Sequential Topic Selection Model with Latent Variable for Topic-Grounded Dialogue

Xiaofei Wen[1,2], **Wei Wei**[1,2*] and **Xian-Ling Mao**[3]

[1]Cognitive Computing and Intelligent Information Processing Laboratory, School of Computer Science and Technology, Huazhong University of Science and Technology
[2]Joint Laboratory of HUST and Pingan Property & Casualty Research (HPL)
[3]Department of Computer Science and Technology, Beijing Institute of Technology
xfwen@hust.edu.cn, weiw@hust.edu.cn, maoxl@bit.edu.cn

Abstract

Recently, topic-grounded dialogue system has attracted significant attention due to its effectiveness in predicting the next topic to yield better responses via the historical context and given topic sequence. However, almost all existing topic prediction solutions focus on only the current conversation and corresponding topic sequence to predict the next conversation topic, without exploiting other topic-guided conversations which may contain relevant topic-transitions to current conversation. To address the problem, in this paper we propose a novel approach, named **S**equential **G**lobal **T**opic **A**ttention (SGTA) to exploit topic transition over all conversations in a subtle way for better modeling post-to-response topic-transition and guiding the response generation to the current conversation. Specifically, we introduce a latent space modeled as a Multivariate Skew-Normal distribution with hybrid kernel functions to flexibly integrate the *global*-level information with sequence-level information, and predict the topic based on the distribution sampling results. We also leverage a topic-aware *prior-posterior* approach for secondary selection of predicted topics, which is utilized to optimize the response generation task. Extensive experiments demonstrate that our model outperforms competitive baselines on prediction and generation tasks.

1 Introduction

Dialog systems have been widely used in a variety of applications. Recent efforts in dialogue systems aim at improving the diversity of agent responses (Zhang et al., 2018; Zhou et al., 2021; Liu et al., 2022c) and endowing agents with the ability to exploit knowledge, express empathy and retain personality (Adiwardana et al., 2020; Roller et al., 2021; Wei et al., 2021; Liu et al., 2022b). However, in many real-world scenarios (*e.g.,* conversational recommendation, shopping guide and psychological counseling), conversational agents need to proactively steer the conversation by smoothly transforming the conversation topic into a specified one. Therefore, topic-grounded controllable conversation has recently attracted extensive attention.

Indeed, topic-grounded controllable dialogue is a task to obtain informative responses through a series of predicted topics and the given context. Recently, topic-grounded conversation works mainly focus on modeling the post-to-response **topic transition** for predicting the next topic to guide the response generation. Many deep learning based approaches are proposed for the task, which utilize topic similarity information as well as additional knowledge information to model dynamic topic transitions (Tang et al., 2019; Wu et al., 2019; Qin et al., 2020; Zhou et al., 2020b; Zhong et al., 2021). These approaches have achieved encouraging results, but they still face the issues as follows. *First*, some of these methods infer the next turn topic only with current turn topic embedding (Tang et al., 2019; Qin et al., 2020; Zhong et al., 2021). However, a topic may be inherently associated with several previous topics. Thus they may suffer from the inability of sufficiently modeling the sequential topic-transitions into the topic embeddings as ignoring the inherent topic dependencies over historical topic sequence. *Second*, almost all previous approaches model the topic transferring information only over the current topic sequence while neglecting the useful topic-transition patterns from other sequences, since frequently co-occurring topics over all topic-sequences are more likely to related to each other. Therefore, fully exploring such information is conceptually advantageous to accurately modeling the topic transition for better infer the next topic.

In this paper, we propose a Transformer-based sequential modeling approach, named **S**equential **G**lobal **T**opic **A**ttention (SGTA), to fully exploit

---

topic-co-occurrences over all topic sequences for better modeling the post-to-response topic transitions and accurately predicting the next topic to the current conversation. Specifically, SGTA consists of three key elements: topic sequence s, global co-occurrence matrix c and latent variable z. The relationships among these elements are elaborated with the graphical model in Figure 1. Specifically, we propose a new latent space based on the value variation of the Transformer and use it to model the contextual relation within current topic sequence. The latent space is modeled as a Multivariate Skew-Normal (MSN) distribution (Azzalini and Valle, 1996) due to the flexibility of its **covariance** parameter to integrate multiple information and the specificity of its **shape** parameter to control the skewness of the distribution. The covariance parameter Σ is designed via task-specific kernel functions for measuring the similarity of pair-wise sequences based on the topic representations. The shape parameter α is used to represent the unique relationships of different topics in a sequence. These two parameters are constructed efficiently using current sequence information and other sequences' *global* information. By sampling the MSN distribution on Transformers, we provide a reparameterization of the MSN distribution to enable amortized inference over the latent space. Based on the sampling results, the model simulates the topic transition relationships in the sequence for predicting the next topic in current conversation.

To properly use predicted words and address fluency and controllability issues in generating responses, we leverage a response-based posterior topic distribution to instruct the model to generate responses based on the predicted topics. The *prior* and *posterior* ideas are widely used in response generation tasks (Lian et al., 2019), which is widely-adapted for avoiding to learn the same conversation patterns to each utterance. Concretely, we utilize the *KL-Div Loss* between the *prior* and *posterior* distributions as part of the overall loss, which allows the model to better adapt to response generation by converging the transition scope of the semantic space.

