic modeling with both the input and output as utterance x's BoW (x^{BoW}). First, x^{BoW} is encoded into its latent discourse role d with the following formula:

$$\pi = gs(f_\pi(x^{BoW})), d = Multi(\pi) \qquad (2)$$

where gs refers to Gumbel softmax function (Lu et al., 2017) to encode the discrete nature of latent discourse d and f_π is another neural perceptron. Afterwards, the decoding process reconstructs x^{BoW} conditioned on d with another fully connected layer:

$$x^{BoW} = f_{\Phi^D}(d) \qquad (3)$$

Here similar to latent topics, we utilize f_{Φ^D}'s weights to compute discourse-word distributions.

3.2 Initiation-Response Pair Prediction

Given topic and discourse representations of a response r (z_r and d_r) and those of its candidate initiation q (z_q and d_q), we further predict how likely they form an initiation-response pair with an utterance matching process. Here we measure the effects of topic consistency and discourse dependency to indicate initiation-response relations.

For topic consistency, we capture how similar the topics of q and r is with the following score:

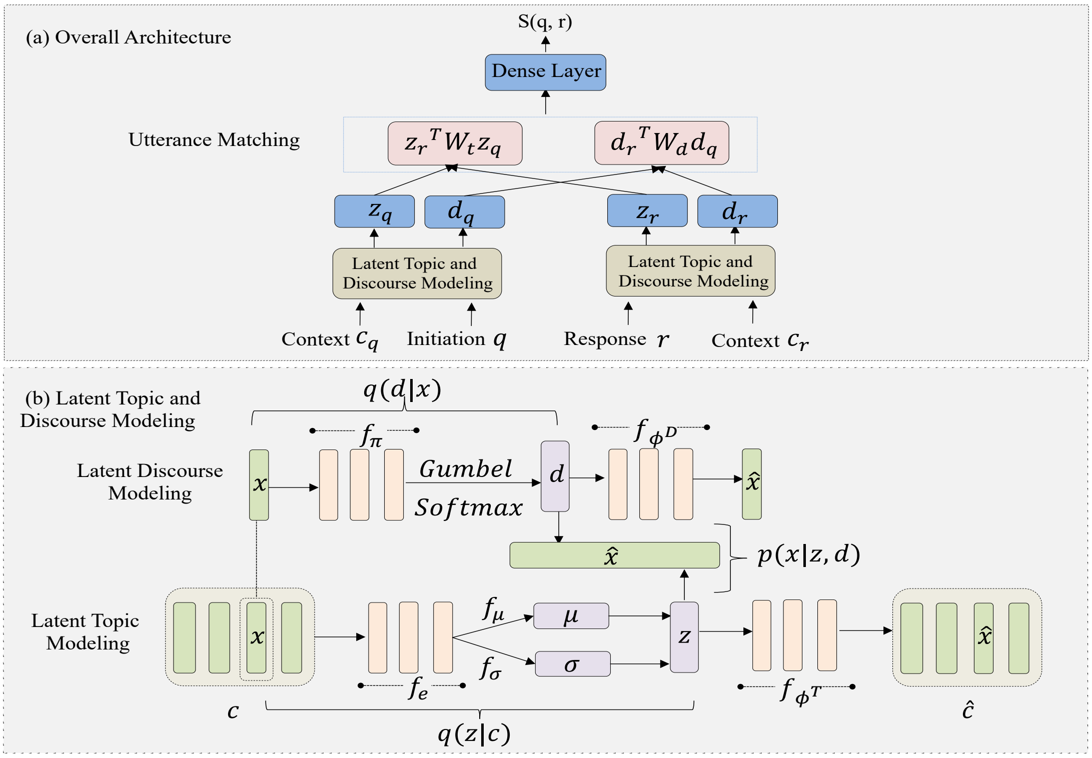


Figure 4: (a) Our model architecture to predict initiation-response pairs. We first learn latent topics and discourse factors for both response r and the candidate initiation q in award of their contexts c_r and c_q and show the detailed learning process in (b) (x denotes q or r.) Then, utterance matching is conducted to measure topic consistency and discourse dependency. Lastly, we predict $S(q,r)$ — the likelihood of r responding to q.

$$S_{topic}(q,r) = z_r^T W_t z_q \tag{4}$$

where W_t is a weight matrix learned to indicate the importance of each topic factor.

