Exemplar Guided Neural Dialogue Generation

Hengyi Cai[1,2*], **Hongshen Chen**[3], **Yonghao Song**[1], **Xiaofang Zhao**[1] and **Dawei Yin**[4]
[1]Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China
[2]University of Chinese Academy of Sciences, Beijing, China
[3]Data Science Lab, JD.com, China
[4]Baidu Inc., China
{caihengyi, songyonghao, zhaoxf}@ict.ac.cn, ac@chenhongshen.com, yindawei@acm.org

Abstract

Humans benefit from previous experiences when taking actions. Similarly, related examples from the training data also provide exemplary information for neural dialogue models when responding to a given input message. However, effectively fusing such exemplary information into dialogue generation is non-trivial: useful exemplars are required to be not only literally-similar, but also topic-related with the given context. Noisy exemplars impair the neural dialogue models understanding the conversation topics and even corrupt the response generation. To address the issues, we propose an exemplar guided neural dialogue generation model where exemplar responses are retrieved in terms of both the text similarity and the topic proximity through a two-stage exemplar retrieval model. In the first stage, a small subset of conversations is retrieved from a training set given a dialogue context. These candidate exemplars are then finely ranked regarding the topical proximity to choose the best-matched exemplar response. To further induce the neural dialogue generation model consulting the exemplar response and the conversation topics more faithfully, we introduce a multi-source sampling mechanism to provide the dialogue model with both local exemplary semantics and global topical guidance during decoding. Empirical evaluations on a large-scale conversation dataset show that the proposed approach significantly outperforms the state-of-the-art in terms of both the quantitative metrics and human evaluations.

1 Introduction

Sequence-to-sequence (SEQ2SEQ) learning [Bahdanau *et al.*, 2015; Sutskever *et al.*, 2014; Cho *et al.*, 2014] has being a state-of-the-art neural network framework for response generation. It treats dialogue generation as a source to target sequence translation problem, where an encoder network [Cho *et al.*, 2014] encodes the context into a vector representing the semantics of the context, and then a decoder network generates the response word-by-word, conditioned on the context vector. Though effective, common wisdom suggests that these models are plagued by the notorious problem of dull, safe responses [Li *et al.*, 2016; Zhang *et al.*, 2018].

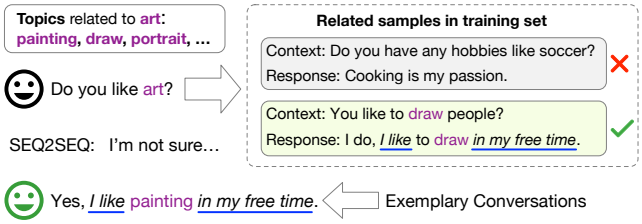

Figure 1: Given an input dialogue context, multiple related samples can be retrieved from the training set according to literal text similarity. The upper one is inappropriate since it shows little relevance with the given post message "Do you like art?" regarding the topic "art". Whereas the lower example correlates well with the given context in terms of the talking topics. The final response "I like painting in my free time" is constructed by referring to such exemplary response template.

This phenomenon occurs partially because that existing models attempt to generate responses for all those conversation contexts based solely on its learnt model parameters [Pandey *et al.*, 2018]. Since human dialogues typically conducted in an open-ended and highly subjective way, capturing all information required to generate responses barely by the model parameters is not necessarily adequate.

Fortunately, exemplary conversations, which can be exploited to improve the response generation, are usually embodied in the training set. As observed in Figure 1, by referring to the exemplar expression "I like ___ in my free time", the user composes an appropriate utterance "I like painting in my free time" responding to the post message "Do you like art?". Such closely related samples provide the model with explicit referable exemplary information that benefit the dialogue generation when responding to a given input message. Based on this observation, it is reasonable to extend the neural dialogue generation model to explicitly take into consideration such relevant exemplary data from the full training set. To augment neural dialogue generation with exemplar conversations, similar examples are first retrieved from training

*Work done at Data Science Lab, JD.com.

data, and the exemplar responses are then fed into the decoder as exemplar vectors to generate the response [Pandey *et al.*, 2018].

Nevertheless, we observe that, effectively fusing such exemplary information into dialogue generation is not that straightforward, but still rather challenging: although the retrieved conversations are similar literally, some of them are topic-unrelated with the given context. Figure 1 shows two exemplar conversations. Both of them are similar to the given context "Do you like art?" literally. However, the upper one discusses "soccer", while the given context and the lower conversation talk about "art". Those topic-unrelated exemplar responses hinder the neural dialogue generation model understanding the exact topic of the conversation, and the model may refer to inappropriate exemplar responses during generation. What's more, simply encoding exemplar responses into hidden vectors further aggravates the inaccurate dialogue topic problem.

