Modeling Topical Relevance for Multi-Turn Dialogue Generation

Hainan Zhang[2*], **Yanyan Lan**[14†], **Liang Pang**[14], **Hongshen Chen**[2], **Zhuoye Ding**[2] and **Dawei Yin**[4]

[1]CAS Key Lab of Network Data Science and Technology, Institute of Computing Technology, CAS
[2]JD.com, Beijing, China
[3]Baidu.com, Beijing, China
[4]University of Chinese Academy of Sciences
{zhanghainan6,chenhongshen,dingzhuoye}@jd.com,
{lanyanyan,pangliang}@ict.ac.cn,yindawei@acm.org

Abstract

Topic drift is a common phenomenon in multi-turn dialogue. Therefore, an ideal dialogue generation models should be able to capture the topic information of each context, detect the relevant context, and produce appropriate responses accordingly. However, existing models usually use word or sentence level similarities to detect the relevant contexts, which fail to well capture the topical level relevance. In this paper, we propose a new model, named STAR-BTM, to tackle this problem. Firstly, the Biterm Topic Model is pre-trained on the whole training dataset. Then, the topic level attention weights are computed based on the topic representation of each context. Finally, the attention weights and the topic distribution are utilized in the decoding process to generate the corresponding responses. Experimental results on both Chinese customer services data and English Ubuntu dialogue data show that STAR-BTM significantly outperforms several state-of-the-art methods, in terms of both metric-based and human evaluations.

1 Introduction

Multi-turn dialogue generation is widely used in many natural language processing (NLP) applications, such as customer services, mobile assistant and chatbots. Given a conversation history containing several contexts, a dialogue generation model is required to automatically output an appropriate response. Therefore, how to fully understand and utilize these contexts is important for designing a good multi-turn dialogue generation model.

Different from single-turn dialogue generation, people usually model the multi-turn dialogue generation in a hierarchical way. A typical example is the Hierarchical Recurrent Encoder-Decoder (HRED) model [Serban *et al.*, 2016; Sordoni *et al.*, 2015]. In the encoding phase, a recurrent neural network (RNN) based encoder is first used to encode each context as a sentence-level vector, and then a hierarchical RNN is utilized to encode these context vectors to a history representation. In the decoding process, another RNN decoder is conducted to generate the response based on the history representation. The parameters of both encoder and decoder are learned by maximizing the averaged likelihood of the training data. However, the desired response is usually only dependent on some relevant contexts, instead of all the contexts. Recently, some works have been proposed to model the relevant contexts by using some similarity measures. For example, Tian et al. [2017] calculates the cosine similarity of the sentence embedding between the current context and the history contexts as the attention weights, Xing et al. [2018] introduces the word and sentence level attention mechanisms to HRED, and Zhang et al. [2019] utilizes the sentences level self-attention mechanism to detect the relevant contexts. However, these similarities are defined on either word or sentence level, which cannot well tackle the topic drift problem in multi-turn dialogue generation.

context1	你好，在吗？(Hello)
context2	有什么问题我可以帮您呢? (What can I do for you?)
context3	商品降价了，我要申请保价 (The product price has dropped. I want a low-price.)
context4	好的，这边帮您申请，商品已经收到了吧? (Ok, I will apply for you. Have you received the product?)
context5	东西收到了发票不在一起吗? (I have received the product without the invoice together.)
context6	开具电子发票不会随货寄出 (The electronic invoice will not be shipped with the goods.)
current context	是发我邮箱吗？(Is it sent to my email？)
response	是的，请您提供邮箱地址，电子发票24小时寄出。 (Yes, please provide your email address， we will send the electronic invoices in 24 hours.)

Table 1: The example from the customer services dataset. The word color indicates the relevant topic word in the contexts and response, showing the topic-drift phenomenon.

Here we give an example conversation, as shown in Table 1. The contexts are of three different topics. The (context1,context2) pair talks about ‘greeting’, the (context3,context4) pair talks about ‘low-price’, and the (context5,...,response) pair talks about ‘invoice’. In this case, using all the contexts indiscriminately will obviously introduce

[*]This work was done when the first author was a Ph.D student at ICT, CAS.

[†]Corresponding Author.

many noises to the decoding process, which will hurt the performance of the multi-turn dialogue generation model. If we use word level similarities to locate the relevant contexts, the current context and context4 in the example will be associated because 'send' and 'receive' are highly similar words, which is clearly wrong. If we use sentence level similarities to locate the relevant contexts, it may still involve the false relevant context4 into consideration.