Likewise, q and r's discourse-level matching score is denoted as $S_{discourse}$ and defined below:

$$S_{discourse}(q,r) = d_r^T W_d d_q \tag{5}$$

where the trainable weight matrix W_d is employed to capture the transition probabilities from q's discourse role to r's ($Pr(d_r \,|\, d_q)$).

Further, to yield the final matching score $S(q,r)$ to estimate how likely r responding to q, we leverage $S_{topic}(q,r)$ and $S_{discourse}(q,r)$ to couple both topic and discourse effects with the weighted sum:

$$S(q,r) = \gamma S_{topic}(q,r) + (1-\gamma) S_{discourse}(q,r) \tag{6}$$

where $\gamma \in [0,1]$ is the parameter balancing the relative contributions of topic and discourse.

3.3 Learning Objectives

Latent Topics and Discourse Modeling Loss. We employ neural variational inference to approximate the posterior distributions over the latent topic z and the latent discourse d.

Encoding Topics and Discourse. To examine how to learn topics and discourse, the cross entropy loss is used to reflect the estimation of z and d from encoding process:

$$L_t = E_{q(z\,|\,c)}[\log p(c\,|\,z)] - KL(q(z\,|\,c)\,||\,p(z)) \tag{7}$$

$$L_d = E_{q(d\,|\,x)}[\log p(x\,|\,d)] - KL(q(d\,|\,x)\,||\,p(d)) \tag{8}$$

KL cost term is added to avoid posterior collapse. For space limitation, we leave out the derivation details and refer the readers to Zhao et al. (2018).

Reconstructing Utterances. For the reconstruction loss to reflect how an utterance can be inferred from z and d, we define the loss L_x as:

$$L_x = E_{q(z\,|\,x)q(d\,|\,c)}[\log p(x\,|\,z,d)] \tag{9}$$

Distinguishing Topics and Discourse. As discussed above, topics and discourse are modeled in different granularity (discourse in utterance only while topics in richer contexts). To further distinguish their respective word distributions, we follow Zeng et al. (2019a) to employ the mutual information to define the mutual dependency of latent topics and discourse:

$$E_{q(z)q(d)}[\log \frac{p(z,d)}{p(z)p(d)}] \tag{10}$$

Further, the mutual information loss (shown in below) is adopted to separate the semantic space of topics and discourse:

$$L_{MI} = E_{q(z)q(d)}[KL(p(d\,|\,z)\,||\,p(d))] \tag{11}$$

Initiation-Response Pair Prediction Loss. To allow positive pairs to obtain higher matching scores than negative, we use hinge loss in training:

$$L_m = \sum_{i=1}^{u} max(0, \lambda - S(q^+, r) + S(q_i^-, r)) \tag{12}$$

where u is the number of negative initiations for each response. λ is a margin parameter and $S(q^+, r)$ and $S(q_i^-, r)$ are the matching scores of a response and its positive and negative initiations.

The Final Objective. Finally, we combine all the effects above and define the overall objective of the entire model as:

$$L = L_t + L_d + L_x + L_m - L_{MI} \tag{13}$$

In the training process, the optimization of final objective L enables the end-to-end exploration of topic and discourse representation and their joint effects to signal pairwise initiation-response relations in conversation structure.

4 Experimental Setup

Data Preprocessing. For *CMV* dataset, the raw data was preprocessed by Tan et al. (2016). We first filter out tokens occurring less than 15 to alleviate sparsity and maintain a vocabulary with $15,182$ tokens. Then, we remove too short (with less than 7 words) and too long utterances (with over 45 words) to better explore utterance-level word statistics for topic and discourse modeling. Next, to form context for quotations and replies (c_q and c_r), we consider all utterances in the original post (from *OH*) as c_q and those in the challenger's comment as c_r. Lastly, the training and test data is separated following Tan et al. (2016), where $6,839$ pairs are used for training and and $1,098$ for test.

For *CS* dataset, we don't remove words and the vocabulary size is $15,407$, with the scale similar to *CMV*. Short utterances with less than 5 words are removed. The Chinese word segmentation and the separation of training and test set has been done by He et al. (2019), with $3,701$ and 576 instances for training and test. Here all utterances in the dialogue thread are used to form both c_q and c_r due to the synchronous nature of *CS* conversations.