Viewing this, in this paper, we propose an Exemplar guided Neural Dialogue generation model—End, where exemplar responses are retrieved in terms of both the text similarity and the topic proximity through a two-stage exemplar retrieval model. In the first stage, a small subset of conversations is retrieved from a training set given a dialogue context. These candidate exemplars are then finely ranked regarding the topical proximity to choose the best-matched exemplar response and guide the dialogue response generation. To further enhance the neural dialogue generation model leveraging the exemplar response and the conversation topics effectively, we introduce a multi-source sampling mechanism during decoding, where the response word can be drawn from the vocabulary embedding space and exemplary collections. The exemplar response is utilized as a soft response template, which can be viewed as local exemplary signals, whereas the dialogue topics serve as global exemplary semantics.

We evaluate the proposed exemplar guided neural dialogue generation model on a real-life conversation dataset. Our experiments reveal that the proposed approach effectively exploits the exemplary information and achieves significant improvement over the strong baselines.

2 Exemplar Guided Neural Dialogue Generation

In this work, we design a neural dialogue system, in which the response generation is guided by the exemplary conversations. Unlike the conventional neural dialogue generation models, the proposed model maintain and actively exploit the training corpus during response generation. As illustrated in Figure 2, End mainly consists of two components: (a) Given an input message x, an exemplary conversation retriever distills the closely related conversations regarding both the text similarity and the topic proximity. (b) An encoder-decoder model generates the final response y under the guidance of the recognized exemplary information.

2.1 Exemplary Conversation Retriever

The proposed exemplary conversation retriever first retrieves a small subset of conversations from a training set and then refine the retrieved conversations based on topic proximity to ameliorate the noisy exemplar issue.

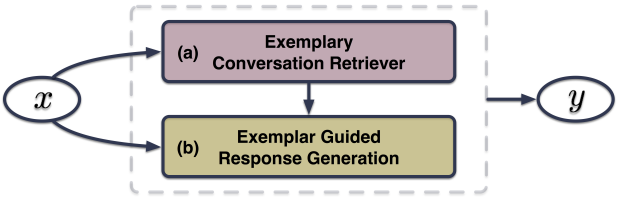

Figure 2: Schematic illustration of our proposed framework.

First-round Exemplar Retrieval. Given a context, the related conversations are gathered by the *exemplar responses retriever*, which recalls the exemplar responses from the training set based on the semantic distance between the given context and the candidate dialogue contexts.

With a huge number of training examples, it is prohibitively expensive at run time to calculate the proximity to the query context iteratively over the whole set of exemplars. We hence index the whole training set in term space. For a query context x, the top-N potential exemplar context-response pairs $(c^{(i)}, r^{(i)})_{i=1}^{N}$ are retrieved by BM25 [Robertson and Zaragoza, 2009]. Note that other sophisticated retrieval models can also be applied in the first round retrieving, e.g., locality sensitive hashing [Indyk and Motwani, 1998].

As aforementioned, retrieving the exemplar responses solely based on the superficial context words is risky, since the noisy exemplars will mislead the model and the resultant irrelevant response usually makes people end the conversation quickly. We therefore finely rank the set of retrieved exemplars and choose the best-matched exemplar response, based on topical proximity. We first introduce the variational topic inference and then elaborate the exemplars reranking.

Latent Topic Inference. We introduce the neural variational topic model [Miao *et al.*, 2017] to approximate the conversation topics of a given dialogue d. d is composed of a context-response pair (x, y). Following Miao *et al.* [2017], we adopt an inference network to parameterize the latent topic distribution θ and a multinomial softmax generative model to reconstruct the conversation based on the topic vectors from the latent topic distribution. More concretely, a latent variable ν is parameterized by an inference network $P(\nu|\mu_{pri}(x), \sigma_{pri}(x))$, which approximates the posterior $Q(\nu|\mu_{pos}(d), \sigma_{pos}(d))$. P and Q are conditioned on a draw from a Gaussian distribution. Outputs of functions μ and σ are parameters of the Gaussian distribution, which are computed using multilayer perceptrons (MLP). We use the reparameterization trick [Kingma and Welling, 2014] to guarantee differentiability when sampling from Gaussian distributions. The topic distribution θ is then built using ν by $\theta = \text{softmax}(\text{Linear}(\nu))$. Given θ, the dialogue d is reconstructed by computing the marginal likelihood:

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

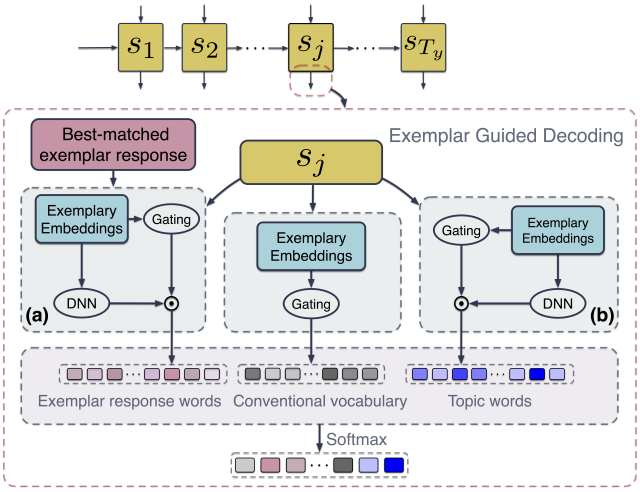

Figure 3: Details of the j-th word generation in the decoder. (a) Decoding with exemplar response. (b) Decoding with context topics. The gating mechanism dynamically controls all the information channels.