We summarize the contributions of this work as:

- To the best of our knowledge, this is the first work that exploits *global*-level topic transitions of all sequence information to learn contextual information for topic-related dialogue.
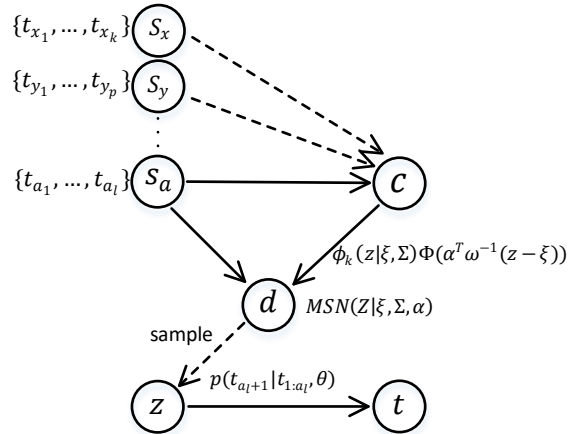
- We propose a unified model to improve the topic

Figure 1: Graphical illustration of topic prediction. Given the current topic sequence s_a and other topic sequences (*e.g.*, s_x and s_y), the global co-occurrence matrix c is obtained with all topic sequences. Parameters of MSN distribution d are constructed from s_a and c in the latent space, and the next turn topic t_{a_l+1} will be predicted based on the latent variable z obtained after sampling.

prediction performance of current sequence by constructing task-specific MSN distribution using sequential information and global information.

- We propose a method that subtly exploits predicted words for response generation and achieves the state-of-the-art on the benchmark dataset compared to competitive baseline methods

2 Related Work

Topic Transition Conversation Recently, several studies attempted to build agents that can actively guide conversations by introducing designated target keyword. Tang et al. (2019) proposed a next-turn keyword predictor and a rule-based keyword selection strategy to solve the topic transition problem, given the conversation history and the target keyword. Qin et al. (2020) improved (Tang et al., 2019) by exploiting the semantic knowledge relation among candidate keywords to achieve smooth keyword transition. Very closely, Zhong et al. (2021) leverages commonsense knowledge for keyword transition prediction through GNN-based models. Inspired by this category of work for controlled topic conversations, Liu et al. (2020) first introduce goals in conversational recommen-
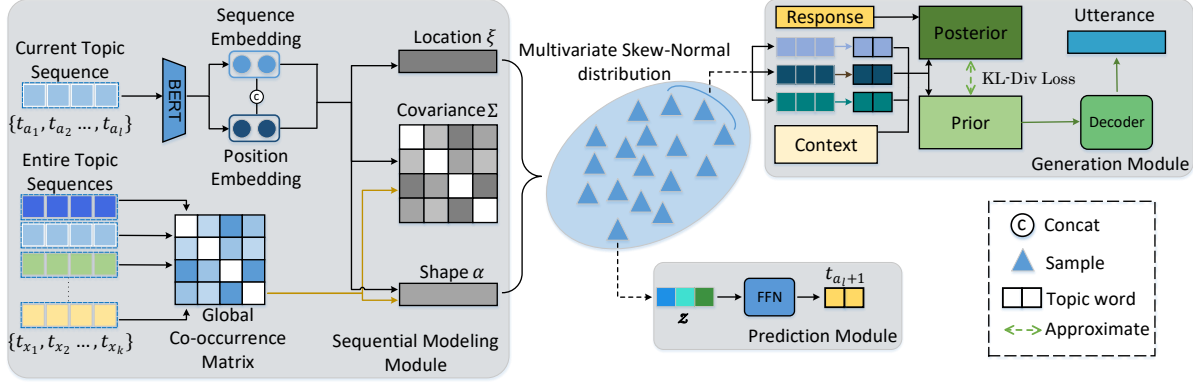
Figure 2: The overall structure of the proposed model SGTA, which consists of (1) a sequential modeling module, (2) a prediction module and (3) a generation module.

dation to enhance the controllability of the conversation process, and Zhou et al. (2020b) utilizes Pre-trained Language Model(PrLM) to capture the given current topic sequence to guide conversational recommendation. Our work follows the task definition in (Zhou et al., 2020b), mainly focusing on topic transition and next-turn topic prediction problem in multi-turn dialogues, given the topic sequence. However, all previous approaches only model the topic-transition information on the current sequence. In contrast, our work learns topic transition information across all sequences to enhance the topic transfer modeling of the current sequence.

Topic-Aware Response Generation Most of early studies on topic-related dialogue fall into two categories, *i.e.,* implicit topic-based (Shang et al., 2015; Serban et al., 2016; Tian et al., 2017) and explicit topic-based (Xing et al., 2017; Wang et al., 2018; Dziri et al., 2019; Liu et al., 2022a). The former aims to model contexts at multiple semantic levels (*i.e.,* topic, style, *etc*) to capture the dynamic conversation topic flow while ignoring the informativeness of responses. The latter generates appropriate utterances on the conversation utterances and given topic information, which relies heavily on manually predefined topic sequences. Due to the excellent performance of the PrLMs on the dialog generation task, Liu et al. (2021) and Zhou et al. (2020b) choose to encode topic information and context by the PrLM to obtain informative and fluent responses. However, it may suffer from topic-noise problem (*i.e.,* raw topic fitting and similar topic selection) when generating responses. In contrast, our work leverages several topics to a *prior* distribution, using *posterior* information on already known

topic to guide the generation of *prior* distribution that affects the generation process.

3 Methodology

We propose a novel model for Topic-guided Dialogue. Figure 2 presents the architecture of our model, which comprises three main components: 1) *Global* Sequential Topic Attention layer; 2) Point-Wise Feed-Forward Network Prediction layer; 3) *Prior* and *Posterior* Response Generation layer. We next present the three components and parameter modeling in detail.

