Topic-diversified Neural Dialogue Generation

Hengyi Cai[*,†], Hongshen Chen[‡], Xiaofang Zhao[†], Dawei Yin[§]
Zhuoye Ding[‡], Yongjun Bao[‡], Weipeng Yan[‡]
[*]University of Chinese Academy of Sciences, Beijing, China
[†]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[‡]JD.com, China
[§]Baidu Inc., China
{caihengyi,zhaoxf}@ict.ac.cn,ac@chenhongshen.com
yindawei@acm.org,{dingzhuoye,baoyongjun,paul.yan}@jd.com

ABSTRACT

Conventional neural dialogue generation approaches, neglecting the topics of the context, are awkward to generate topic-specific responses when confronted with multi-topic conversations. To address the issues of dialogue generation in multi-topic scenarios, in this paper, we propose a **T**opic-diversified **N**eural **D**ialog generation framework—TND, to leverage the common ground and the difference across multiple topics of conversations. In particular, an encoder first transforms the input context into a hidden context representation vector; a topic indicator is applied to figure out the topical information of the given context; topic-wised responses are then generated through multiple topic-specific generators, where each generator dynamically injects the effects of the topical information into the response generation; and the most appropriate response is finally selected referring to the topical proximity between the context and the response. Extensive experiments on a large scale conversation dataset show that the proposed framework surpasses the state-of-the-art baselines on both the automatic evaluations and human evaluations, and the visualization further verifies the effectiveness of TND as well.

CCS CONCEPTS

• **Computing methodologies** → **Discourse, dialogue and pragmatics**.

ACM Reference Format:
Hengyi Cai[*,†], Hongshen Chen[‡], Xiaofang Zhao[†], Dawei Yin[§] and Zhuoye Ding[‡], Yongjun Bao[‡], Weipeng Yan[‡]. 2020. Topic-diversified Neural Dialogue Generation. In *Proceedings of AIIS'20.* ACM, New York, NY, USA, 8 pages. https://doi.org/TBA

1 INTRODUCTION

With the advent of large scale corpus and deep learning techniques, sequence-to-sequence (SEQ2SEQ) [1, 4, 27] based open-domain conversation systems prevail [23, 25] in recent years, and generate responses for a given context in an end-to-end way.

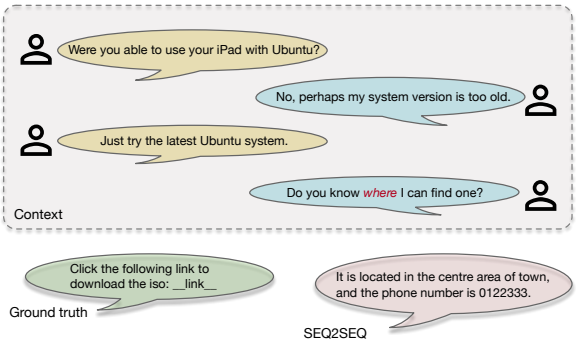

Figure 1: A representative case generated by SEQ2SEQ in multi-topic conversational scenarios.

Although it is referred as "open-domain", large sums of efforts in open-domain dialogue systems have been made for generating better responses within a single specific scenario actually, regardless of the complicated and diversified conversation topics. Moreover, when directly applying SEQ2SEQ to multi-topic conversational scenarios, the model sometimes misunderstands the topics of the context, hence generates generic or even irrelevant response. In Figure 1, users are discussing "Ubuntu". However, SEQ2SEQ mistakenly respond a real-word address instead of a hyperlink address, due to its inability to identify the current conversational topics from the dialogue context.

Inspired by the phenomenon that when human responds for a specific context, one may first identify the topics of the context, and then generate a proper response within this topic. In this paper, we propose a **T**opic-diversified **N**eural **D**ialog generation framework—TND, to enable the model to identify the topics of the input context and respond more appropriately by leveraging the common ground and the difference across multiple topics of conversations. Specifically, an encoder transforms the input context (previous utterances) into a hidden context representation vector. Then a topic indicator figures out the latent topics of the given context. One straightforward topic indicator classifies the input context into a specific scenario. Though effective, this method relies on the manually labeled topic annotations of the context and suffers from the label unbalance problem [9]. Therefore, we further propose a latent