We argue that context relevance should be computed at the topic level, to better tackle the topic drift problem in multi-turn dialogue generation. From both linguistic and cognitive perspective, topic is the high level cluster of knowledge, which can describe the relationship of sentences in the context, and has an important role in human dialogue for directing focus of attention. In this paper, we propose a new model, namely STAR-BTM, to model the Short-text Topic-level Attention Relevance with Biterm Topic Model (BTM) [Yan *et al.*, 2013]. Specifically, we first pre-train the BTM model on the whole training data, which split every customer-server pair in the context as a short document. Then, we use the BTM to get each sentence topic distribution and calculate the topic distribution similarity between the current context and each history context as the relevance attention. Finally, we utilize the relevance attention and the topic distribution to conduct the decoding process. The BTM model and the text generation model are jointly learned to improve their performances in this process.

In our experiments, we use two public datasets to evaluate our proposed models, i.e., Chinese customer services and English Ubuntu dialogue corpus. The experimental results show that STAR-BTM generates more informative and suitable responses than traditional HRED models and its attention variants, in terms of both metric-based evaluation and human evaluation. Besides, we have shown the relevant attention words, indicating that STAR-BTM obtains coherent results with human's understanding.

2 Related Work

Recently, multi-turn dialogue generation has gained more attention in both research community and industry, compared with the single-turn dialogue generation [Li *et al.*, 2017; Mou *et al.*, 2017; Zhang *et al.*, 2018a; Zhang *et al.*, 2018b]. One of the reasons is that it is closely related to the real application, such as chatbot and customer service. More importantly, multi-turn dialogue generation needs to consider more information and constraints [Chen *et al.*, 2018; Zhang *et al.*, 2018c; Zhang *et al.*, 2019; Wu *et al.*, 2017; Zhou *et al.*, 2016], which brings more challenges for the researchers in this area. To better model the historical information, Serban et al. [Serban *et al.*, 2016] propose the HRED model, which uses a hierarchical encoder-decoder framework to model all the contexts information. With the widespread use of HRED, more and more variant models have been proposed. For example, Serban et al. [Serban *et al.*, 2017b; Serban *et al.*, 2017a] propose Variable HRED (VHRED) and MrRNN which utilize the latent variables as intermediate states to generate diverse responses.

However, it is unreasonable to use all the contexts indiscriminately for the multi-turn dialogue generation task, since the responses are usually only associated with a portion of the previous contexts. Therefore, some researchers try to use the similarity measure to define the relevance of the context. Tian et al. [Tian *et al.*, 2017] propose a weighted sequence (WSeq) attention model for HRED, which uses the cosine similarity as the attention weight to measure the correlation of the contexts. But this model only uses the unsupervised sentence level representation, which fails to capture some detailed semantic information. Recently, Xing et al. [Xing *et al.*, 2018] introduced the traditional attention mechanism [Bahdanau *et al.*, 2015] into HRED, named hierarchical recurrent attention network (HRAN). In this model, the weight of attention is calculated based on the current state, the sentence level representation and the word level representation. However, the word level attention may introduce some noisy relevant contexts. Shen et al. [Chen *et al.*, 2018] propose to introduce the memory network into the VHRED model, so that the model can remember the context information. Theoretically, it can retrieve some relevant information from the memory in the decoding phase, however, it is not clearly whether and how the system accurately extracts the relevant contexts. Zhang et al. [Zhang *et al.*, 2019] proposed to use the sentence level self-attention to model the long distance dependency of contexts, to detect the relevant context for the multi-turn dialogue generation. Though it has the ability to tackle the position bias problem, the sentence level self-attention is still limited in capturing the topic level relevant contexts.

The motivation of this paper is to detect the topic level attention relevance for multi-turn dialogue generation. It is a more proper way to deal with the topic draft problem, as compared with the traditional word or sentence level methods. Some previous works[Xing *et al.*, 2017; Xing *et al.*, 2018] have been proposed to use topic model in dialogue generation. They maily use the topic model to provide some topic related words for generation, while our work focuses on detecting the topic level relevant contexts.