where the log-likelihood of a word w_i can be factorized as:

$$\begin{aligned}\log p(w_i|\beta,\theta) &= \log \sum_{z_i}[p(w_i|\beta_{z_i})p(z_i|\theta)] \\ &= \log(\theta\cdot\beta^{\mathrm{T}}).\end{aligned} \tag{2}$$

z_i is the topic assignment and β is the topics-words similarity matrix. We further introduce $F \in \mathbb{R}^{M\times H}$ as the topic word embeddings, $\Lambda \in \mathbb{R}^{K\times H}$ as the topic embeddings and generate the topic-words similarity matrix β by: $\beta_k = \mathrm{softmax}(F\cdot\Lambda_k^{\mathrm{T}})$, where K represents the topic number, M denotes the number of topic words and H stands for the embedding size.

Exemplar Response Refining. To refine the retrieved exemplar responses using the latent topics and make the dialogue generation robust to the noisy exemplars, we finely rank the set of retrieved exemplars and choose the best-matched exemplar response, based on topical proximity. For each candidate exemplar context-response $(c^{(i)}, r^{(i)})$, the ranking function is as follows:

$$\mathrm{Score}((c^{(i)}, r^{(i)})) = \theta_i\cdot\theta_x^{\mathrm{T}}, \tag{3}$$

where θ_i is the topic proportion of the exemplar context-response pair $(c^{(i)}, r^{(i)})$, computed through the latent topic inference network, and θ_x is the topic proportion of the query context. The retrieved exemplar response with the highest ranking score—r_t, will be adopted to guide the response generation.

2.2 Exemplar Guided Response Generation

Given an input sequence $\{w_1, w_2, ..., w_n\}$, we adopt a bidirectional RNN to transform the discrete tokens into hidden representations. The final hidden states in two directions are then concatenated to form a sentence representation $\mathbf{h} = [\overrightarrow{\mathbf{h}}_n^{\mathrm{T}}; \overleftarrow{\mathbf{h}}_n^{\mathrm{T}}]$. For the input context x and the retrieved exemplar response r, they are both in the form of sequences, and will be transformed into their corresponding distributed representations $\mathbf{h}_x$ and $\mathbf{h}_r$ through separate bidirectional LSTMs respectively.

The decoder generates the response sequentially through a forward RNN. For the j-th word, the scoring function relies on the decoding hidden state $\mathbf{s}_j$ and the exemplary embeddings which consist of the exemplar response r_t and the topic embedding $\mathbf{t}$ computed as $\mathbf{t} = \theta\Lambda$. The architecture of the decoder is shown in Figure 3.

Vanilla Decoder. The vanilla decoder simply generates the response word y_j conditioned on the context x and the previous generated words $y_{[1:j-1]}$. Then, the probability of the response y is as follows:

$$\begin{aligned}p(y|x) &= \prod_{j=1}^{T_y} p(y_j|y_{[1:j-1]}; x) \\ &= \prod_{j=1}^{T_y} p(y_j|\mathbf{s}_j),\end{aligned} \tag{4}$$

where T_y is the length of response y and $\mathbf{s}_j$ denotes a combination of the source context information and the recurrent hidden state up to time step j. The vanilla decoder does not exploit any exemplary information. All required information are conveyed through the hidden $\mathbf{s}$. Relying merely on the context hidden states, the model often gets in trouble for generating appropriate responses.

Decoding with Exemplar Response. Conventional practice exploiting the exemplar response simply encodes the exemplars as hidden vectors, which may lead to the loss of exemplary information. We hence employ the exemplar response as a soft language template, allowing the response word to be drawn from the exemplary collections. As shown in Figure 3.(a), we integrate the exemplar response r_t into the response generation. In decoding, the generation probability $p(y_j)$ can be defined as:

$$p(y_j) = p_{\Omega_V}(y_j) + p_\Gamma(y_j), \tag{5}$$

where $p_{\Omega_V}(y_j)$ and $p_\Gamma(y_j)$ are the probabilities of generating y_j from the conventional vocabulary Ω_V and exemplar response r_t, respectively, and are computed as:

$$\begin{aligned}p_{\Omega_V}(y_j = w) &= \begin{cases}\frac{1}{Z_1}e^{\Psi_{\Omega_V}(w)}, & w\in\Omega_V \\ 0, & w\notin\Omega_V\end{cases} \\ p_\Gamma(y_j = w) &= \begin{cases}\frac{1}{Z_1}\sum_{m:r_t^m=w} e^{\Psi_\Gamma(r_t^m)}, & w\in r_t \\ 0, & w\notin r_t\end{cases}\end{aligned} \tag{6}$$