AIIS'20, July 30, 2020, Virtual Event, China

ACM ISBN TBA…$TBA
https://doi.org/TBA

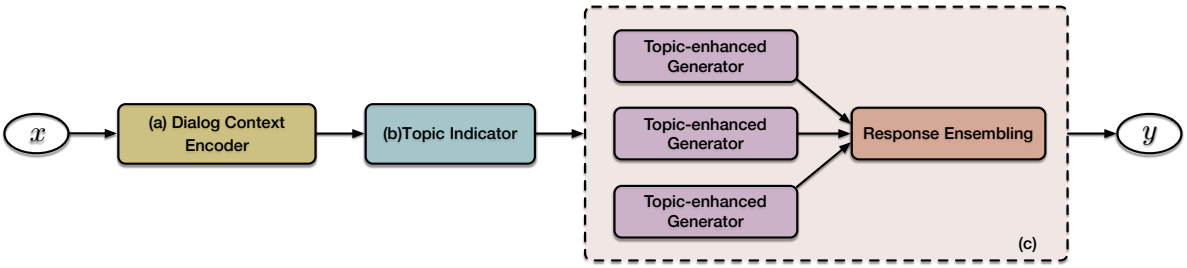

Figure 2: Model architecture.

topic indicator, which models a generative process of a dialogue during training, to distill the latent topics of the context. Finally, topic-wised responses are generated through multiple ensembled topic-specific generators. Each generator dynamically controls the effects of the topic information. The final response is then selected according to the topical proximity between the context and the response. Extensive experiments show that our model outperforms the state-of-the-art baselines on both the automatic evaluations and human evaluations.

2 TOPIC-DIVERSIFIED NEURAL DIALOGUE GENERATION

2.1 Overview

As illustrated in Figure 2, the topic-diversified neural dialogue generation framework handles the multi-topic conversations through the following procedures:

(a) Given a context $x = \{w_1, w_2, \cdots, w_{T_x}\}$, consisting of T_x words, an encoder (Figure 2 (a)) transforms the discrete tokens into context hidden representation $\mathbf{h}_x$. Typically, we utilize the bi-directional LSTM as the context encoder.
(b) Then, a topic indicator (Figure 2 (b)) identifies the topics of the context x according to $\mathbf{h}_x$.
(c) Finally, a response is generated by first producing topic-wised responses through multiple ensembled topic-enhanced generators, and then choosing the best one according to the topical proximity between the context and the response (Figure 2 (c)).

2.2 Topic Indicator

2.2.1 Topic Classification. One straightforward approach to identify the specific topic of an input context is topic classification. Here, we denote it as a *topic classification indicator*. In particular, we first apply a multi-layer perceptron (MLP) over the context representation $\mathbf{h}_x$, and then predict the topics through a softmax function. The topic classification indicator is computed as:

$$\xi_{clas} = \text{softmax}(\text{MLP}(\mathbf{h}_x)). \tag{1}$$

In response generation, we utilize the output probability distribution ξ_{clas} as a topic distribution vector.

Though topic classification is quite simple and straightforward, training the topic classifier heavily relies on the manually labeled topic annotations. Since the conversation scenarios and topics are numerous, annotating a large corpus is prohibitively expensive. What's more, the topic classification indicator suffers from the label unbalance problem [9]. In realistic corpus, the number of context-response pairs in each topic varies dramatically, which leads the topic classifier prone to the dominated topics.

2.2.2 Latent Topic Inference. We hence design a *latent topic indicator*, based on a neural topic inference network, to infer the latent topics without any explicit annotations.

Inspired by neural topic model [18], we infer the latent topics by modeling the generative process of a dialogue:

(a) As illustrated in Figure 3, the semantics of the given dialogue are modeled using a latent variable ν.
(b) Then, we construct the topic proportion θ from the latent variable ν with a softmax transformation.
(c) Finally, the dialogue d is reconstructed with θ through distribution $p(w_i|\beta_{z_i})$, where d represents a context-response pair (x, y), w_i is the i-th word in d, z_i is a topic assignment sampled from θ, and β_{z_i} is the topic-word distribution of assignment z_i.

Specifically, we sample the latent variable ν from the given context x by $P(\nu|x) = \mathcal{N}(\mu_{prior}, \sigma^2_{prior})$, where $\mathcal{N}(\mu_{prior}, \sigma^2_{prior})$ is the multivariate Gaussian distribution with mean μ_{prior} and covariance σ^2_{prior}. Following Kingma and Welling [13], we reparameterize the latent variable ν using a Gaussian noise ϵ by $\nu = \mu_{prior} + \epsilon \cdot \sigma_{prior}$. Given the bag-of-words representation of the context x as input, we compute μ_{prior} and σ^2_{prior} using multi-layer perceptrons. We perform neural variational inference [19], by employing an inference network $Q(\nu|d) = \mathcal{N}(\mu_{posterior}, \sigma^2_{posterior})$ to approximate the intractable true posterior $p(\nu|d)$, where $\mu_{posterior}$ and $\sigma^2_{posterior}$ are computed similarly as the prior.