3 STAR-BTM

In this section, we will describe our Short-text Topic Attention Relevance with Biterm Topic Model (STAR-BTM) in detail, with the architecture shown in Figure 1. STAR-BTM consists of three modules, i.e. , the pre-trained BTM model, topic level attention module and the joint learning decoder. Firstly, we pre-train the BTM model on the whole training data, to obtain the topic word distribution of each context. Secondly, the topic level attention is calculated as the similarity between the topic distributions of the current context and each history context. After that, the attention weights are multiplied with the hierarchical hidden state in HRED to obtain the history representation. Finally, the history representation and the topic distribution of the current context are concatenated to decode the response step by step.

From the architecture, we can see that STAR-BTM introduces the short text topic model into the HRED model, to incorporate the topic level relevant contexts to the decoding process. It is clear that the topic level distribution can provide more specific topic information than only using the word and

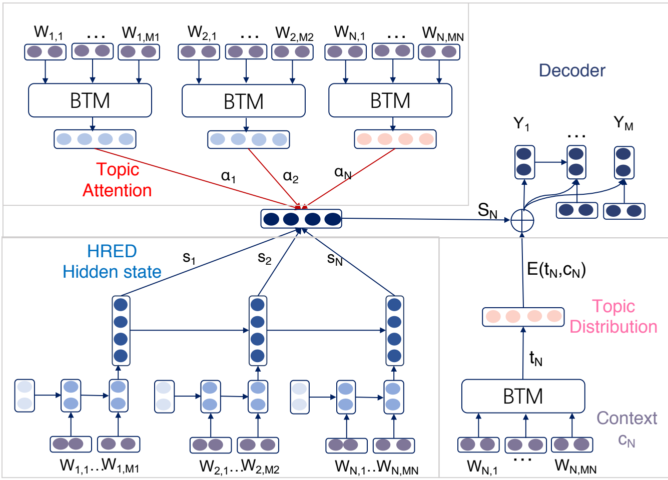

Figure 1: The architecture of STAR-BTM.

sentence level representations. What is more, the topic model firstly 'sees' the whole data globally by the pre-training techniques, and is then fine-tuned by the joint learning technique with the generation model.

3.1 Pre-train BTM Module

We use the pre-trained BTM model on the whole training data to obtain the topic distribution. The pre-trained model on training data can be viewed as the background knowledge, which supplies additional information for the current dialogue session. Like human dialogue in reality, the background knowledge about potential topics will help to detect actual focus of attention model.

BTM [Yan *et al.*, 2013] is a widely used topic model especially designed for short text, which is briefly introduced as follows. For each co-occurrence biterm $b = (w_i, w_j)$ of word w_i and w_j, the joint probability of b is written as:

$$P(b) = \sum_t P(t)P(w_i|t)P(w_j|t),$$

where t stands for a topic.

To infer the topics in a document, BTM assumes that the topic proportions of a document equals to the expectation of the topic proportions of biterms generated from the document. Then we have,

$$P(t|d) = \sum_b P(t|b)P(b|d), \tag{1}$$

where d is a document.

Both $P(t|b)$ and $P(b|d)$ can be calculated via Bayes' formula as follows.

$$P(t|b) = \frac{P(t)P(w_i|t)P(w_j|t)}{\sum_t P(t)P(w_i|t)P(w_j|t)},$$

$$P(b|d) = \frac{n_d(b)}{\sum_b n_d(b)},$$

where $n_d(b)$ is the frequency of the biterm b in the document d. The parameters inference is based on the Gibbs Sampling.

Now we introduce how we apply BTM in our work. Firstly, we split the whole training data $\mathcal{D} = \{(C,Y) = (c_1, \dots, c_N, Y)\}$ to context pairs, i.e. $\mathcal{D} = \{(c_1, c_2), (c_3, c_4), \dots, (c_N, Y)\}$. In the training process, we treat each context pair as one document for BTM. This is reasonable because each pair can be viewed as a single-turn dialogue, and the input and output of a single-turn dialogue are usually of the same topic. After utilizing the Gibbs Sampling, we obtain the word distribution of each topic $P(w_i|t)$ and the topic distribution $P(t)$. In the inferring process, given each sentence c_i in $\mathcal{D}$, the topic of c_i is computed by $P(t_i) = \arg\max_t P(t|c_i)$ in Equation 1.