where $Z_1 = \sum_{w\in\Omega_V} e^{\Psi_{\Omega_V}(w)} + \sum_{w\in r_t} e^{\Psi_\Gamma(w)}$ is the normalization term. Ψ_{Ω_V} and Ψ_Γ are the scoring functions and r_t^m stands for the m-th word in r_t. Ψ_{Ω_V} and $\Psi_\Gamma(r_t^m)$ are defined as:

$$\Psi_{\Omega_V}(w) = \mathbf{w}^{\mathrm{T}}\rho_V(\mathbf{s}_j); \quad \Psi_\Gamma(r_t^m) = \mathbf{w}^{\mathrm{T}}\rho_\Gamma(\mathbf{s}_j, \mathbf{h}_{r_t}), \tag{7}$$

where ρ_V and ρ_Γ are non-linear transformation functions, like multi-layer perceptrons, to project the input into the scoring vector. $\mathbf{h}_{r_t}$ is the hidden representation of exemplar response r_t, and $\mathbf{w}$ is a one-hot indicator vector of word w.

Decoding with Dialogue Topics. To further provide the dialogue response generation with global topical exemplary semantics, we extend the response generation by sampling response words from the topic words vocabulary Ω_Λ. When generating a response word, the model predicts the word probabilities by referring to both the conventional vocabulary and the topic words vocabulary, as illustrated in Figure 3.(b). Then, the generation probability $p(y_j)$ can be defined as:

$$p(y_j) = p_{\Omega_V}(y_j) + p_{\Omega_\Lambda}(y_j), \tag{8}$$

where $p_{\Omega_\Lambda}(y_j)$ is defined by:

$$p_{\Omega_\Lambda}(y_j = w) = \begin{cases} \frac{1}{Z_2} e^{\Psi_{\Omega_\Lambda}(w)}, & w \in \Omega_\Lambda \\ 0, & w \notin \Omega_\Lambda \end{cases}. \tag{9}$$

$Z_2 = \sum_{w\in\Omega_V} e^{\Psi_{\Omega_V}(w)} + \sum_{w\in\Omega_\Lambda} e^{\Psi_{\Omega_\Lambda}(w)}$ is the normalization term. $\Psi_{\Omega_\Lambda}(w)$ is computed by:

$$\Psi_{\Omega_\Lambda}(w) = \mathbf{w}^{\mathrm{T}} \rho_\Lambda(\mathbf{s}_j, \mathbf{t}), \tag{10}$$

where ρ_Λ is a non-linear transformation function, like multilayer perceptrons, to project the input $\mathbf{s}_j$ and $\mathbf{t}$ into the scoring vectors. $\mathbf{w}$ is a one-hot indicator vector of word w. $p_{\Omega_V}(y_j = w)$ is formulated similarly as in Eq.(6).

Exemplar-Enhanced Gating. In order to dynamically control the effects of the exemplary information in the process of dialogue response generation, we further introduce a gating mechanism for the scoring functions. We utilize the exemplary embeddings, including the exemplar response $\mathbf{h}_{r_t}$ and topic embedding $\mathbf{t}$, together with the decoding hidden state $\mathbf{s}_j$, to perform gating. The scoring functions are updated as gated scoring functions:

$$\begin{aligned}
\Psi_{\Omega_V}(w) =& \mathcal{G}_V(g_j = 0|\mathbf{h}_{r_t}, \mathbf{t}, \mathbf{s}_j)\mathbf{w}^{\mathrm{T}}\rho_V(\mathbf{s}_j) + \\
& \mathcal{G}_V(g_j = 1|\mathbf{h}_{r_t}, \mathbf{t}, \mathbf{s}_j)\mathbf{w}^{\mathrm{T}}\rho_V^g(\mathbf{s}_j, \mathbf{h}_{r_t}, \mathbf{t}) \\
\Psi_\Gamma(w) =& \mathcal{G}_\Gamma(g_j = 0|\mathbf{h}_{r_t}, \mathbf{t}, \mathbf{s}_j)\mathbf{w}^{\mathrm{T}}\rho_\Gamma(\mathbf{s}_j, \mathbf{h}_{r_t}) + \\
& \mathcal{G}_\Gamma(g_j = 1|\mathbf{h}_{r_t}, \mathbf{t}, \mathbf{s}_j)\mathbf{w}^{\mathrm{T}}\rho_\Gamma^g(\mathbf{s}_j, \mathbf{h}_{r_t}, \mathbf{t}) \\
\Psi_{\Omega_\Lambda}(w) =& \mathcal{G}_\Lambda(g_j = 0|\mathbf{h}_{r_t}, \mathbf{t}, \mathbf{s}_j)\mathbf{w}^{\mathrm{T}}\rho_\Lambda(\mathbf{s}_j, \mathbf{t}) + \\
& \mathcal{G}_\Lambda(g_j = 1|\mathbf{h}_{r_t}, \mathbf{t}, \mathbf{s}_j)\mathbf{w}^{\mathrm{T}}\rho_\Lambda^g(\mathbf{s}_j, \mathbf{h}_{r_t}, \mathbf{t})
\end{aligned}, \tag{11}$$

where $\mathcal{G}_V$, $\mathcal{G}_\Gamma$ and $\mathcal{G}_\Lambda$ are the gating functions, which can be implemented as simple as a sigmoid function or as a gated recurrent unit. At each time step, the gating functions control whether or not the next response word is generated, referring to the exemplar response and topic information. When $\mathcal{G}(g_j = 0)$, it indicates that the decoder hidden state $\mathbf{s}_j$ is informative enough to score the next response words, while $\mathcal{G}(g_j = 1)$ denotes that the exemplary information should be taken more into account.