We create the topic distribution θ with latent variable ν through a softmax transformation:

$$\theta = g(\nu) = \text{softmax}(\nu \cdot \mathbf{W}_\nu), \tag{2}$$

where $\mathbf{W}_\nu$ stands for the trainable parameter.

Finally, the dialogue d is reconstructed using the topic proportion θ. We compute the marginal likelihood of a dialogue d as:

$$p(d) = \int_\theta p(\theta) \prod_{i=1}^{|d|} \sum_{z_i} p(w_i|\beta_{z_i}) p(z_i|\theta) d\theta. \tag{3}$$

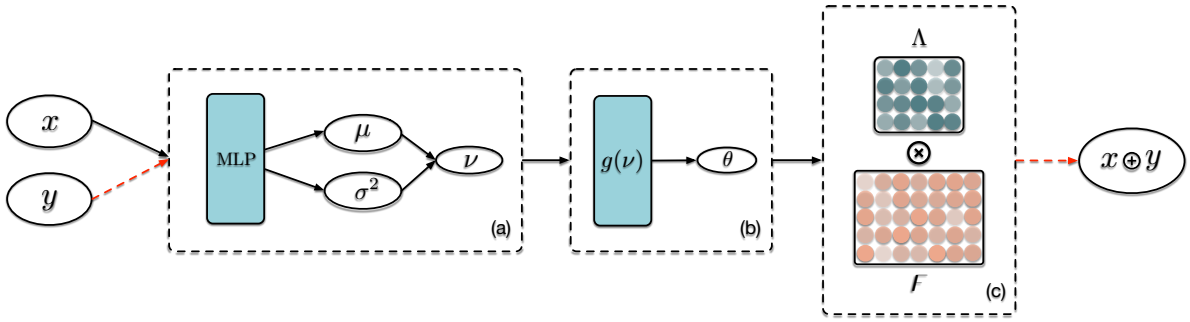

Figure 3: Latent topic indicator. The dashed red lines only appear in training.

We further integrate out the topic assignment z_i, and formulate the log-likelihood of a word w_i in dialogue d as:

$$\begin{aligned} \log p(w_i|\beta,\theta) &= \log \sum_{z_i} [p(w_i|\beta_{z_i}) p(z_i|\theta)] \\ &= \log(\theta \cdot \beta^{\mathrm{T}}) \end{aligned}, \tag{4}$$

where $\beta = \{\beta_1, \beta_2, \cdots, \beta_K\}$, and we formulate the topic-word distribution β_k by:

$$\beta_k = \mathrm{softmax}(F \cdot \Lambda_k^{\mathrm{T}}),$$

where $F \in \mathbb{R}^{M \times H}$ is the topical word embedding matrix with M topical words and embedding size H. $\Lambda \in \mathbb{R}^{K \times H}$ is the topic embedding matrix for K topics.

Unlike the conventional latent variable dialogue generation models [6, 24, 35], where the latent variable ν is directly utilized to guide the response generation, we project the latent variable ν into the topic proportion θ. During response generation, θ can be employed as a topic distribution vector ξ_{latent}.

2.3 Topic-Enhanced Response Generation

Topic-diversified neural dialogue generation model generates a topic-specific response on two stages:

(1) The individual topic-specific decoder generates the response by dynamically taking account the topic distribution of the context.
(2) Multiple generators are ensembled to generate candidate topic-specific responses and then the most appropriate response is selected with respect to the topic proximity between the context and responses.

2.3.1 Topic-Enhanced Decoding. Given the topic distribution vector ξ (ξ can be either ξ_{clas} or ξ_{latent}), we first apply an approximate embedding layer to transform the topic distribution vector into the context topic embedding $\mathbf{e}$:

$$\mathbf{e} = \mathbf{W}_d \xi, \tag{5}$$

where $\mathbf{W}_d$ denotes the topic embeddings. At each time step j, the generator dynamically refers to the topical information. The decoding hidden state $\mathbf{h}'_j$ is formulated according to both the recurrent decoder hidden state $\mathbf{h}_j$ and the context topic embedding $\mathbf{e}$:

$$\mathbf{h}'_j = \mathbf{h}_j + \mathbf{g}_j \odot (\mathbf{W}_f \mathbf{e}), \tag{6}$$

where $\mathbf{W}_f$ is a parameter matrix. $\mathbf{g}_j$ is a gating mechanism, which dynamically controls the effects of the topical information during response generation. $\mathbf{g}_j$ is calculated as:

$$\mathbf{g}_j = \sigma(\mathbf{W}_e \mathbf{e} + \mathbf{W}_h \mathbf{h}'_j), \tag{7}$$

where $\mathbf{W}_e$ and $\mathbf{W}_h$ are parameter matrices. $\sigma(\cdot)$ denotes the sigmoid function.