The BTM model is more suitable for the dialogue generation task than the traditional topic models, such as Latent Dirichlet Allocation(LDA) model. That is because the dialogue has the characteristic of short text with omitted information, which makes LDA not reliable any more. BTM uses word co-occurrence as the core information to determine the topic. So it only depends on the semantic dominance of local co-occurrence information, breaks the document boundary, uses the information of the entire corpus instead of a single document to overcome the sparse problem in short text topic modeling.

3.2 Topic-level Attention Module

We define the context data as $C = \{c_1, \dots, c_N\}$, and each sentence in C is defined as $c_i = \{x_1^{(i)}, \dots, x_M^{(i)}\}$. Given the sentence c_i as input, the RNN model first maps the input sequence c_i to the fixed dimension vector h_M^i as follows:

$$h_k^{(i)} = f(h_{k-1}^{(i)}, w_k^{(i)}),$$

where $w_k^{(i)}$ is the word vector of $x_k^{(i)}$, $h_k^{(i)}$ represents the hidden state vector of the RNN at time k, which combines $w_k^{(i)}$ and $h_{k-1}^{(i)}$. We obtained the state representation set of the contexts $\{h^{(1)}, \dots, h^{(N)}\}$.

Then we use a high-level RNN model to take the context state representation set $\{h^{(1)}, \dots, h^{(N)}\}$ as input, and obtain the high-level context representation vector s_k:

$$s_k = f(s_{k-1}, h^{(k)}),$$

where $h^{(k)}$ is the vector representation of the k-th sentence, and s_k represents the state vector of the high-level RNN at time k, which combines $h^{(k)}$ and s_{k-1}. We obtained the output of the high-level RNN at each step: $\{s_1, \dots, s_N\}$.

Given the context data $C = \{c_1, \dots, c_N\}$, we obtained the topic for each sentence as $T = \{t_1, \dots, t_N\}$ through the pre-trained BTM model. We define attention weights as:

$$\alpha_i = \frac{E(t_i, c_i)E(t_N, c_N)}{|E(t_i, c_i)| \cdot |E(t_N, c_N)|},$$

where $E(t_i, c_i)$ is the sum of the word distribution for topic t_i and the projected word distribution for context c_i, which is defined as the product of the word distribution for topic t_i and the one-hot representation of context c_i.

Finally, we obtain the softmax attention weights α_i' and the context vector S_N as:

$$\alpha_i' = \frac{\alpha_i}{\sum_{j=1}^N \alpha_j}, \quad S_N = \sum_{i=1}^N \alpha_i' \times s_i.$$

3.3 Joint Learning Decoder

We conduct another RNN as the decoder to generate the response $Y = \{y_1, \cdots, y_M\}$. Given the context vector S_N, the topic distribution of the current context D_N and the previously generated word $y_1, \cdots, y_{i-1}$, the decoder predicts the probability of the next word y_i by converting the joint probability into a conditional probability through a chain rule in probability theory. We use the topic distribution of the current context D_N in decoder for the reason that it could supply the topic information to generate more relevant response.

Given a set of training data $\mathcal{D} = \{(C, T, Y)\}$, STAR-BTM assumes that the data is conditionally independent, and samples from the probability P_g, and uses the following negative log likelihood as a minimized objective function:

$$\mathcal{L} = -\sum_{(C,T,Y)\in\mathcal{D}} \log P_g(Y|C, T),$$

where C is the context, T is the topic distribution of C and Y is the real response.

4 Experiment

In this section, we conducted experiments on the Chinese customer service dataset and the English Ubuntu conversation dataset to verify the effectiveness of our proposed method.

4.1 Experimental Settings

We first introduce experimental settings, including datasets, baselines, parameter settings, and evaluation measures.

Datasets

We utilize two public multi-turn dialogue datasets in our experiments, which are widely used in the evaluation of multi-turn dialogue generation task. The Chinese customer service dataset, named JDC, consists of 515,686 history-response pairs published by the JD contest [1]. We randomly divided the corpus into training, validation and testing, each contains 500,000, 7843, and 7843 pairs, respectively. The Ubuntu conversation dataset [2] is extracted from the Ubuntu Q&A forum, called Ubuntu [Lowe *et al.*, 2015]. We utilize the official scripts for tokenizing, stemming and morphing, and remove the duplicates and sentence whose length is less than 5 or greater than 50. Finally, we obtain 3,980,000, 10,000, and 10,000 history-response pairs for training, validation and testing, respectively.