Exemplar Guided Decoder. The full version of the exemplar guided exemplar decoder jointly utilizes all the proposed mechanisms to generate the final response. When generating a word y_j, both the exemplar response and topic words are integrated through gated multi-source sampling mechanisms. The generation probability of y_j can be finalized as:

$$p(y_j) = p_{\Omega_V}(y_j) + p_\Gamma(y_j) + p_{\Omega_\Lambda}(y_j), \tag{12}$$

and $Z = \sum_{w\in\Omega_V} e^{\Psi_{\Omega_V}(w)} + \sum_{w\in r_t} e^{\Psi_\Gamma(w)} + \sum_{w\in\Omega_\Lambda} e^{\Psi_{\Omega_\Lambda}(w)}$ is used to normalize the scores. Figure 3 details the j-th word generation in the proposed decoder.

Optimizing. End are trained to maximize the generation likelihood of the given parallel corpus as well as the variational lower bound of the latent topic inference:

$$\begin{aligned}
\mathcal{J} \approx & \sum_{j=1}^{T_y} \log p(y_j|y_{[1:j-1]}; \mathbf{h}_x, \mathbf{h}_{r_t}, \mathbf{t}) + \sum_{i=1}^{|d|} \log p(w_i|\beta, \nu) \\
& - D_{KL}(Q(\nu|\mu_{pos}(d), \sigma_{pos}(d))\|P(\nu|\mu_{pri}(x), \sigma_{pri}(x)))
\end{aligned}, \tag{13}$$

where the first term is the conventional response generation objective, the second term is the dialogue generation objective in latent topic inference, and the third term is the KL divergence between two Gaussian distributions.

3 Experiments

Dataset. To validate our model's effectiveness, we construct an open-domain conversation corpus spanning over several public available dialogue dataset, including a movie discussions dataset collected from Reddit [Dodge *et al.*, 2015], and a Ubuntu technical corpus [Lowe *et al.*, 2015] discussing about the usage of Ubuntu. These datasets are widely used in dialogue researches [Pandey *et al.*, 2018]. 57,402 context-response pairs are sampled for training, 3,000 for validation and 3,000 for testing.

Hyper Parameters and Reproducibility. Our model is implemented using ParlAI [Miller *et al.*, 2017]. We truncate all context utterances to length 100 and response utterance to length 50. We take the most frequent 20,000 words as conventional vocabulary. Regarding model implementations, the RNNs in the encoder and the decoder utilize 2-layer LSTM structures with 256 hidden cells for each layer. The latent variable size is set to 64. The size of latent topics is set to 10. The dimensions of word embedding and topic embedding matrix are set to 300. Top-10 candidate exemplar responses are retrieved by the exemplar responses retriever in the first round retrieving. The Adam [Kingma and Ba, 2014] optimizer with a learning rate of 0.001 is used to train the models. We use early stopping with log-likelihood on the validation set as the stopping criteria.

Baselines. We compare the proposed End with the following state-of-the-art baselines. 1) **SEQ2SEQ+Attention**: Attention-based sequence-to-sequence model [Bahdanau *et al.*, 2015] is a representative baseline. It is denoted as SEQ2SEQ hereafter; 2) **CVAE**: Latent variable conversational model [Clark and Cao, 2017; Zhao *et al.*, 2017] is a derivative of the SEQ2SEQ model in which it incorporates a latent variable at the sentence-level to inject stochasticity and diversity; 3) **LAED**: A recurrent encoder-decoder model [Zhao *et al.*, 2018] using discrete latent actions for interpretable neural dialogue generation; 4) **EED**: A conversation model [Pandey *et al.*, 2018] that utilize similar examples from training data to generate responses; 5) **CopyNet**: An attention-based sequence-to-sequence model augmented with copy mechanism [Gu *et al.*, 2016]; 6) **TAS2S**: TAS2S [Xing *et al.*, 2017] incorporates the topic 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.

Models	Relevance (%)				Informativeness (%)		
	BLEU	Ave.	Gre.	Ext.	Dist-1	Dist-2	Dist-3
SEQ2SEQ	0.8097	72.34	65.43	39.67	0.3662	1.1	1.984
CVAE	1.059	73.7	65.52	40.84	0.5207	2.042	4.131
LAED	1.182	74.13	66.23	41.11	0.5861	2.38	4.582
EED	1.259	74.81	65.95	39.44	0.2186	0.6267	1.054
CopyNet	0.9179	74.27	66.19	42.19	0.8357	2.501	4.354
TAS2S	0.8845	74.81	66.11	42.18	0.7999	2.863	5.575
END	**1.281**	**74.97***	**66.38**	**42.69***	**2.057***	**6.292***	**10.33***

Table 1: Evaluations on relevance and informativeness metrics (%). "*" denotes that result is statistically significant with $p < 0.01$.