For both datasets, 10% data is further sampled from the training set for validation.

Model Settings. The hyperparameters are tuned on validation set. For the number of topics (K) and discourse roles (D), we set $K = 50, D = 5$ for *CMV* dataset and $K = 10, D = 3$ for *CS*. Max margin weight λ is set to 10 (Eq. 12) and $\gamma = 0.5$ for balancing topic consistency and discourse dependency (Eq. 6). In model training, we set the batch size to 32, dropout probability to 0.5, and the maximum epoch number to 200 (with early stop). The trainable parameters are optimized via stochastic gradient descent with learning rate decay, whose initial learning rate is set to 0.1.

Evaluation Metrics. In evaluation, we examine whether the positive initiations can be ranked higher than negative for each response. Two widely-used information retrieval metrics *Hits@N* and Mean Reciprocal Rank (*MRR*) are adopted. For *Hits@N* we only measure the hits at the top two retrieved initiations, i.e., $N = 1, 2$.

Comparison Models. We first consider three non-neural baselines that rank initiations based on: 1) POSITION, where earlier utterances are ranked higher for *CMV* while later is higher for *CS*; 2) EMBEDDING_SIM — the cosine similarity between a response and an initiation utterance measured by the average word embeddings from Glove (Pennington et al., 2014); 3) LDA_DISC — using cross entropy to discriminate initiation's and response's topic distributions inferred by latent Dirichlet allocation (LDA) (Du et al., 2017).

We also compare with the following neural models proposed by previous work: 1) MALSTM (Mueller and Thyagarajan, 2016) designed for sentence-level semantic matching (LSTM for utterance encoding and Manhattan distance for matching); 2) COATTENTION (Ji et al., 2018b) proposed for pairwise argument quality evaluation, where a co-attention network learns alignment representations and a BiGRU layer computes similarity for matching. 3) RPN (He et al., 2019), the state-of-the-art model for question-answer pairing in dialogues that ranks initiations by recurrent pointer networks (RPN).

In addition, we consider the following neural matching models with a fully connected layer to score initiation-response pairs and the following encoders for utterance-level representation learning: RNN (henceforth Match_RNN), autoencoder (henceforth Match_AE), variational autoencoder (henceforth Match_VAE), and discrete variational autoencoder (Zhao et al., 2018) (henceforth Match_DVAE).

Further, to study the relative contributions of topic consistency and discourse dependency, we compare with our two ablations, one only explores the topic effects (henceforth Topic_Only) and the other discourse (henceforth Discourse_Only).

5 Results and Discussions

5.1 Main Comparison Results

The overall results are shown in Table 2. Several interesting observations can be drawn.

• *All models yield generally better performance on CS than CMV.* It shows that initiation-response links are more difficult to be identified on dialogues in argumentative than everyday styles.

• *Neural networks perform better than non-neural baselines.* Initiation-response pair prediction is challenging, where shallow features from position, word embeddings, and LDA-based latent topics cannot guarantee good performance. Neural models explore deeper semantic features and hence provide better results.

• *Autoencoders can learn useful representations.* It is observed that models based on autoencoders perform generally better than other neural models. This shows that autoencoders are effective in encoding utterances compared with other alternatives, such as RNN.

• *Topics contribute more on CMV while discourse is more useful in CS.* Topic_Only performs much better than Discourse_Only on *CMV*, while the opposite is observed on *CS*. It is probably because of the richer context in *CMV* to learn latent topics (with more words per conversation as shown in Table 1), while the synchronous *CS* dialogues exhibits richer discourse word patterns from back and forth interactions between participants and hence allow better discourse modeling.

• *Our model significantly outperforms all comparisons.* This shows that the joint effects of topics and discourse can usefully indicate the relations of initiations and responses in conversation context.

5.2 Effects of Topics and Discourse

We have shown the joint effects of topics and discourse to signal initiation-response relations. Here we further analyze what we learn for topic and discourse representations.