Baseline Methods and Parameter Settings

We used seven baseline methods for comparison, including the traditional Seq2Seq [Sutskever *et al.*, 2014], HRED [Serban *et al.*, 2016], VHRED [Serban *et al.*, 2017b], Weighted Sequence with Concat (WSeq) [Tian *et al.*, 2017], Hierarchical Recurrent Attention Network (HRAN) [Xing *et al.*, 2018], Hierarchical Hidden Variational Memory Network (HVMN) [Chen *et al.*, 2018] and Relevant Context with Self-Attention (ReCoSa) [Zhang *et al.*, 2019]. To fairly compare the topic-level attention model with self-attention model, we extend our STAR-BTM to the ReCoSa scenario, named ReCoSa-BTM, where the topic embedding is concatenated with the sentence representation.

For JDC, the Jieba tool is utilized for Chinese word segmentation, and its vocabulary size is set to 68,521. For Ubuntu, we set the vocabulary size to 15,000. To fairly compare our model with all baselines, the number of hidden nodes is all set to 512 and the batch size set to 32. The max length of sentence is set to 50 and the max number of dialogue turns is set to 15. The number of topics in BTM is set to 8. We use the Adam for gradient optimization in our experiments. The learning rate is set to 0.0001. We run all models on the Tesla K80 GPU with Tensorflow.[3]

Evaluation Measures

We use both quantitative evaluation and human judgements in our experiments. Specifically, we use the traditional indicators, i.e., PPL and BLEU [Xing *et al.*, 2017] to evaluate the quality of generated responses [Chen *et al.*, 2018; Tian *et al.*, 2017; Xing *et al.*, 2018]. And we also use the *distinct* value [Li *et al.*, 2016a; Li *et al.*, 2016b] to evaluate the degree of diversity of generation responses. They are widely used in NLP and multi-turn dialogue generation tasks [Chen *et al.*, 2018; Tian *et al.*, 2017; Xing *et al.*, 2018; Zhang *et al.*, 2018c; Zhang *et al.*, 2018a; Zhang *et al.*, 2018b].

For human evaluation, given the 300 randomly sampled context and its generated responses from all the models, we invited three annotators (all CS majored students) to compare the STAR-BTM model with the baseline methods, e.g. win and loss, based on the coherence of the generated response with respect to the contexts. In particular, the win tag indicates that the response generated by STAR-BTM is more relevant than the baseline model. In order to compare the informativeness of the response generated by the models, we also require the annotators to label the informativeness of each model. If the response generated by STAR-BTM is more informative than the baseline, the annotator will label 1, otherwise label 0.

4.2 Experimental Results

Experimental results on two datasets are demonstrate below.

Metric-based Evaluation

The metric-based evaluation results are shown in Table 2. From the results, we can see that the models which detect the relevant contexts, such as WSeq, HRAN, HVMN and ReCoSa, are superior to the traditional HRED baseline models in terms of BLEU, PPL and *distinct*. This is mainly because these models further consider the attention of the relevant context information rather than all the contexts in the optimization process. HRAN introduces the traditional attention mechanism to learn the important context sentences. HVMN utilizes the memory network to remember the context information. ReCoSa uses the self-attention to detect the relevant contexts. But their effects are limited since they do not consider the topical level relevance. Our proposed STAR-BTM and ReCoSa-BTM have shown good results. Taking the BLEU value on the JDC dataset as an example, the BLEU value of the STAR-BTM and ReCoSa-BTM are 13.386 and

[1] http://jddc.jd.com/

[2] https://github.com/rkadlec/ubuntu-ranking-dataset-creator

[3] https://github.com/zhanghainan/STAR-BTM

JDC Dataset				
Model	PPL	BLEU	distinct-1	distinct-2
SEQ2SEQ	20.287	11.458	1.069	3.587
HRED	21.264	12.087	1.101	3.809
VHRED	22.287	11.501	1.174	3.695
WSeq	21.824	12.529	1.042	3.917
HRAN	20.573	12.278	**1.313**	5.753
HVMN	22.242	13.125	0.878	3.993
STAR-BTM	**20.267**	**13.386**	0.937	**5.816**
ReCoSa	17.282	13.797	1.135	6.590
ReCoSa-BTM	18.432	**13.912**	1.180	**6.739**
Ubuntu Dataset				
Model	PPL	BLEU	distinct-1	distinct-2
SEQ2SEQ	104.899	0.4245	0.808	1.120
HRED	115.008	0.6051	1.045	2.724
VHRED	186.793	0.5229	1.342	2.887
WSeq	141.599	0.9074	1.024	2.878
HRAN	110.278	0.6117	1.399	3.075
HVMN	164.022	0.7549	1.607	3.245
STAR-BTM	**104.893**	**1.3303**	1.601	**4.525**
ReCoSa	96.057	1.6485	1.718	3.768
ReCoSa-BTM	96.124	**1.932**	**1.723**	**4.734**

Table 2: The metric-based evaluation results(%).