3.1 Problem Statement

Let $\mathcal{C} = \{u_1, u_2, \ldots, u_{|\mathcal{C}|}\}$ be a multi-turn conversation; let $\mathcal{S} = \{t_1, t_2, \ldots, t_l\}$ be the topic sequence of conversation $\mathcal{C}$, and let $\mathcal{T} = \{t_1, t_2, \ldots, t_{|\mathcal{T}|}\}$ be all of topics.

Given the conversation $\mathcal{C}$ and topic sequence $\mathcal{S}$, the task of topic guide conversation aims to predict the top-N topics ($1 \leq N \leq |\mathcal{T}|$) from $\mathcal{T}$ at turn k and generate appropriate response $u_{|\mathcal{C}|+1}$.

3.2 *Global* Sequential Topic Attention layer

In this section, we will describe how to perform serialized embedding of topics in Section 3.2.1 and how to construct MSN distribution based on sequential information and global information in Section 3.2.2.

3.2.1 Embedding layer

To obtain a better semantic representation of the topic words in the sequence, we encode the topic sequences using BERT (Devlin et al., 2019); also, to bring the model to aware previous topics information while addressing the sequence, we add

position embedding for each topic. We use the latest n topics in $\mathcal{S}$, where $n \geq l$, and we pad constant zero vector to make each sequence length equal. Specifically, the topic embedding matrix is defined as $\mathbf{E} \in \mathbb{R}^{|\mathcal{T}| \times d}$, where d denotes the topic embeddings' dimensionality and $\mathbf{E}$ is the estimated by BERT. We extract the input matrix $\widehat{\mathbf{E}} \in \mathbb{R}^{n \times d}$, where $\widehat{\mathbf{E}}_s = \mathbf{E}_{T_s}$. Inspired by Kang and McAuley (2018), we inject a learnable position embedding $\mathbf{P} \in \mathbb{R}^{n \times d}$ into the input embedding:

$$\widehat{X} = [e_i \parallel p_{l-i+1}] \quad (1)$$

where $e_i \in \widehat{\mathbf{E}}_s$ and $p_i \in \mathbf{P}$. We make the reverse location embedding injection due to the unfixed length of the topic sequence. Compared with the regular forward location, the distance between each topic and the predicted topic contains certain effective information (Wang et al., 2020; Zou et al., 2022).

3.2.2 Multivariate Skew-Normal Distribution

The core of our model is constructed based on Transformer's Multi-Head Attention, we choose Transformer due to its excellent ability in sequence modeling problem (Kang and McAuley, 2018; Kim et al., 2020; Ji et al., 2020). For the Scaled-Dot Attention in Multi-Head Attention, given the Equation (2):

$$\text{Attn}\,(Q, K, V) = \text{softmax}(\frac{QK^T}{\sqrt{d_k}})V \quad (2)$$

Since the key to attention mechanism is querying via the alignment score, we turn the alignment score (*i.e.,* scaled dot product of Q and K^T) into a latent variable z, which allows us to adapt the model to apply the mandatory information flexibly, including *global*-level transition and topic relativities. Despite the superior results demonstrated by multivariate normal distribution in modeling the covariance matrix (Fisher and Sun, 2011), which exhibits the intra-sequence topic correlation, due to its forced symmetric shape of the density curve (*i.e.,* suffering from modeling skewness), we devise z to follow the Multivariate Skewed Normal (MSN) distribution (Azzalini and Valle, 1996). According to Azzalini and Capitanio (1999), $z \in \mathbb{R}^k$ is continuous with density function as:

$$f(z) = 2\phi_k(z - \xi; \Sigma)\Phi\{\alpha^T\omega^{-1}(z - \xi)\} \quad (3)$$

where $\Sigma = \omega\psi\omega$ is the covariance matrix and $\psi \in \mathbb{R}^{k \times k}$ denotes correlation parameter, as well

as $\xi = (\xi_1, ..., \xi_k)^T$, $\omega = \text{diag}(\omega_1, ..., \omega_k)$ and $\alpha \in \mathbb{R}^k$ denote the location, scale and shape parameters respectively. Moreover, ϕ_k is the k-dimensional multivariate normal density with the mean ξ and the covariance Σ, and $\Phi(\cdot)$ is the $N(0, 1)$ distribution function. For clarity, we denote the above distribution obediently as Equation (4) and follow the original notation:

$$z \sim MSN(Z \mid \xi, \Sigma, \alpha) \quad (4)$$

As we need to sample the softmax parameter values from MSN distribution by which to revise the alignment score, we adapt Equation (2) to derive the following equation:

$$\text{H} = \text{SGTA}(T_{seq}, G_b) = \text{softmax}(Z)V \quad (5)$$

where T_{seq} denotes the input topic sequence, G_b denotes the *global*-level topic transition co-occurrence matrix, which is constructed by statistically calculating the topic co-occurrence in the whole dataset. Furthermore, $\text{H} = \{h_1, ..., h_n\}$ is denoted as hidden layer output with $h_i \in \mathbb{R}^d$. All parameters modeling will be explained in detail in subsequent sections.

Location The Location ξ represents the mean of the distribution. Considering that we demand the deterministic alignment scores with maximum likelihood to facilitate sampling the MSN distribution, we manipulate the alignment score as:

$$\xi = \text{LeakyReLU}(\frac{(T_{seq}\mathbf{W}_l^Q)(T_{seq}\mathbf{W}_l^K)}{\sqrt{d}}) \quad (6)$$

To keep the standard deviation positive, we choose LeakyReLU as the activation function.