2.3.2 Multi-topic Ensembling. Intuitively, humans would like to express in specific styles for conversations with different topics and tend to switch between topics. Hence, we ensemble multiple generators to generate candidate topic-specific responses and then select the most appropriate response based on the topical proximity. In particular, for each candidate response y_i, the ranking score is defined as:

$$\mathrm{Score}(y_i) = \xi_i \xi_x^{\mathrm{T}}, \tag{8}$$

where ξ_i and ξ_x are the topic distribution vectors of the response y_i and context x, respectively. The response with the highest ranking score will be adopted as the final response.

2.4 Training & Inference

The topic-diversified dialogue generation model figures out the context topics of the given dialogue with the topic indicator, as well as generates a proper response with respect to the specified topics. To build a unified topic-diversified neural dialogue generation model, we jointly optimize the topic indicator and the ensembled multiple response generators during training.

For the model utilizing the topic classification indicator, we simply combine the training objectives of the topic classification and response generation:

$$\mathcal{J} = \log p(\ell_x|x) + \log p(y|x), \tag{9}$$

where ℓ_x is the topic label for the context x.

Regarding the model augmented with the latent topic indicator, given the definitions in Eq.(3), similarly as Kingma and Welling [13], Miao et al. [18], we formulate a variational lower bound for the generation likelihood, modeling the dialogue reconstruction in

the latent topic indicator and response generation as:

$$\begin{aligned}\mathcal{J} =& \mathbb{E}_{Q(\theta|d)}[\sum_{i=1}^{|d|}[\log \sum_{z_i}[p(w_i|\beta_{z_i})p(z_i|\theta)]] + \log p(y|x)] \\ & - D_{KL}(Q(\theta|d)||P(\theta|x)) \\ \approx & \sum_{i=1}^{|d|} \log p(w_i|\beta, v) + \sum_{j=1}^{T_y} \log p(m_j|m_{j-1}, \mathbf{h}'_j, x) \\ & - D_{KL}(Q(v|d)||P(v|x))\end{aligned}, \quad (10)$$

where $y = \{m_1, m_2, \cdots, m_{T_y}\}$, and the prior estimation of the latent variable v, $P(v|x)$, approximates the posterior $Q(v|d)$, by minimizing the KL divergence between two distributions. We also adopt the previously proposed techniques, KL annealing [2] and Bag-of-words loss [35], to ameliorate the vanishing latent variable problem.

3 EXPERIMENT SETTINGS

3.1 Datasets

To validate our model's effectiveness of handling the multi-topic conversational scenarios, we utilize multiple public available corpora as the experimental dataset, comprising of a movie discussions dataset collected from Reddit [7] and an Ubuntu technical corpus [17] mainly discussing the usage of Ubuntu. 1,019,644 (context, response) pairs were sampled from these datasets, including 1,017,244 pairs for training, and 2,400 for testing.

3.2 Comparison Models

We compare the proposed topic-diversified neural dialogue generation framework with the following state-of-the-art baselines.

- **SEQ2SEQ+Attention**: Attention-based sequence-to-sequence model [1] is a representative baseline. It is denoted as SEQ2SEQ hereafter.
- **CVAE**: Latent variable conversational model [6, 35] is a derivative of the SEQ2SEQ model in which it incorporates a latent variable at the sentence-level to inject stochasticity and diversity.
- **DOM-SEQ2SEQ**: A domain-aware conversational model consisting of multiple domain-targeted SEQ2SEQ models for responses generation and a domain classifier for responses ranking [5].
- **TA-SEQ2SEQ**: TA-SEQ2SEQ [32] incorporates the topical information into the response generation, where the topics are learned from a separate LDA model to enrich the context, resulting with more informative and interesting responses.

3.3 Hyper Parameters

We implemented our model with ParlAI[1] [20]. The response vocabulary size is set to 20,000, and all the out-of-vocabulary words are transformed into a special token UNK. We set the dimension of word embeddings to 300 and all the models (our models and the baselines) use the pretrained word embeddings obtained by running the FastText [11] tool on the training dataset. Utterance lengths are truncated at 60. The learning optimizer used is the Adam [12] with an initial learning rate of 0.001. The L2 regularization is set to 10^{-5}. Regarding the model implementation, we stacked two layers of bi-directional LSTM structures for the encoder and two layers of left-to-right LSTMs for the decoder. All models share the same hidden representation dimension, which is set to 256. The dimension of the latent variable is set to 64. In the latent topic indicator, the MLP model utilizes 512, 512 nodes in the first two layers, respectively. Following Miao et al. [18], the most frequent 2,511 words are taken as the topical word vocabulary used in the latent topic indicator by stemming, filtering stopwords from the dataset. We trained a Twitter LDA model to prepare the topical words used in TA-SEQ2SEQ and set its model-specific parameters following the original paper [32]. All the models are trained with early-stopping, i.e., we stop the model training if the loss does not decrease after 10 validations. We finally report the evaluation scores on the test set.