JDC Dataset				
model	STAR-BTM vs.			kappa
	win (%)	loss (%)	inform. (%)	
SEQ2SEQ	55.32	2.12	73.79	0.356
HRED	48.93	6.38	70.87	0.383
VHRED	48.94	8.51	69.98	0.392
WSeq	44.68	8.5	66.99	0.378
HRAN	34.04	10.64	60.19	0.401
HVMN	27.66	12.77	61.02	0.379
ReCoSa	25.34	20.71	55.63	0.358
Ubuntu Dataset				
model	STAR-BTM vs.			kappa
	win (%)	loss (%)	inform. (%)	
SEQ2SEQ	51.46	3.88	72.60	0.398
HRED	48.54	6.80	71.23	0.410
VHRED	48.44	6.76	69.18	0.423
WSeq	40.78	6.80	67.80	0.415
HRAN	32.04	11.65	61.16	0.422
HVMN	25.24	13.59	60.27	0.414
ReCoSa	20.14	15.33	56.15	0.409

Table 3: The human evaluation on JDC and Ubuntu.

13.912, which are significantly better than that of HVMN and ReCoSa, i.e., 13.125 and 13.797. The *distinct* value of our model is also higher than other baseline models, indicting that our model can generate more diverse responses. We also conducted a significance test. The results show that the improvement of our model is significant in both Chinese and English datasets with p-value < 0.01. In summary, our STAR-BTM and ReCoSa-BTM model are able to generate higher quality and more diverse responses than the baselines.

Human Evaluation

The results of human evaluation are shown in Table 3. The percentage of win, loss, and informativeness(inform.), as compared with the baseline models, are given to evaluate the quality and the informativeness of the generated responses by STAR-BTM. From the experimental results, the percentage of win is greater than the loss, indicating that our STAR-BTM model is significantly better than the baseline methods. Taking JDC as an example, STAR-BTM obtains a preference gain (i.e., the win ratio minus the loss ratio) of 36.18 %, 23.4 %, 14.89 % and 4.63%, respectively, as compared with WSeq, HRAN, HVMN and ReCoSa. In addition, the percentage of informativeness is more than 50 percent, as compared with WSeq, HRAN, HVMN and ReCoSa, i.e.,66.99%，60.19%, 61.02% and 55.63%, respectively, showing that topic level information is effective for the multi-turn dialogue generation task and our STAR-BTM can generate interesting response with more information. The Kappa [Fleiss, 1971] value demonstrates the consistency of different annotators.

Case Study

To facilitate a better understanding of our model, we present some examples in Table 4, and show the top 10 words of each topic in the Table 5. From the results, we can see that why the topic level attention model performs better than the model only using the word and sentence level representation. Taking the example1 in Table 4 as an example, it easy to generate common responses by using only sentence level representation, such as '*What can I do for you?*' and '*Yes*'. However, our topic level attention model has the ability to generate more relevant and informative responses, such as '*Based on the submitted after-sales service form*' and '*Yes, you need apply after-sales and select lack*'. This is mainly because the topic level attention is able to associate some important information such as '补发(send a new one for a replacement)' and '售后(after-sales)' by topic modeling, which are usually hard to be captured by traditional word or sentence level similarities. These results indicate the advantage of modeling topic level relevance.

We also show the top 10 words of each topic from the BTM model on the two dataset, as shown in Table 5. Take the JDC dataset as an example, from the results, we can see that BTM model distinguishes eight topics, i.e., '配送(shipping), 发票(invoice), 退款(refund), 售后(after-sale), 催单(reminder), 保价(low-price),缺货(out-of-stock) and 感谢(thanks)'. For each topic, the top 10 words represent the core information of the topic. Take the example1 in the Table 4 as an example, since the '补发(send a new one for a replacement)' and '售后(after-sales)' are the 15-th and second word in the same topic 4, respectively, the model can generate 'submitted after-sales service form' based on the topic level attention. In the example2, the current context is about the 'gateway' with topic 'network', so the topic distribution can supply some additional topic information, such as 'restart', 'dhcp' and 'router'. In a word, our STAR-BTM and ReCoSa-BTM model can supply the critical topic information to improve the informativeness of the generated response.