Covariance The covariance Σ indicates the relation between two different topics, which is formed by two subparameters ω and ψ as we mentioned before. For scale ω, given a topic sequence $\mathcal{S}$ with topic position $\mathcal{S}_p^i$ and $\mathcal{S}_p^j$, we normalize them to x_i and x_j and infer the variance $\omega_i^2, \omega_j^2 \in \mathbb{R}_+$ of z respectively, by amortization inference as in Equation (7).

$$\omega_m = \text{LeakyReLU}(\frac{(x_n\mathbf{W}_\omega^Q)(x_m\mathbf{W}_\omega^K)}{\sqrt{d}}) \quad (7)$$

As for correlation ψ, we adopt the kernel function approach for the mixing calculation, since it can efficiently and nonlinearly compute the inner

product of samples in the feature space and calculate high-dimensional distance measures. It's well known that proximal distance metric illustrates proximate relationship. Especially, we comprise a hybrid kernel function for building task-aware metric. Different from Tang et al. (2019) and Ji et al. (2020), we consider global co-occurrence level topic transition information to construct kernel functions for metrics, nevertheless they build only based on the embedding similarity, which is difficult to obtain sufficient information. We then describe the specific implementation of kernel function.

• **Co-occurrence kernel** is determined by the number of co-occurrence between pairwise topics, which is linearly dependent with the topic pairs. The co-occurrence kernel is designed as:

$$k_{co}(x_i, x_j) = \omega_i \omega_j \log(\frac{P_{ij}^\beta}{p_i p_j}) \qquad (8)$$

where P_{ij} is the co-occurrence value of x_i and x_j in the common sequence and p_i, p_j are their individual occurrence values. β is the factor that determines the impact degree of the co-occurrence value P_{ij}.

• **Topic pair kernel** is determined by topic transition pattern pair representation, which is strongly correlated with the topic pairs association. The topic pair kernel $k_{tp}(x_i, x_j)$ is designed as:

$$k_{tp} = \omega_i \omega_j (\exp(-\gamma \|x_i - x_j\|^2) + x_i x_j) \quad (9)$$

which combines both Gaussian kernel and . In particular, the Gaussian kernel is primarily used to characterize the similarity between samples, and $\gamma > 0$ is the unique hyperparameter of Gaussian kernel function.

To make full use of the above two part information, we design the above two kernel functions to be summed as shown in Equation (10):

$$k(x_i, x_j) = k_{co}(x_i, x_j) + \eta r k_{tp}(x_i, x_j) \quad (10)$$

where $r = \text{softmax}(x\mathbf{W}_x + \mathbf{b}_x)$, $\mathbf{W}_x \in \mathbb{R}^{d \times d}$, $\mathbf{b}_x \in \mathbb{R}^d$ and η denote the learnable parameters. After relation modeling, we set the correlation matrix $\psi_{ij} = \frac{k(x_i, x_j)}{\omega_i \omega_j}$ and substitute ψ into $\Sigma = \omega \psi \omega$ to infer the *covariance* Σ.

Shape The shape parameter α reflects the relation of each topic in the sequence with the last topic. It contains the consideration for relative position information, while this correlation helps the model to learn the implicit transition pattern between *topic*-level and *position*-level relations. Specifically, we let $\alpha_i = s_{n-i} \frac{\widehat{\alpha}_i}{\max(\widehat{\alpha})}$ represents correlation between the final topic t_n and topic t_i, where $\widehat{\alpha}$ is a ratio parameter which mirrors the estimation and consideration at *topic*-level, while s_{n-i} is a relative scaling parameter with *position*-level information.

We use the co-occurrence matrix G_b to calculate the ratio parameter $\widehat{\alpha}_i$, divided into intra-sequence level and global level parts, representing the sum of linear arrangement from t_i to t_n and influence factors summation of the top-m frequently co-occurring topics with t_i, respectively. $g_{i,j}$ is the i-th row and j-th column value of G_b (*i.e.,* the value of t_i and t_j). The detailed formula for $\widehat{\alpha}_i$ is given below:

$$\widehat{\alpha}_i = \underbrace{\sum_{j=1}^n g_{i,j} g_{j,n}}_{\textit{intra-sequence level}} + \underbrace{\sum_{l=1}^m g_{i,k_l} g_{k_l,n}}_{\textit{global level}} \quad (11)$$

Equation (11) calculates $\widehat{\alpha}_i$ by utilizing the sum of two dot products, which exhibits the correlation between t_i and t_n, where g_{i,k_l} stands for the l-th element after decreasing the order of all $g_{i,k}$ values with $k \in |\mathcal{T}|$, and m is an adjustable variable. Inspired by Ji et al. (2020), we utilize the average of the pairwise dot product sums of the remaining topics in the sequence to replace $g_{j,j}$ since it is an invalid value in matrix G_b, which can be expressed as:

$$g_{j,j} \Longleftarrow \underbrace{\frac{\sum_{p \in \{1,...,n\} \setminus j} g_{j,p}}{n-1}}_{\textit{average of remaining co-occurrence}} \quad (12)$$

Similar to ω in section *Covariance*, we define the scaling parameter s_{n-i} as Eq. (13):

$$s_{n-i} = \text{LeakyReLU}(\frac{(x_n \mathbf{W}_s^Q)(x_i \mathbf{W}_s^K)|n-i|}{\sqrt{d}}) \quad (13)$$