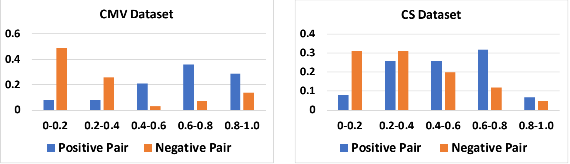


Figure 5: The distribution of topic similarity in the CMV dataset (a) and CS (b). X-axis shows similarity intervals and y-axis indicates proportions. For each interval, positive pair results are displayed on the left (in blue) and negative on the right (in orange).

Topic Effects. We first analyze the effects of topic consistency and compute the cosine similarity of the latent topics we learn for responses (z_r) and candidate initiations (z_q). The distributions over positive and negative pairs are shown in Figure 5. For both datasets, our model generally assigns higher topic similarity for positive pairs than negative, probably because responses tend to follow the concern of initiations and are hence likely to contain similar topic words. We also observe a proportion drop in very similar positive pairs ($sim > 0.8$), indicating that most responses do not echo what were said in initiations, though their topics might be similar. Nevertheless, negative pairs exhibit different distributions compared with the positive ones. Our model is able to capture such features in topic consistency modeling (Eq. 4), which might help in distinguishing positive and negative initiations for a response.

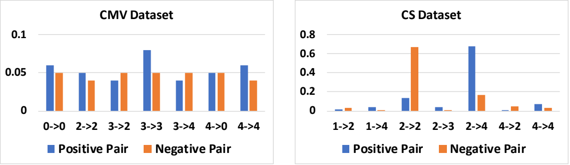


Figure 6: The transition distributions of discourse roles from initiations to responses, *CMV* in (a) and *CS* in (b). Only the top 5 transitions observed in positive (on the left in blue) and negative pairs (on the right in orange) are displayed. X-axis: initiation-response discourse roles ($d_q \rightarrow d_r$); Y-axis: proportions.

Discourse Effects. We then discuss how discourse dependency affects the prediction of initiation-response pairs. The transition distributions of discourse roles from initiations to responses ($d_q \rightarrow d_r$) are shown in Figure 6. As can

2*Models	CMV Dataset			CS Dataset		
	Hits@1	Hits@2	MRR	Hits@1	Hits@2	MRR
Non-Neural Models						
Position	24.68*	24.68*	24.68*	49.13*	49.13*	49.13*
Embedding_Sim	22.77*	45.00*	48.66*	17.01*	39.06*	44.04*
LDA_Disc	24.68*	42.99*	47.77*	26.39*	49.65*	52.40*
Neural Models						
MaLSTM (Mueller and Thyagarajan, 2016)	29.87*	42.99*	50.91*	43.58*	72.40*	65.80*
CoAttention (Ji et al., 2018b)	47.72*	68.31*	67.26*	51.56*	79.17*	71.77*
RPN (He et al., 2019)	46.45*	67.21*	66.22*	52.95*	80.21*	72.69*
Match_RNN	49.45*	71.58*	68.79*	50.00*	80.38*	71.13*
Match_AE	51.82*	74.77‡	70.50*	52.78*	82.12‡	72.88*
Match_VAE	53.19*	73.95‡	71.11‡	52.60*	81.42*	72.70*
Match_DVAE	47.45*	69.95*	67.34*	53.82*	82.81‡	73.65*
Ablations						
Topic_Only	58.20	76.14	73.78	42.53*	69.10*	64.11*
Discourse_Only	41.44*	63.02*	62.20*	48.96*	76.04*	69.76*
Our model	**59.74**	**76.23**	**74.41**	**64.93**	**84.20**	**79.23**

Table 2: Comparison results on two datasets and our model achieves the best results under all settings. * and ‡ indicates that our model significantly outperforms the comparison model (* for $p<0.01$ and ‡ for $p<0.05$, both measured with Wilcoxon signed rank test).

be seen, the discourse transition distributions in *CS* dataset are diverse for positive and negative pairs. It may help explain why discourse can better signal initiation-response pairs on *CS* compared with *CMV* (observed from Discourse_Only's performance in Table 2). For *CMV*, there are slightly different distributions for positive and negative pairs. For this reason, topic factors may contribute more than discourse (seen via comparing Discourse_Only and Topic_Only on *CMV*). This also indicates that discourse modeling for argumentative dialogues is challenging, which may require the learning of more complex features other than word statistics and is beyond the capacity of our model.