5 Conclusion

In this paper, we propose a new multi-turn dialogue generation model, namely STAR-BTM. The motivation comes from the fact that topic drift is a common phenomenon in multi-turn dialogue. The existing models usually use word or sentence level similarities to detect the relevant contexts, which ignore the topic level relevance. Our core idea is to utilize topic models to detect the relevant context information and generate a suitable response accordingly. Specifically, STAR-BTM first pre-trains a Biterm Topic Model on the whole train-

Example1		
context1	我收到货，一个镜子碎了	I received the goods, a mirror was broken.
context2	可以**申请售后**	You can **apply for after-sale**
context3	可以安排**补发**吗?	Can you **send a new one for a replacement**?
response	您已经**申请售后**了吗?	Have you ever **apply for the after-sale**?
Seq2Seq	有什么能帮到您的吗?	What can I do for you?
HRED	亲爱哒，请您稍等哦，小红人这就帮您去查	Dear, please wait, I will help you check
VHRED	您好，为了更好的为您服务，请您提供下京东的订单号	Hello, in order to better serve you, please provide the order number
WSeq	可以的	Yes
HRAN	这个订单是吗	Is this order?
HVMN	ORDERID 这个是吗?	ORDERID, this is yours?
ReCoSa	可以的，为您申请呢	Yes, I will apply for you!
STAR-BTM	以提交的**售后服务单**为准的	Based on the submitted **after-sales service form**
ReCoSa-BTM	可以的，需要申请售后换货，选择缺件	Yes, you need apply after-sales and select lack

Example2	
context1	I set up a dhcp server on my ubuntu machine . Client connect to it and get a dhcp address . However , those client cann't connect the network
context2	Have you set the default gateway up in the dhcpd config ? Same subnet on all machine ?
context3	Same subnet . How do I configure a default gateway ?
response	Option router 192.168.129.10 in your dhcpd.conf on the dhcp server
Seq2Seq	I don't know
HRED	I don't know what I have to do with the same time
VHRED	I have a lot of UNK
WSeq	I don't know , but I don't know what I do that
HRAN	You can use the ubuntu server
HVMN	Subnet will work
ReCoSa	How about the dhcp server?
STAR-BTM	Try restart it
ReCoSa-BTM	Dhcp server with router

Table 4: The generated responses from the STAR-BTM model on JDC dataset.

Topic	Topic top 10 words in JDC dataset.									
1	订单 order	配送 delivery	请 please	商品 item	时间 time	站点 site	联系 contact	电话 phone	亲爱 dear	地址 address
2	发票 invoice	地址 address	订单 order	修改 modification	电子 electronic	开具 issue	需要 need	电话 phone	号 number	姓名 name
3	工作日 work day	退款 refunds	订单 order	账 accounts	取消 cancellations	申请 applications	支付 payments	成功 successes	商品 goods	请 please
4	申请 apply	售后 after-sale	点击 click	端 end	提交 submit	客户服务 customer service	审核 review	链接 link	返修 rework	补发 replacement
5	订单 order	站点 site	时间 time	日期 date	下单 order	ORDERID	编号 number	催促 urging	信息 information	订单号 order number
6	商品 products	金额 money	保价 low-price	姓名 name	手机 mobile	申请 apply	快照 snapshot	订单 order	查询 inquiries	请 please
7	查询 inquire	帮 help	调货 delivery	问题 problem	处理 deal	缺货 out of stock	订单号 order number	提供 offer	采购 purchase	请 please
8	!	帮到 help	谢谢 thank	支持 support	感谢您 thank you	评价 evaluation	客气 kind	妹子 I	请 please	祝您 wish you
Topic	**Topic top10 words in Ubuntu dataset**									
1	import	each	not	old	noth	would	than	of	thinic	retri
2	cover	adhoc	version	each	retri	alt	benefit	would	ubuntu	apt_preferec
3	from	cover	alt	or	consid	ed	link	we	window	minut
4	run	desktop	cover	kick	distribut	browser	old	show	laptop	ars
5	each	show	instead	from	irc	over	saw	rpm	mockup	out
6	not	libxt-dev	big	a	by	reason	aha	cover	interest	!
7	896	on	system	cover	restart	not	urgent	violat	overst	ping
8	kxb	but	charg	alway	polici	f	_my_	aha	ugh	zealous