Loss function Given the above construction of the distribution parameters, the latent variable z needs to be inferred according to MSN distribution. Following Kingma and Welling (2013), z can be inferred by optimizing the lower bound on the evidence of Jensen's inequality for the marginal logarithm $p(y_n)$ while predicting the $(n + 1)$-th topic. We present the loss function as shown in

Eq. (14).

$$\mathcal{L}_z(\theta) = \mathbb{E}_z \left[\log p\left(y_n \mid z \right) \right]$$

$$\leq \log \int p\left(y_n \mid z \right) p(z) dz = \log p\left(y_n \right)$$

$$(14)$$

We also design the co-occurrence loss L_{rank} by listwise ranking loss (Cao et al., 2007) to match topic-co-occurrence relevance and ranking consistency. The total loss $\mathcal{L}$ is defined as the sum of the co-occurrence loss $\mathcal{L}_{rank}$ and the latent variable loss $\mathcal{L}_z$:

$$\mathcal{L} = \mathcal{L}_z + \delta \mathcal{L}_{rank} \qquad (15)$$

where δ is a hyperparameter. We also use the reparameterization trick (Kingma et al., 2015) to ensure that the model is trainable.

3.3 Point-Wise Feed-Forward Network Prediction layer

We apply the PointWise Feed-Forward network in Transformer to the output of the model and incorporate location dependent information. The pointwise feedforward network consists of two linear layers and the activation layer. The output $F = \{\text{FFN}(h_1), ..., \text{FFN}(h_d)\}$. We also stacks b self-attentive blocks to adaptively and hierarchically extract previous consumed topic information and learn complex topic transition patterns.

After the above adapted query attention, we predict the next topic (given the first n topics) based on $F_n^{(b)}$:

$$r_{i,n} = F_n^{(b)} \mathbf{E}_i \qquad (16)$$

where $r_{i,n}$ is the relevance of topic t_i being the next topic given the first n topics (*i.e.*, $t_1, ..., t_n$), and $\mathbf{E}_i$ is the BERT embedding of topic t_i. Empirically, high relevance represents a compact transfer relation, so we use ranking r to make predictions for next topic t_i.

3.4 *Prior* and *Posterior* Response Generation layer

After obtaining the predicted topic words, it is essential to employ them flexibly to generate smooth response for completing exhaustive dialogue interaction. Most of the previous methods select topics for generating responses based on the similarity between the previous topic and next turn's topic or context, which can be regarded as the topic *prior* distribution, however there are actual conversation scenarios in which multiple candidate topics are pertinent for the previous topic transitions (*e.g.,*

movie→music and *movie→friend*). Both *music* and *friend* can be considered as the next topic word for *movie*. Nevertheless, with the *posterior* distribution constructed from query and response, model can be more empirical in selecting the appropriate topics for generation.

Particularly, during training, we design that the posterior distribution $p(t = t_i|q, r)$ containing the response information r is approximated by the prior distribution $p(t = t_i|q)$ containing only historical information q, which includes context c and history topic sequence $t_{1:n-1}$ (*i.e.*, $q = \{c, t_{1:n-1}\}$). We evaluate the approximation loss with *KL-Div Loss*, as shown in the following equation:

$$\mathcal{L}_{KL}(\theta) = \sum_{i=1}^{N} p(t = t_i|q, r) \log \frac{p(t = t_i|q, r)}{p(t = t_i|q)}$$

$$(17)$$

The posterior distribution can guide the model to generate natural responses according to secondary selected topic, and converging the semantic transition range while generation. Compared to the previous method, $BOW\,Loss$ serves as a part of the overall loss for evaluating each word of the predicted responses, while allowing for a better fit to the true responses, and meanwhile ensure the controllability of the generated utterances. Following Lian et al. (2019), the $BOW\,Loss$ is defined as:

$$\mathcal{L}_{BOW}(\theta) = -\mathbf{E}_{t_i \sim p(t|c,r)} \sum_{j=1}^{m} \log p(r_j|t_i) \quad (18)$$

where θ denotes the model parameters in Eq. (17) & Eq. (18). Additionally, we apply Transformer (Vaswani et al., 2017) as the decoder of our model to generate outstanding and convincing responses. For name consistency, we denote this generation method via SGTA in the whole paper.

4 Experiment

4.1 Dataset

TG-ReDial dataset The TG-ReDial dataset (Zhou et al., 2020b) is composed of 10,000 conversations between seekers and recommenders. It contains a total of 129,392 utterances from 1,482 users, covering 2,571 topics. The dataset is constructed in a topic-guided manner, where each conversation includes a topic sequence as well as a conversation target, and both parties communicate sequentially according to the topic sequence to accomplish the target and thus achieve the recommendation. It is

Model	with target			without target		
	Hit@1	**Hit@3**	**Hit@5**	**Hit@1**	**Hit@3**	**Hit@5**
PMI	0.0349	0.0927	0.1290	-	-	-
Kernel	0.0418	0.0957	0.1125	-	-	-
DKRN	0.4015	0.4821	0.5068	-	-	-
CKC	0.5914	0.7857	0.8265	0.3269	0.4505	0.4961
MGCG	0.5861	0.7528	0.8094	0.3157	0.4386	0.4483
Conversation-BERT	0.6072	0.7893	0.8105	0.2952	0.4099	0.4585
Topic-BERT	0.6104	0.7966	0.8147	0.4185	0.5520	0.5971
TG-CRS	0.6128	0.8157	0.8302	0.4352	0.5726	0.6291
SGTA	**0.6208***	**0.8523***	**0.8671***	**0.4487***	**0.6030***	**0.6546***

Table 1: Automatic evaluation of topic predictions task. **Bold** text indicates the best result. Significant improvements compared to the best baseline are marked with $*$ (t-test, $p < 0.05$).

notable that the conversation targets are ***optional*** in our task setting as additional information in the model input, and we will give a detailed experimental comparison later. On average, each conversation in the TG-ReDial dataset has 7.9 topics and a utterance contains 19 words.