6 Related Work

Our work is in the line with prior efforts to detect initiation-response pairs. Wang and Rosé (2010) explore how topic features discovered via latent semantic analysis (LSA) work in this task, largely ignoring the effects of discourse roles. On the contrary, our study shows that both topics and discourse are helpful to identify who respond to whom in conversation structure. Other related work (Jamison and Gurevych, 2014; Du et al., 2017; Chen et al., 2017) focus on the design of hand-crafted features. Recently, there exists a growing attention over how neural framework perform to identify replying relations in conversation discourse (Guo et al., 2018; He et al., 2019). However, they ignore the effects of latent topics and discourse to structure a conversation, which are extensively studied here and shown useful to indicate initiation-response relations in experiments.

We are also inspired by the previous approaches to discover latent topics and discourse in conversations contexts. Many of them employ probabilistic graphical models in LDA-fashion to explore word statistics (Ritter et al., 2010; Li et al., 2018a; Zeng et al., 2018). We take the advantage of the recent progress to explore conversation representations via variational autoencoders (VAE) (Miao et al., 2017; Zhao et al., 2018; Zeng et al., 2019a), allowing to capture topic and discourse factors in an unsupervised manner. However, their effects to signal user interactions in conversation structure have never been studied before, which is a gap our work fills in.

7 Conclusion

This work explores the effects of latent topics and discourse roles to signal initiation-response relations that structure a conversation. We first employ a VAE-based neural model to capture topic and discourse representations in an unsupervised manner. Then, topic consistency and discourse dependency are further exploited to predict how likely an utterance responds to an initiation. Extensive experiments on large-scale datasets containing asynchronous English argumentative conversations (from the *CMV* subreddit) and synchronous Chinese customer service dialogues (from *Wangwang* platform) show that our model significantly outperform the previous state-of-the-art models.

Limitations

The model's performance relies on preprocessing steps, such as token filtering and utterance length restrictions, which could potentially introduce bias or eliminate valuable information. To address this issue, the use of modern tokenizers and large language models may be beneficial. Additionaly, in terms of multi-lingual generalizability, the model's ability to identify initiation-response pairs in asynchronous English argumentative conversations and synchronous Chinese customer service dialogues may not readily transfer to other languages.

Acknowledgments

This work is supported by National Natural Science Foundation of China (No. 62176058) and National Key RD Program of China (2023YFF1204800). The project's computational resources are supported by CFFF platform of Fudan University.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. 2002. Latent dirichlet allocation. In *Advances in neural information processing systems*, pages 601–608.

Jun Chen, Chaokun Wang, Heran Lin, Weiping Wang, Zhipeng Cai, and Jianmin Wang. 2017. Learning the structures of online asynchronous conversations. In *International Conference on Database Systems for Advanced Applications*, pages 19–34. Springer.

Wenchao Du, Pascal Poupart, and Wei Xu. 2017. Discovering conversational dependencies between messages in dialogs. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Gaoyang Guo, Chaokun Wang, Jun Chen, and Pengcheng Ge. 2018. Who is answering to whom? finding "reply-to" relations in group chats with long short-term memory networks. In *Proceedings of the 7th International Conference on Emerging Databases*, pages 161–171. Springer.

Shizhu He, Kang Liu, and Weiting An. 2019. Learning to align question and answer utterances in customer service conversation with recurrent pointer networks. In *Thirty-Third AAAI Conference on Artificial Intelligence*.

Emily Jamison and Iryna Gurevych. 2014. Adjacency pair recognition in wikipedia discussions using lexical pairs. In *Proceedings of the 28th Pacific Asia Conference on Language, Information and Computing*.

Lu Ji, Zhongyu Wei, Xiangkun Hu, Yang Liu, Qi Zhang, and Xuanjing Huang. 2018a. Incorporating argument-level interactions for persuasion comments evaluation using co-attention model. In *Proceedings of the 27th International Conference on Computational Linguistics, COLING 2018, Santa Fe, New Mexico, USA, August 20-26, 2018*, pages 3703–3714.

Lu Ji, Zhongyu Wei, Xiangkun Hu, Yang Liu, Qi Zhang, and Xuanjing Huang. 2018b. Incorporating argument-level interactions for persuasion comments evaluation using co-attention model. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3703–3714.