Automatic Evaluation Metrics. The BLEU [Papineni *et al.*, 2002] metric is employed to measure the response quality. Besides, in order to evaluate the semantic relevance between the generated response and the ground-truth response, we also adopted the embedding-based similarity metrics proposed by Liu *et al.* [2016]: Embedding Average (**Ave.**), Embedding Extrema (**Ext.**) and Embedding Greedy (**Gre.**). To measure informativeness and diversity of the response, we also exploited the Distinct-n metrics (n={1,2,3}).

Overall Performance. In Table 1, we compare the results of our model with all the baselines in terms of both the relevance metrics and the informativeness metrics. Overall, we observe that our model exceeds all the comparison models on automatic evaluation metrics.

For relevance metrics, CVAE, LAED and TAS2S surpass the original attention-based SEQ2SEQ baseline regarding the BLEU score and embedding-based evaluation, which is consistent with the reports in Xing *et al.* [2017]. It indicates that both the latent variable and the topic information slightly enable SEQ2SEQ generating more appropriate responses. EED exhibits competitive BLEU score improvement among baselines, implying that the exemplar response is helpful to promote the response relevance. CopyNet also improves the embedding-based metrics a lot compared with SEQ2SEQ, owing to its ability to copy words from the context. The improvements of END over SEQ2SEQ are even larger than the baseline models, which demonstrates the benefits of exploiting the exemplary information from the training corpus. As the topic information is automatically inferred during response generation, the error accumulation problem is reduced, comparing with exploiting the fixed pretrained topic information as in Xing *et al.* [2017]. In terms of informativeness, CVAE, LAED, CopyNet and TAS2S also achieve better performances comparing to SEQ2SEQ, whereas our model presents much larger improvements in Distinct-{1,2,3} metrics. It implies that, under the guidance of exemplary information, our model is more adept at generating diverse dialogue responses. We also conducted significance tests with t-test for relevance metrics and Sign-test [Dixon and Mood, 1946] for Distinct metrics. END significantly outperforms the baselines on the majority of metrics with p-value < 0.01.

Model Ablation. To examine the effectiveness of the exemplar response and topic information in response generation, we conducted model ablations by removing particular modules from END. As shown in Table 2, we observe that without either the exemplar response or conversation topics, the performance drops rapidly with respect to all the evaluation metrics. It verifies the effectiveness of decoding with the exemplary information. Note that the performance drops when the topic information excludes from the exemplar decoding, affirming that the conversation topics are helpful to refine the retrieved exemplars for response generation. In line (3) of Table 2, when both the exemplar response and conversation topics together incorporated in response generation, compared to the decoding with either the exemplar response or conversation topics, the model obtains much better performance. Finally, the exemplar-enhanced gating mechanism further improves the performance and achieves the best results (line (4) in Table 2).

END Ablations	Relevance (%)				Informativeness (%)		
	BLEU	Ave.	Gre.	Ext.	Dist-1	Dist-2	Dist-3
(1) *w/o* Exemplar	**1.302**	73.42	65.26	41.75	1.348	3.835	6.233
(2) *w/o* Topic	1.160	74.36	66.04	41.41	1.661	4.976	8.433
(3) *w/o* Gating	1.228	74.42	66.17	42.60	1.814	5.299	8.662
(4) Full Model	1.281	**74.97**	**66.38**	**42.69**	**2.057**	**6.292**	**10.33**

Table 2: Ablation study on the END framework.

Opponent	Win	Loss	Tie	Kappa
END vs. SEQ2SEQ	57%	21%	22%	0.6996
END vs. CVAE	56.2%	19.8%	24%	0.6231
END vs. LAED	57.4%	19.6%	23%	0.5932
END vs. EED	55.7%	20.9%	23.4%	0.6783
END vs. CopyNet	55.6%	23.7%	20.7%	0.5819
END vs. TAS2S	52.5%	22%	25.5%	0.6647

Table 3: The results of human evaluation.

Human Evaluation. We also carried out the human study through comparisons between our model and the baselines, following Wang *et al.* [2018]. For each case, given a context-response pair, two generated responses were provided, one is from our model and the other is from the comparison model. We randomly selected 500 samples from the test set. Three well-educated students were invited to rate which one is better considering four factors: context relevance, logical consistency, fluency and informativeness. Notice that cases with different evaluations are marked as "tie". Table 3 summarizes the results of subjective evaluation. The kappa scores indicate that the annotators came to a fair agreement in the judgment. Not surprisingly, as shown in Table 3, END consistently outperforms all the comparison systems. However, we notice that CVAE and LAED exhibit insignificant improvements comparing with SEQ2SEQ. We analyzed the bad cases and found that CVAE still suffers from the general, meaningless responses. Meanwhile, though LAED generates more informative responses, some responses are irrelevant to the given context. Augmented with the exemplar response and inferred latent topics, END introduces a competitive boost in response quality, which is in line with the automatic evaluation, confirming the superior performance of the proposed method.