[1]http://parl.ai/

3.4 Automatic Evaluation Metrics

Automatic evaluating the performance of dialogue models is non-trivial. Liu et al. [16] reported that word-overlap automatic metrics like BLEU [22] or ROUGE [15] are not well correlated with human evaluations. We hence adopted three embedding-based similarity metrics proposed by Liu et al. [16] to evaluate the semantic relevance between the generated response and the ground-truth response: Embedding Average (**Average**), Embedding Extrema (**Extrema**) and Embedding Greedy (**Greedy**) [8, 21]: Embedding Average computes the sentence embedding by averaging all the constituent words in the sentence; for Embedding Extrema, vector extrema score computes the maximum or minimum value of each dimension of word embeddings in the response; and in Embedding Greedy, greedy matching score actually finds the most similar word between the predicted response and the ground truth response.

The embedding-based metrics are alternatives to word-overlap based metrics which actually take the meaning of each word into consideration by calculating the similarity between the generated response and the ground truth response in the embedding space. We employed FastText [11] to obtain the word embeddings.

We also measure informativeness and diversity of the response utilizing the distinct-1 and distinct-2 metrics, following Li et al. [14]. These metrics measure the ratio of distinct unigrams and bigrams in the entire generated responses. A higher ratio of distinct unigrams or bigrams denotes more informative and diverse responses.

4 EXPERIMENT RESULTS

4.1 Topic Classification *vs* Latent Topic Inference

The topic indicator identifies the topic information of a given context. The topic classification indicator (TND-*clas*) classifies the context into a specific topic and is optimized with topic annotations. The latent topic indicator (TND-*latent*) infers the latent topics of the context and is trained by modeling a generative process of the dialogue. Table 1 compares the performance of both topic indicators. We observe that the topic-diversified neural dialogue generation model utilizing latent topic indicator (Table 1 (b)) consistently outperforms the model augmented with a topic classification indicator (Table 1 (a)) on all the automatic evaluation metrics. When viewing the results, we find that TND-*clas* suffers from the label unbalance

Model	Embedding-based metrics (%)			Informativeness metrics (%)	
	Average	Greedy	Extrema	Distinct-1	Distinct-2
SEQ2SEQ	72.22	88.57	42.68	0.3033	0.6051
(a) TND-*clas*	73	89.76	44.55	0.6263	1.429
(b) TND-*latent*	75.22	90.22	45.82	0.7409	2.083
(c) TND (w/o topic-enhanced decoding)	74.36	89.68	45.68	0.6991	1.805
(d) TND (w/o response ensembling)	73.47	89.64	43.64	0.4155	1.012

Table 1: Automatic evaluation results (%) of model variants.

Model	Embedding-based metrics (%)			Informativeness metrics (%)	
	Average	Greedy	Extrema	Distinct-1	Distinct-2
SEQ2SEQ	72.22	88.57	42.68	0.3033	0.6051
CVAE	72	88.45	44	0.6592	1.568
DOM-SEQ2SEQ	73.5	87.38	44.34	0.5295	1.147
TA-SEQ2SEQ	75.16	87	44.43	0.5674	1.234
TND	**75.22**	**90.22**	**45.82**	**0.7409**	**2.083**

Table 2: Evaluations on embedding-based and informativeness metrics (%). The best performance is in boldface.

problem. Therefore it sometimes fails to recognize the topics of the context. Moreover, topic classification requires the predefined topic annotations during training, which further limits the model extendability. Therefore, we exploit the latent topic indicator henceforth. Specifically, the latent topic indicator clusters similar contexts according to the inferred topic distribution, which reflects the natural proximity of the contexts and enables information sharing across different topics.

4.2 Effects of Topic-Enhanced Decoding

We propose two mechanisms in topic-enhanced response generation. The first one is to utilize the topic distribution vector as an additional signal to enhance the response decoding process. To evaluate its effectiveness, we compare a model without using the topic distribution vector for the response generation. In Table 1 (c), we observe the performance drop in comparison with TND-*latent*. This illustrates that the topic information effectively improves response generation.