4.2 Implementation details

We implemented SGTA and related baseline experiments in PyTorch. The default parameters for all experiments are set as follows: we set the batch size to 16 and the embedding size is set to 768. We used two self-attention blocks and one head for sequence modeling following Kang and McAuley (2018). It is worth noting that we constructed the co-occurrence matrix only from the training set data. We used Adam optimizer (Kingma and Ba, 2015) with a learning rate initialized to $1e^{-5}$ and the dropout rate is set to 0.1.

4.3 Baselines

Our baselines for assess the performance of topic prediction and response generation come in two groups:

Topic Guide Prediction: (1) PMI and Kernel (Tang et al., 2019) employ the forcing strategy of selecting keywords with higher similarity to the target. (2) DKRN (Qin et al., 2020) introduces a graph routing mechanism for keyword search based on Tang et al. (2019). (3) MGCG (Liu et al., 2020) adopt a completion judgment mechanism in the exploration of topic sequence. (4) CKC (Zhong et al., 2021) follows the work of (1) and (2), introduces the commonsense graph and utilizes the

GNN method to select keywords. (5) Conversation-BERT, Topic-BERT and TG-CRS (Zhou et al., 2020b) respectively input the context, historical topic sequence and the concatenation of them to BERT for encoding and predicting.

The above baselines are briefly divided into two sub-categories, *similarity*-based (*i.e.,* PMI, Kernel, DKRN and CKC) and *sequence*-based (*i.e.,* MGCG, Conversation-BERT, Topic-BERT and TG-CRS). The former predicts based on the similarity of next turn topic and target while the latter predicts based on the topic sequence and context.

Topic Response Generation: (1) KBRD (Chen et al., 2019) enhances the transformers' decoder with keywords searched on the knowledge graph. (2) KGSF (Zhou et al., 2020a) incorporates both word-oriented and entity-oriented knowledge graphs to enhance response. (3) MGCG adopts multi-task learning framework to predict topics and generates responses with posterior distributions. (4) Transformer (Vaswani et al., 2017) is a widely used multi-head attention encoder-decoder structure. (5) TG-CRS utilizes the predicted word concating context input to GPT-2 (Radford et al.) to generate responses.

4.4 Evaluation metrics

Automatic evaluation To evaluate the topic prediction task, following the previous work (Tang et al., 2019; Zhou et al., 2020b; Liu et al., 2020), we adopt Hit@$k(k = 1, 3, 5)$ as metric for ranking all the possible topics (topic keywords recall at position k in all keywords). To evaluate the response generation task, following Zhou et al. (2020b), we

Model	Automatic					Human	
	PPL	BLEU-1	BLEU-2	Distinct-1	Distinct-2	Fluency	Informativeness
KBRD	28.022	0.221	0.028	0.004	0.008	1.16	1.30
KGSF	40.758	0.239	0.042	0.015	0.064	1.35	1.35
MGCG	12.386	0.256	0.061	0.012	0.041	1.31	1.26
Transformer	32.856	0.287	0.071	0.013	0.083	1.35	1.31
TG-CRS	**7.223**	0.280	0.065	0.021	0.094	**1.42**	1.43
SGTA	8.539	**0.301***	**0.080***	**0.023***	**0.124***	1.41	**1.47**

Table 2: Automatic evaluation and human evaluation of response generation task. **Bold** text indicates the best result. Significant improvements compared to the best baseline are marked with $*$ (t-test, $p < 0.05$).

adopt perplexity (PPL), BLEU-n($n = 1, 2$) and Distinct-n($n = 1, 2$) for examining the fluency and informativeness of the responses.

Human evaluation For generation tasks, it is necessary to employ manual evaluation to test the ability of the models to make topic-related informative responses. We randomly select 100 dialogues from our model and baseline, recruit four annotators to evaluate several models in two aspects, *i.e., fluency* and *informativeness*. The former measures whether the generated responses are fluent, while the latter measures whether the system introduces relevant topics in response. Score ranges from 0 to 2, and the final average score is calculated as the result.

4.5 Performance on Topic Prediction

Table 1 shows the experimental results of topic prediction task. PMI and Kernel do not perform well, since they cannot consider the topic sequence context. CKC performs significantly better than other *similarity*-based baselines as it exploits the external information in the commonsense knowledge graphs. We notice that TG-CRS outperforms the other baselines since it jointly models dialogue context, topic sequence, and user profile. SGTA gives an increase of 1.3%, 4.5%, and 4.4% over TG-CRS in terms of Hit@1/3/5 respectively with consideration of target information, as SGTA effectively leverages the topic-transition patterns in other topic sequences. It should be clarified that since user-level information is not taken into account in our task, our experimental set-up that contains the target *differs from* the original set-up, where we concatenate the target word embedding as additional information with the topic sequence. It is easy to figure that conversation topic transfer tends to be more uncontrollable without considering target information, where only CKC can make predic-

tions in the *similarity*-based approach due to its having sufficient external information as a prediction base. Moreover, in the *sequence*-based methods, Topic-BERT is fairly better than Conversation-BERT, which shows the effectiveness of topic sequence. SGTA benefits from better topic sequence modeling, improves over the best baseline by 3.1%, 5.4% and 4.2% on Hit@1/3/5, respectively.