Case Study. Table 4 lists several responses generated by END and the baselines. The exemplar responses and the top-5 inferred topic words are also presented. The exemplar responses and discovered topic words contain informative con-

	Input context (C), exemplar response (E), top-5 inferred topic words (T) and the reference (R)	Generated responses
1	**C**: I need to book a train for Friday please. **E**: Booking was successful. Reference number is: UNK. Can I assist you further today? **T**: **trains**, leaving, **travel**, tickets, arrives **R**: Great. Where are you headed?	**SEQ2SEQ**: I have booked you a taxi. The reference number is UNK. **CVAE**: I have a train of the train that leaves at the train that arrives at the train. **LAED**: I'll try that, I'll try it. **EED**: You're welcome, have a great day! **CopyNet**: I will book it for you. **TAS2S**: I'm not sure if you're using the Ubuntu-desktop. **END**: There are a number of **trains**. Would you like to **travel**?
2	**C**: My first one was from the main download page, my second one was from the official torrent. How are they even wrong? **E**: Try to **download** Ubuntu **from** a different mirror, then burn a new ISO. **T**: http, works, error, click, file **R**: Get a new iso like i said.	**SEQ2SEQ**: I'm not sure... I'm not sure if it's a bug. **CVAE**: I'm not sure it's a UNK. **LAED**: I'm not sure how to do it. **EED**: OK, but I'm not sure how to do that... **CopyNet**: I think it's a good one. **TAS2S**: I'm not sure if you're using the Ubuntu-desktop **END**: I'm trying to **download** some stuff **from** the repo.
3	**C**: The end of a movie called Akira. **E**: I just disagree **with you**. I think the filming was good but the story was way too empty. **T**: imdb, **movies**, **watch**, comedy, youtube **R**: What is the movie about?	**SEQ2SEQ**: I'm not sure that's the point. **CVAE**: I'm not sure that's the same thing. **LAED**: I'm not sure if it's a good movie. **EED**: I think it's the best movie I've seen. **CopyNet**: I'm not sure what the movie is. **TAS2S**: I'm not sure if you're talking about the UNK. **END**: I wanna **watch** **movies** **with you**.

Table 4: Test samples of END and the baselines. The reference is the ground-truth response in the dataset. Common words among exemplar responses, inferred topic words and responses generated by END are highlighted.

tents that are utilized by the END model for generating responses. In the first example (line 1 in Table 4), we notice that the topic words "trains" and "travel" are decoded into the response. This is in concert with our intuition that latent topic inference helps to provide END with informative topical information. As for example 2 in Table 4, the exemplar response provides the model with a soft template "download ___ from ___" and influences *how to say it*. Regarding line 3 in Table 4, END benefits from both the inferred topic words and the exemplar response, and composes an appropriate phrase "watch movies with you" by consulting the topic words "movies, watch" and the exemplar phrase "with you". In general, we found that END is able to effectively fuse such exemplary information into the dialogue response generation.

4 Related Work

To improve the neural dialogue systems, prior art typically focuses its attention on elaborately exploiting the given conversation for response generation, by using latent variables [Serban *et al.*, 2017; Clark and Cao, 2017; Zhao *et al.*, 2017; Zhao *et al.*, 2018], hierarchical history modeling [Serban *et al.*, 2016; Chen *et al.*, 2018], input-dependent parameterization [Cai *et al.*, 2019] or predicting keywords from context [Yao *et al.*, 2017; Wang *et al.*, 2018].

In contrast to the above models, our model takes into account more information from the whole training set than a current dialogue context for generating the final responses. Pandey *et al.* [2018] encoded the exemplar response into a hidden vector, which may lead to the loss of information, while we utilize the exemplar response as a template through a copying mechanism [Gu *et al.*, 2016]. To ameliorate the problem of noisy exemplars, we also refine the retrieved exemplar responses using the inferred latent topics. While the principal idea of both the papers remains similar, the difference lies in the mechanism of gathering and incorporating the retrieved data. Xing *et al.* [2017] incorporated topic information into the SEQ2SEQ dialogue response generation. Wang *et al.* [2017] biased the generation process with a topic restriction. However, their topic information is obtained through pre-trained models. Yao *et al.*; Wang *et al.* [2017; 2018] leveraged the predicted keywords to boost the response informativeness, which does not involve topic modeling actually. While in our model, the latent topics are automatically inferred from the given dialogue and the model is trained within a unified framework in an end-to-end fashion. Another difference is that, they only utilized the topic words to guide the response generation, while we enhance the response generation with both exemplar responses and latent topics.