Yunhao Jiao, Cheng Li, Fei Wu, and Qiaozhu Mei. 2018. Find the conversation killers: A predictive study of thread-ending posts. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web, WWW 2018, Lyon, France, April 23-27, 2018*, pages 1145–1154.

Diederik P Kingma and Max Welling. 2013. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

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

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

Jiasen Lu, Anitha Kannan, Jianwei Yang, Devi Parikh, and Dhruv Batra. 2017. Best of both worlds: Transferring knowledge from discriminative learning to a generative visual dialog model. In *Advances in Neural Information Processing Systems*, pages 314–324.

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

Jonas Mueller and Aditya Thyagarajan. 2016. Siamese recurrent architectures for learning sentence similarity. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Kechen Qin, Lu Wang, and Joseph Kim. 2017. Joint modeling of content and discourse relations in dialogues. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 974–984.

Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082*.

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

Emanuel A Schegloff. 2007. *Sequence organization in interaction: A primer in conversation analysis I*, volume 1. Cambridge University Press.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624. International World Wide Web Conferences Steering Committee.

Yi-Chia Wang and Carolyn P Rosé. 2010. Making conversational structure explicit: identification of initiation-response pairs within online discussions. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 673–676. Association for Computational Linguistics.

Jichuan Zeng, Jing Li, Yulan He, Cuiyun Gao, Michael R Lyu, and Irwin King. 2019a. What you say and how you say it: Joint modeling of topics and discourse in microblog conversations. *Transactions of the Association for Computational Linguistics*, 7:267–281.

Xingshan Zeng, Jing Li, Lu Wang, Nicholas Beauchamp, Sarah Shugars, and Kam-Fai Wong. 2018. 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, Volume 1 (Long Papers)*, pages 375–385.

Xingshan Zeng, Jing Li, Lu Wang, and Kam-Fai Wong. 2019b. Neural conversation recommendation with online interaction modeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 4632–4642.

Tiancheng Zhao, Kyusong Lee, and Maxine Eskenazi. 2018. Unsupervised discrete sentence representation learning for interpretable neural dialog generation. *arXiv preprint arXiv:1804.08069*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv preprint arXiv:1703.10960*.

A Position Distribution of Datasets

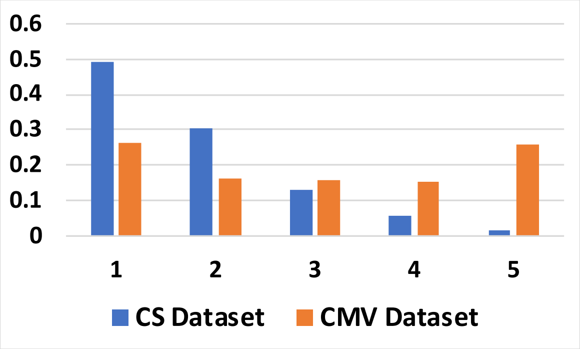


Figure 7: The distribution over relative positions of initiations and responses. X-axis: initiations' utterance order counted from responses (only considering customer's or *OH*'s turns). Y-axis: proportions.

We analyze the relative positions of initiations and responses and show the distribution of their intermediate utterance number in Figure 7. As can be seen, large proportion of responses do not interact with the closest utterance, though *CS* sellers do respond more to newer questions, probably because of recency effects in in synchronous dialogues — people's attention tends to be drawn by new information. However, in asynchronous forum discussions, *CMV* challengers are more likely to quote the opening points in *OH*'s post. Another possible reason is that most key arguments are located at the beginning of a post.

B Further Discussions

B.1 Parameter Analysis.

Here we present in-depth analyses of our model and start with the discussion of two important parameters — the number of topics (K) and discourse (D).

Varying Topic Number. Figure 8 (a) shows how Hits@1 scores change over varying number of topics (K). For comparison, we also display MATCH_DVAE's results, the best comparison model in Table 2. For relatively large K, our model performs consistently better than MATCH_DVAE. We also find that the our trend on both datasets are not monologues, where the best performance is attained at $K = 50$ for *CMV* and $K = 10$ for *CS*. This implies that the topics in customer service dialogues are limited (focusing on products) while participants may discuss wide range of topics in social media debates.