4.6 Performance on Response Generation

Table 2 exhibits the generation performance of automatic and manual metrics on the TG-REDIAL dataset. For automatic evaluation, we find that TG-CRS performs best in the baselines, indicating the robustness of the pre-trained model for the generation task. KGSF outperforms KBRD and MGCG in terms of diversity because it combines KG-enhanced item and word embedding to generate the utterance. Transformer performs better on the BLEU metric due to its word-by-word attention mechanism. SGTA performs best on BLEU and Distinct using a prior and posterior approximation without depending on pre-trained language models, improving 7.5%, 23%, 9.5% and 32% over TG-CRS at BLEU-1/2 as well as Distinct-1/2 respectively, though PPL is slightly weaker than TG-CRS's effect. As for the manual results, SGTA improves 2.8% compared to TG-CRS in the informativeness dimension, owing to the fact that the posterior approach makes model learn reasonable topic choices easily, and also SGTA's performance is in the first tier in terms of fluency.

4.7 Ablation study

We conduct ablation study to assess the importance of global co-occurrence (*w/o global*), intra-sequence position (*w/o intra-pos*), topic sequence context (*w/o sequen*), as well as the MSN distri-

Model	with target		
	Hit@1	Hit@3	Hit@5
SGTA	0.6208*	0.8523*	0.8671*
w/o global	0.5852	0.7863	0.8121
w/o intra-pos	0.6081	0.8034	0.8256
w/o sequen	0.5749	0.7687	0.7905
w/o MSN	0.5895	0.7993	0.8203

Table 3: Results of the ablation study for topic prediction task.

Model	with target		
	Hit@1	Hit@3	Hit@5
SGTA	0.6208*	0.8523*	0.8671*
w/o MSN	0.5895	0.7993	0.8203
w/o MSN-ξ	0.6097	0.8107	0.8343
w/o MSN-Σ	0.5964	0.8039	0.8267
w/o MSN-α	0.5982	0.8051	0.8255

Table 4: Results of the ablation study for MSN parameters.

bution (*w/o MSN*), and results are presented in Table 3. Concretely, after removing global topic co-occurrence information (*w/o global*), the average reduction in our model effect is about for 6.6%, which demonstrates the desirability of extracting global information. On this basis, the effect of the model is improved by only removing the intra-sequence position information (*w/o intra-pos*), but it is still lower than our model by 4.2%. Moreover, the experiment to remove the MSN distribution (*w/o MSN*) show that the information selection with skewness is effective under the premise of reasonable modeling parameters. Finally, experiments with the removing of sequential transformer modeling (*w/o sequen*) show the key role of both in the model structure.

We also conduct ablation study to assess the performance of MSN distribution three parameters: Location ξ (*w/o MSN-ξ*), Covariance Σ (*w/o MSN-Σ*) and Shape α (*w/o MSN-α*) and Table 4 shows the results. We use the experimental design with the parameter set to zero, and each ablation study represents the effect of removing only the individual parameter separately. Specifically, shape parameter α and the covariance parameter Σ perform more important effects in the distribution construction.

5 Conclusion

In this paper, we focus on topic-grounded controllable dialogue tasks. We propose a new approach named SGTA. It models the latent space as a MSN distribution utilizing global information, intra-sequence semantic and position information, which allows the model to better integrate the relationships between information and make topic predictions based on the results of distribution sampling. We also utilize a *prior-posterior* distribution approach to generate a new topic response. Extensive experiments on TG-ReDial dataset show that SGTA achieves state-of-the-art performance on prediction and generation tasks.

Acknowledgement

Limitations

Our work suffers from the following limitations: (1) lack of experimental results on a more topic conversation dataset. Although there are few existing topic conversation datasets and TG-ReDial is an explicit dataset at the topic level, experiments on more datasets are necessary for the generalizability of the method. (2) Our topic prediction has yet to be improved in mining global information at the dataset level. After sampling and verifying the experimental results we believe that the extraction by topic word co-occurrence has some shortcomings, such as words with more co-occurrence may not appear under the current round, which requires the integration of context as well as semantics.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

A Azzalini and A Capitanio. 1999. Statistical applications of the multivariate skew normal distribution.

Journal of the Royal Statistical Society. Series B, Statistical Methodology, pages 579–602.

Adelchi Azzalini and A Dalla Valle. 1996. The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726.

Zhe Cao, Tao Qin, Tie-Yan Liu, Ming-Feng Tsai, and Hang Li. 2007. Learning to rank: from pairwise approach to listwise approach. In *Proceedings of the 24th international conference on Machine learning*, pages 129–136.

Qibin Chen, Junyang Lin, Yichang Zhang, Ming Ding, Yukuo Cen, Hongxia Yang, and Jie Tang. 2019. Towards knowledge-based recommender dialog system. 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)*, pages 1803–1813.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar Zaiane. 2019. Augmenting neural response generation with context-aware topical attention. In *Proceedings of the First Workshop on NLP for Conversational AI*, pages 18–31, Florence, Italy. Association for Computational Linguistics.