5 Conclusion

In this work, we present END—a novel neural dialogue generation model which considers not only a given conversation context, but also a set of relevant exemplary conversations from the training corpus in the process of response generation. To provide the dialogue model with beneficial exemplars, the proposed approach adopts a two-stage exemplar retrieval model: in the first stage, a small subset of conversations is retrieved from a training set given a dialogue context; these candidate exemplars are then refined regarding the topical proximity to choose the best-matched exemplar response. To effectively fuse such exemplary information into dialogue response generation, we further introduce a multi-source sampling mechanism to provide the dialogue model with both local exemplary semantics and global topical guidance during response decoding. Extensive experiments show that the proposed model outperforms the state-of-the-art baselines and is capable of generating more informative and relevant responses.

Acknowledgements

We would like to thank all the reviewers for their insightful and valuable comments and suggestions. Hongshen Chen and Xiaofang Zhao are the corresponding authors.

References

[Bahdanau *et al.*, 2015] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In *ICLR*, 2015.

[Cai *et al.*, 2019] Hengyi Cai, Hongshen Chen, Cheng Zhang, Yonghao Song, Xiaofang Zhao, and Dawei Yin. Adaptive parameterization for neural dialogue generation. In *EMNLP-IJCNLP*, 2019.

[Chen *et al.*, 2018] Hongshen Chen, Zhaochun Ren, Jiliang Tang, Yihong Eric Zhao, and Dawei Yin. Hierarchical variational memory network for dialogue generation. In *WWW*, 2018.

[Cho *et al.*, 2014] Kyunghyun Cho, Bart van Merrienboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using rnn encoder–decoder for statistical machine translation. In *EMNLP*, 2014.

[Clark and Cao, 2017] Stephen Clark and Kris Cao. Latent variable dialogue models and their diversity. In *EACL*, 2017.

[Dixon and Mood, 1946] Wilfrid J Dixon and Alexander M Mood. The statistical sign test. *Journal of the American Statistical Association*, 41(236):557–566, 1946.

[Dodge *et al.*, 2015] Jesse Dodge, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander H. Miller, Arthur Szlam, and Jason Weston. Evaluating prerequisite qualities for learning end-to-end dialog systems. *CoRR*, abs/1511.06931, 2015.

[Gu *et al.*, 2016] Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O. K. Li. Incorporating copying mechanism in sequence-to-sequence learning. In *ACL*, 2016.

[Indyk and Motwani, 1998] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: Towards removing the curse of dimensionality. In *STOC*, 1998.

[Kingma and Ba, 2014] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. In *ICLR*, 2014.

[Kingma and Welling, 2014] Diederik P. Kingma and Max Welling. Auto-encoding variational bayes. In *ICLR*, 2014.

[Li *et al.*, 2016] Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In *NAACL-HLT*, 2016.

[Liu *et al.*, 2016] Chia-Wei Liu, Ryan Lowe, Iulian Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *EMNLP*, 2016.

[Lowe *et al.*, 2015] Ryan Lowe, Nissan Pow, Iulian Serban, and Joelle Pineau. The ubuntu dialogue corpus: A large dataset for research in unstructured multi-turn dialogue systems. In *SIGDIAL*, 2015.

[Miao *et al.*, 2017] Yishu Miao, Edward Grefenstette, and Phil Blunsom. Discovering discrete latent topics with neural variational inference. In *ICML*, 2017.

[Miller *et al.*, 2017] A. H. Miller, W. Feng, A. Fisch, J. Lu, D. Batra, A. Bordes, D. Parikh, and J. Weston. Parlai: A dialog research software platform. *arXiv preprint arXiv:1705.06476*, 2017.

[Pandey *et al.*, 2018] Gaurav Pandey, Danish Contractor, Vineet Kumar, and Sachindra Joshi. Exemplar encoder-decoder for neural conversation generation. In *ACL*, 2018.

[Papineni *et al.*, 2002] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, 2002.

[Robertson and Zaragoza, 2009] Stephen E. Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.

[Serban *et al.*, 2016] Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, 2016.

[Serban *et al.*, 2017] Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C. Courville, and Yoshua Bengio. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, 2017.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *NIPS*, 2014.

[Wang *et al.*, 2017] Di Wang, Nebojsa Jojic, Chris Brockett, and Eric Nyberg. Steering output style and topic in neural response generation. In *EMNLP*, 2017.

[Wang *et al.*, 2018] Wenjie Wang, Minlie Huang, Xin-Shun Xu, Fumin Shen, and Liqiang Nie. Chat more: Deepening and widening the chatting topic via A deep model. In *SIGIR*, 2018.

[Xing *et al.*, 2017] Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. Topic aware neural response generation. In *AAAI*, 2017.

[Yao *et al.*, 2017] Lili Yao, Yaoyuan Zhang, Yansong Feng, Dongyan Zhao, and Rui Yan. Towards implicit content-introducing for generative short-text conversation systems. In *EMNLP*, 2017.

[Zhang *et al.*, 2018] Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *ACL*, 2018.

[Zhao *et al.*, 2017] Tiancheng Zhao, Ran Zhao, and Maxine Eskénazi. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *ACL*, 2017.

[Zhao *et al.*, 2018] Tiancheng Zhao, Kyusong Lee, and Maxine Eskénazi. Unsupervised discrete sentence representation learning for interpretable neural dialog generation. In *ACL*, 2018.