Varying Discourse Number. The results for varying discourse number (D) are displayed in Figure 8 (b). Similar to K, our model exhibits consistently better results than MATCH_DVAE for $D > 1$. It is also observed that *CS* is more sensitive to D compared with *CMV*, indicating that discourse factors largely affect the initiation-response prediction results on CS.

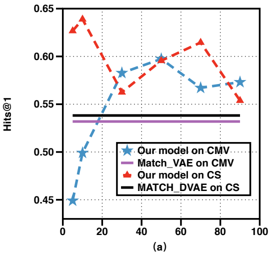


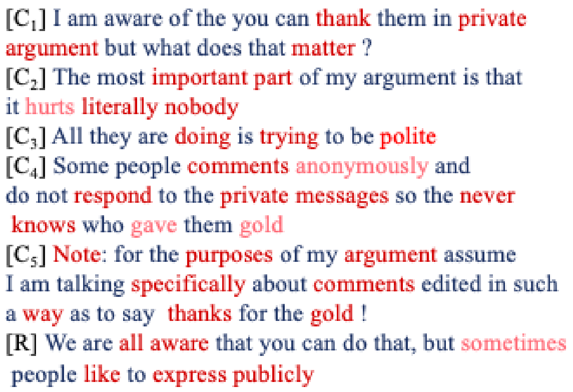
Figure 8: Hits@1 over varying number of topics and discourse X-axis: topic number (K in (a)) and discourse number (D in (b)). Y-axis: Hits@1 score. Blue and red curves: our model on *CMV* and *CS*. Purple and black lines: MATCH_DVAE on *CMV* and *CS*.

B.2 Case Study.

To further examine what we learn to represent topics and discourse, we take the *CMV* conversation snippet in Figure 1 as an example to analyze the topic and discourse words assigned by our model. Recall that R answers C_1's question suggested by the shared pronoun "that" and the similar topics they concern. Figure 9 shows the visualization results and displays topic words in red and discourse in blue. It is observed that our model is able to separate topic words (e.g., "thank", "private", and "public) from discourse (e.g., "that", "what", and "?"), which may resulting in coherent topic and discourse distributions and indicative representations to signal initiations-response relations. Interestingly, discourse words are mostly stop words and punctuation. Their meaningful clusters exhibiting different statistic patterns might usefully indicate varying discourse behaviors in conversations, which is consistent with the findings from previous studies (Li et al., 2018b; Zeng et al., 2019a).

B.3 Downstream Task.

In Introduction, we mentioned that the detection of initiation-response pairs may contribute to a better understanding of conversation structure and hence benefit downstream applications. Here we take the prediction of argument persuasiveness as an example to discuss whether the representations learned by our model can advance the state-of-the-art performance on this task. Table 3 shows the performance of the non-neural baseline (Tan et al., 2016), the state-of-the-art model (Ji et al., 2018b), and Ji et al. (2018b) incorporating the topic and discourse representations we learn (z and d). The dataset is also collected from *CMV* and argument quality is labeled by Δ (given by *OH* to indicate the successful persuasion). It is seen that the latent topics and discourse learned to signal initiation-response relations can indeed help to predict argument quality, suggesting that the persuasiveness of arguments are closely related to the structure of who respond to whom in argumentation processes.



[C$_1$] I am aware of the you can thank them in private argument but what does that matter ?
[C$_2$] The most important part of my argument is that it hurts literally nobody
[C$_3$] All they are doing is trying to be polite
[C$_4$] Some people comments anonymously and do not respond to the private messages so the never knows who gave them gold
[C$_5$] Note: for the purposes of my argument assume I am talking specifically about comments edited in such a way as to say thanks for the gold !
[R] We are all aware that you can do that, but sometimes people like to express publicly

Figure 9: Visualization of the topic and discourse word assignment for the *CMV* conversation snippet in Figure 1. The blue words are prone to indicate discourse ($p(w \mid d) > p(w \mid z)$) while red topic. Darker colors indicate higher confidence.

Models	Pairwise accuracy
Tan et al. (2016) (*baseline*)	65.70
Ji et al. (2018b) (*SOTA*)	70.45
Ji et al. (2018b)+Our model	74.12

Table 3: The pairwise accuracy to predict argument persuasiveness. The results in the first two rows were reported in their original paper. Our representations help advance the state of the art (SOTA).