Thomas J Fisher and Xiaoqian Sun. 2011. Improved stein-type shrinkage estimators for the high-dimensional multivariate normal covariance matrix. *Computational Statistics & Data Analysis*, 55(5):1909–1918.

Mingi Ji, Weonyoung Joo, Kyungwoo Song, Yoon-Yeong Kim, and Il-Chul Moon. 2020. Sequential recommendation with relation-aware kernelized self-attention. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 4304–4311.

Wang-Cheng Kang and Julian McAuley. 2018. Self-attentive sequential recommendation. In *2018 IEEE International Conference on Data Mining (ICDM)*, pages 197–206. IEEE.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential latent knowledge selection for knowledge-grounded dialogue. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *ICLR (Poster)*.

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

Durk P Kingma, Tim Salimans, and Max Welling. 2015. Variational dropout and the local reparameterization trick. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

Rongzhong Lian, Min Xie, Fan Wang, Jinhua Peng, and Hua Wu. 2019. Learning to select knowledge for response generation in dialog systems. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*, pages 5081–5087. International Joint Conferences on Artificial Intelligence Organization.

Jiayi Liu, Wei Wei, Zhixuan Chu, Xing Gao, Ji Zhang, Tan Yan, and Yulin Kang. 2022a. Incorporating casual analysis into diversified and logical response generation. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 378–388, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Yifan Liu, Wei Wei, Jiayi Liu, Xianling Mao, Rui Fang, and Dangyang Chen. 2022b. Improving personality consistency in conversation by persona extending. In *Proceedings of the 31st ACM International Conference on Information amp; Knowledge Management*, CIKM '22, pages 1350–1359, New York, NY, USA. Association for Computing Machinery.

Yuhang Liu, Wei Wei, Daowan Peng, and Feida Zhu. 2022c. Declaration-based prompt tuning for visual question answering. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*, pages 3264–3270. International Joint Conferences on Artificial Intelligence Organization. Main Track.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, and Wanxiang Che. 2021. DuRecDial 2.0: A bilingual parallel corpus for conversational recommendation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4335–4347, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zeming Liu, Haifeng Wang, Zheng-Yu Niu, Hua Wu, Wanxiang Che, and Ting Liu. 2020. Towards conversational recommendation over multi-type dialogs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1036–1049, Online. Association for Computational Linguistics.

Jinghui Qin, Zheng Ye, Jianheng Tang, and Xiaodan Liang. 2020. Dynamic knowledge routing network for target-guided open-domain conversation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 8657–8664.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online. Association for Computational Linguistics.

Iulian Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 30.

Lifeng Shang, Zhengdong Lu, and Hang Li. 2015. Neural responding machine for short-text conversation. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1577–1586, Beijing, China. Association for Computational Linguistics.

Jianheng Tang, Tiancheng Zhao, Chenyan Xiong, Xiaodan Liang, Eric Xing, and Zhiting Hu. 2019. Target-guided open-domain conversation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5624–5634, Florence, Italy. Association for Computational Linguistics.

Zhiliang Tian, Rui Yan, Lili Mou, Yiping Song, Yansong Feng, and Dongyan Zhao. 2017. How to make context more useful? an empirical study on context-aware neural conversational models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 231–236, Vancouver, Canada. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

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 *The 41st international acm sigir conference on research & development in information retrieval*, pages 255–264.

Ziyang Wang, Wei Wei, Gao Cong, Xiao-Li Li, Xian-Ling Mao, and Minghui Qiu. 2020. Global context enhanced graph neural networks for session-based recommendation. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 169–178.

Wei Wei, Jiayi Liu, Xianling Mao, Guibing Guo, Feida Zhu, Pan Zhou, Yuchong Hu, and Shanshan Feng. 2021. Target-guided emotion-aware chat machine. *ACM Trans. Inf. Syst.*, 39(4):43:1–43:24.

Wenquan Wu, Zhen Guo, Xiangyang Zhou, Hua Wu, Xiyuan Zhang, Rongzhong Lian, and Haifeng Wang. 2019. Proactive human-machine conversation with explicit conversation goal. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3794–3804, Florence, Italy. Association for Computational Linguistics.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Yizhe Zhang, Michel Galley, Jianfeng Gao, Zhe Gan, Xiujun Li, Chris Brockett, and Bill Dolan. 2018. Generating informative and diverse conversational responses via adversarial information maximization. *Advances in Neural Information Processing Systems*, 31.

Peixiang Zhong, Yong Liu, Hao Wang, and Chunyan Miao. 2021. Keyword-guided neural conversational model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 14568–14576.

Kun Zhou, Wayne Xin Zhao, Shuqing Bian, Yuanhang Zhou, Ji-Rong Wen, and Jingsong Yu. 2020a. Improving conversational recommender systems via knowledge graph based semantic fusion. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1006–1014.

Kun Zhou, Yuanhang Zhou, Wayne Xin Zhao, Xiaoke Wang, and Ji-Rong Wen. 2020b. Towards topic-guided conversational recommender system. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4128–4139, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Wangchunshu Zhou, Qifei Li, and Chenle Li. 2021. Learning from perturbations: Diverse and informative dialogue generation with inverse adversarial training. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 694–703, Online. Association for Computational Linguistics.

Ding Zou, Wei Wei, Xian-Ling Mao, Ziyang Wang, Minghui Qiu, Feida Zhu, and Xin Cao. 2022. Multi-level cross-view contrastive learning for knowledge-aware recommender system. In *SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, Madrid, Spain, July 11 - 15, 2022*, pages 1358–1368. ACM.