4.3 Effects of Response Ensembling

To further exploit the topic information, we ensemble multiple topic-specific response generators and choose the final response according to the topical proximity between the context and the response. In Table 1 (d), we notice that TND without response ensembling performs even worse than TND without topic-enhanced decoding. We conjecture that the response ensembling generates multi-perspective responses of a given context, which improves the robustness of the model when the topic information of the context is not clear enough.

4.4 Overall Performance

Table 2 reports our results on the automatic metrics. In general, we observe that the proposed TND surpasses all the compared baselines on the automatic evaluation metrics. Comparing with CVAE, where it infers a latent variable from the context and injects the variability into the response generation, the proposed TND is able to further infer the latent topic distribution of the given context from the latent variable. The improvements of TND against CVAE demonstrate that the latent topic indicator not only enhances the interpretability, but also improves the performance by a large margin.

DOM-SEQ2SEQ simply assigns a domain-specific SEQ2SEQ model for the given context, and our model performs much better than DOM-SEQ2SEQ, which once again verifies the superiority of the inferred latent topic information and the unified multi-topic structure. TA-SEQ2SEQ exploits the external topic information for response generation, and thus achieves better performances comparing to SEQ2SEQ. Whereas our model is able to automatically infer the latent conversational topics, the error accumulation problem is reduced, comparing with exploiting the fixed pretrained topic information as in Xing et al. [32].

Opponent	Win	Loss	Tie	Kappa
TND vs. SEQ2SEQ	43.5%	10.5%	46%	0.5826
TND vs. CVAE	30.5%	11.5%	58%	0.7322
TND vs. DOM-SEQ2SEQ	24%	9.5%	66.5%	0.5547
TND vs. TA-SEQ2SEQ	27.5%	14%	58.5%	0.5132

Table 3: Human evaluation results.

Context	Inferred Topic Distribution	Response
I am still deciding, can you verify whether or not they have free parking?		**SEQ2SEQ**: I have no idea what you are looking for. **CVAE**: Yes, it is a guesthouse in the north. **DOM-SEQ2SEQ**: Yes, it is located at 154 chesterton road. **TA-SEQ2SEQ**: I have a few options for you. **TND**: Yes, it does have free parking.
Yes, I'd like to know the entrance fee, please?		**SEQ2SEQ**: The address is UNK. **CVAE**: The address is unit su43, grande arcade, saint andrews street. **DOM-SEQ2SEQ**: The postcode is cb23bu. Is there anything else i can help you with? **TA-SEQ2SEQ**: The phone number is 01223 __unk__. **TND**: The entrance fee is free.
The beginning scene was 1997 and Galahad said something about the mistake 17 years ago to Arthur so it's 2014 in the movie.		**SEQ2SEQ**: I'm not sure why it was a movie. **CVAE**: I'm not sure what you're talking about. **DOM-SEQ2SEQ**: I'm not sure that's a movie. **TA-SEQ2SEQ**: I'm not sure if it's a good movie. **TND**: That's the best movie.
Regardless, I think the Bond series has netted a greater profit than F&F , just due to the length of time.		**SEQ2SEQ**: I think it was a good movie. **CVAE**: I'm not sure what you're talking about. **DOM-SEQ2SEQ**: I'm not sure that's the best movie. **TA-SEQ2SEQ**: I'm not sure what you mean... **TND**: I really agree with that.
My question is what is the recommended program to install to get the installer thinking it is interfacing with the internet? have you loaded the repos?		**SEQ2SEQ**: I'm not sure, I'm not sure what I'm looking for. **CVAE**: I'm not sure what I'm trying to do. **DOM-SEQ2SEQ**: The postcode is cb41da. The postcode is cb41da. **TA-SEQ2SEQ**: I'm not sure if it's a bot or something... **TND**: It's not working to a new install.

Table 4: Test samples of our model and the baselines. The bar graphs show the topic distributions.

Movie	Train	Ubuntu	Restaurant	Network
watch	travel	install	food	router
movie	arrives	grub	priced	wireless
youtube	train	apt	moderate	ip
scene	leaves	kernel	restaurants	address
trailer	leaving	nvidia	expensive	phone
movies	london	cd	indian	network
watched	departing	sudo	serves	card
film	arrive	installed	centre	eth0
dc	admission	boot	range	dhcp
batman	station	ubuntu	parking	wifi

Table 5: Topics by the words with top-10 highest probability discovered by the latent topic indicator.

4.5 Human Evaluation and Case Study

We further validate the effectiveness of TND by carrying out human evaluations following Wang et al. [29]. We randomly selected 200 samples from the test set. For each case, given a context-response pair, a pair of generated responses ($response_1$, $response_2$) are provided, one is from the proposed model TND and the other is from the comparison model. Three well-educated evaluators, who have no knowledge about which system the response is from, are required to rate among win ($response_1$ is better) loss ($response_2$ is better) and tie (they are equally good or bad), considering four factors: context relevance, logical consistency, fluency and informativeness. Note that cases with different rating options will be counted as "tie".

Table 3 summarizes the results of subjective evaluation. TND surpasses all the comparison models, which is consistent with the automatic evaluation results. The kappa scores indicate that the annotators came to a fair agreement in the judgment.

Table 4 lists several real cases in the test set. We find that the topic distributions are well inferred by the latent topic indicator. The responses generated by our model are more relevant to the given context in the topic perspectives.

4.6 Visualization

We further dive into the latent topic indicator to get some insights of how it improves the response generation. We observe that the topic distribution distilled by the latent topic indicator is highly interpretable. Table 5 shows the topics by the words with top-10 highest probabilities learned by the latent topic indicator. For the convenience of visualization, the topic number is set to 5. Latent topic indicator effectively induces the commonness and variations across different topics.

5 RELATED WORK

Wen et al. [30] proposed a multi-domain adaptation schema by first training a model with an out-of-domain dataset and then fine-tuning on a small in-domain dataset. In our model, we build a

unified framework for multi-topic conversations. It learns to generate responses for different topics simultaneously. Choudhary et al. [5] built a domain-aware char-bot which generates responses for different domains by utilizing multiple domain-specific SEQ2SEQ models and then re-ranking the generated responses through a domain classifier. In order to improve the informativeness of the response, Xing et al. [32] incorporated topic information into the sequence-to-sequence dialogue response generation. Wang et al. [28] biased the generation process with a topic restriction. However, their topic information is obtained through pre-trained models. Jin et al. [10], Wang et al. [29], Yao et al. [33] leveraged the predicted keywords to boost the response informativeness, which does not involve topic modeling actually. The topic indicator in our model is capable of not only classifying the context into specific topics but also inferring the latent topic information within a unified framework, which enables the topical information sharing between multi-topic conversations. Furthermore, instead of utilizing a vanilla decoder, the topic-enhanced decoder in our model dynamically controls the effects of the topical information during response generation.

With respect to the latent variable neural dialogue generation models [6, 24, 26, 35], previous works directly introduced the variational autoencoder [13] into the dialogue response generation to address the generic response issue of deterministic generation process. They argue that the variational latent variable injects the stochasticity and diversity into the response generation. Although the latent variable is able to promote the response diversity, the meaning representations of the latent variables remain uninterpretable. Chen et al. [3] used the latent variable to randomly access the relevant dialogue history. Zhao et al. [34] attempted to discover interpretable latent variables by using discrete latent actions. Wen and Luong [31] augmented the response generation with the latent topics, whereas their latent topics also remain uninterpretable. The latent topic inference network in our model projects the latent variable into an explainable topic distribution, which enhances the model with a more interpretable latent variable, and increases the response qualities as well. The latent variable is optimized through a generative process of the given dialogue.

6 CONCLUSION

In this work, we propose a topic-diversified neural dialogue generation framework—TND, which leverages the common ground and the difference across conversation topics by identifying the topics of the input context and enhancing the response generation. Specifically, two variants of topic indicator are provided to figure out the topics of the context, where a topic classification indicator classifies the context into a specific topic, and a latent topic indicator infers the latent topics by modeling a generative process of the given dialogue. Topic-wised responses are generated through multiple topic-specific generators, while the topical information is controlled dynamically to guide the response generation. The final response is selected according to the topical proximity between the context and generated responses. Extensive experiments on a large scale conversation dataset show that TND outperforms the state-of-the-art baselines on both the automatic evaluation metrics and human evaluations. The visualization of the inferred topics further demonstrates the superiority of the proposed framework.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *ICLR*.

[2] Samuel R. Bowman, Luke Vilnis, Oriol Vinyals, Andrew M. Dai, Rafal Józefowicz, and Samy Bengio. 2016. Generating Sentences from a Continuous Space. In *SIGNLL*. 10–21.

[3] Hongshen Chen, Zhaochun Ren, Jiliang Tang, Yihong Eric Zhao, and Dawei Yin. 2018. Hierarchical Variational Memory Network for Dialogue Generation. In *WWW*. 1653–1662.

[4] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In *EMNLP*. 1724–1734.

[5] Sajal Choudhary, Prerna Srivastava, Lyle H. Ungar, and João Sedoc. 2017. Domain Aware Neural Dialog System. *CoRR* abs/1708.00897 (2017). arXiv:1708.00897

[6] Stephen Clark and Kris Cao. 2017. Latent Variable Dialogue Models and their Diversity. In *EACL*. 182–187.

[7] Jesse Dodge, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander H. Miller, Arthur Szlam, and Jason Weston. 2015. Evaluating Prerequisite Qualities for Learning End-to-End Dialog Systems. *CoRR* abs/1511.06931 (2015). arXiv:1511.06931

[8] Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *NIPS*, Vol. 2.

[9] Yves Grandvalet, Johnny Mariéthoz, and Samy Bengio. 2005. A Probabilistic Interpretation of SVMs with an Application to Unbalanced Classification. In *NIPS*. 467–474.

[10] Xisen Jin, Wenqiang Lei, Zhaochun Ren, Hongshen Chen, Shangsong Liang, Yihong Zhao, and Dawei Yin. 2018. Explicit State Tracking with Semi-Supervision for Neural Dialogue Generation. In *CIKM*.

[11] Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. 2016. FastText.zip: Compressing text classification models. *arXiv preprint arXiv:1612.03651* (2016).

[12] Diederik P. Kingma and Jimmy Ba. 2014. Adam: A Method for Stochastic Optimization. In *ICLR*.

[13] Diederik P. Kingma and Max Welling. 2013. Auto-Encoding Variational Bayes. *CoRR* abs/1312.6114 (2013). arXiv:1312.6114

[14] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A Diversity-Promoting Objective Function for Neural Conversation Models. In *NAACL-HLT*. 110–119.

[15] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of summaries. In *ACL*. 10.

[16] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How NOT To Evaluate Your Dialogue System: An Empirical Study of Unsupervised Evaluation Metrics for Dialogue Response Generation. In *EMNLP*. 2122–2132.

[17] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. 2015. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems. In *SIGDIAL*. 285–294.

[18] Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering Discrete Latent Topics with Neural Variational Inference. In *ICML*, Vol. 70. 2410–2419.

[19] Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural Variational Inference for Text Processing. In *ICML*. 1727–1736.

[20] A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. 2017. ParlAI: A Dialog Research Software Platform. *arXiv preprint arXiv:1705.06476* (2017).

[21] Jeff Mitchell and Mirella Lapata. 2008. Vector-based Models of Semantic Composition. In *ACL*. 236–244.

[22] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In *ACL*. 311–318.

[23] Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C. Courville, and Joelle Pineau. 2016. Building End-To-End Dialogue Systems Using Generative Hierarchical Neural Network Models. In *AAAI*. 3776–3784.

[24] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. 2017. A Hierarchical Latent Variable Encoder-Decoder Model for Generating Dialogues. In *AAAI*. 3295–3301.

[25] Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural Responding Machine for Short-Text Conversation. In *ACL*. 1577–1586.

[26] Xiaoyu Shen, Hui Su, Yanran Li, Wenjie Li, Shuzi Niu, Yang Zhao, Akiko Aizawa, and Guoping Long. 2017. A Conditional Variational Framework for Dialog Generation. In *ACL*. 504–509.

[27] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to Sequence Learning with Neural Networks. In *NIPS*. 3104–3112.

[28] Di Wang, Nebojsa Jojic, Chris Brockett, and Eric Nyberg. 2017. Steering Output Style and Topic in Neural Response Generation. In *EMNLP*.

[29] Wenjie Wang, Minlie Huang, Xin-Shun Xu, Fumin Shen, and Liqiang Nie. 2018. Chat More: Deepening and Widening the Chatting Topic via A Deep Model. In *SIGIR*. 255–264.

[30] Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Lina Maria Rojas-Barahona, Pei-Hao Su, David Vandyke, and Steve J. Young. 2016. Multi-domain Neural Network Language Generation for Spoken Dialogue Systems. In *NAACL-HLT*. 120–129.

[31] Tsung-Hsien Wen and Minh-Thang Luong. 2018. Latent Topic Conversational Models. *CoRR* abs/1809.07070 (2018). arXiv:1809.07070

[32] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic Aware Neural Response Generation. In *AAAI*.

[33] Lili Yao, Yaoyuan Zhang, Yansong Feng, Dongyan Zhao, and Rui Yan. 2017. Towards Implicit Content-Introducing for Generative Short-Text Conversation Systems. In *EMNLP*.

[34] Tiancheng Zhao, Kyusong Lee, and Maxine Eskénazi. 2018. Unsupervised Discrete Sentence Representation Learning for Interpretable Neural Dialog Generation. In *ACL*. 1098–1107.

[35] Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. 2017. Learning Discourse-level Diversity for Neural Dialog Models using Conditional Variational Autoencoders. In *ACL*. 654–664.