Table 5: The top10 words for each topic from the BTM model on JDC dataset.

ing data, and then incorporate the topic level attention weights to the decoding process for generation.We conduct extensive experiments on both Chinese customer services dataset and English Ubuntu dialogue dataset. The experimental results show that our model significantly outperforms existing HRED models and its attention variants. Therefore, we obtain the conclusion that the topic-level information can be useful for improving the quality of multi-turn dialogue generation, by using proper topic model, such as BTM.

In future work, we plan to further investigate the proposed STAR-BTM model. For example, some personal information can be introduced to supply more relevant information for personalized modeling. In addition, some knowledges like concerned entities can be considered in the relevant contexts to further improve the quality of generated response.

Acknowledgments

This work was supported by Beijing Academy of Artificial Intelligence (BAAI) under Grants No. BAAI2019ZD0306, and BAAI2020ZJ0303, the National Natural Science Foundation of China (NSFC) under Grants No. 61722211, 61773362, 61872338, and 61906180, the Youth Innovation Promotion Association CAS under Grants No. 20144310, and 2016102, the National Key R&D Program of China under Grants No. 2016QY02D0405, the Lenovo-CAS Joint Lab Youth Scientist Project, and the Foundation and Frontier Research Key Program of Chongqing Science and Technology Commission (No. cstc2017jcyjBX0059), and the Tencent AI Lab Rhino-Bird Focused Research Program (No. JR202033).

References

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

[Chen *et al.*, 2018] Hongshen Chen, Zhaochun Ren, Jiliang Tang, Yihong Eric Zhao, and Dawei Yin. Hierarchical variational memory network for dialogue generation. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pages 1653–1662, 2018.

[Fleiss, 1971] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *American Psychological Association*, 1971.

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

[Li *et al.*, 2016b] Jiwei Li, Will Monroe, Alan Ritter, and Galley et al. Deep reinforcement learning for dialogue generation. *EMNLP*, 2016.

[Li *et al.*, 2017] Jiwei Li, Will Monroe, Tianlin Shi, Alan Ritter, and Dan Jurafsky. Adversarial learning for neural dialogue generation. *EMNLP*, 2017.

[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. *Computer Science*, 2015.

[Mou *et al.*, 2017] Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. *ACL*, 2017.

[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 *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[Serban *et al.*, 2017a] Iulian Vlad Serban, Tim Klinger, Gerald Tesauro, Kartik Talamadupula, Bowen Zhou, Yoshua Bengio, and Aaron Courville. Multiresolution recurrent neural networks: An application to dialogue response generation. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017.

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

[Sordoni *et al.*, 2015] Alessandro Sordoni, Yoshua Bengio, Hossein Vahabi, Christina Lioma, Jakob Grue Simonsen, and Jian-Yun Nie. A hierarchical recurrent encoder-decoder for generative context-aware query suggestion. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, pages 553–562, 2015.

[Sutskever *et al.*, 2014] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In *NIPS*, pages 3104–3112, 2014.

[Tian *et al.*, 2017] Zhiliang Tian, Rui Yan, Lili Mou, Yiping Song, Yansong Feng, and Dongyan Zhao. 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)*, volume 2, pages 231–236, 2017.

[Wu *et al.*, 2017] Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, 2017.

[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*, pages 3351–3357, 2017.

[Xing *et al.*, 2018] Chen Xing, Yu Wu, Wei Wu, Yalou Huang, and Ming Zhou. Hierarchical recurrent attention network for response generation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[Yan *et al.*, 2013] Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1445–1456. ACM, 2013.

[Zhang *et al.*, 2018a] Hainan Zhang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. Reinforcing coherence for sequence to sequence model in dialogue generation. In *International Joint Conference on Artificial Intelligence*, pages 4567–4573, 2018.

[Zhang *et al.*, 2018b] Hainan Zhang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. Tailored sequence to sequence models to different conversation scenarios. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1479–1488, 2018.

[Zhang *et al.*, 2018c] Weinan Zhang, Yiming Cui, Yifa Wang, Qingfu Zhu, Lingzhi Li, Lianqiang Zhou, and Ting Liu. Context-sensitive generation of open-domain conversational responses. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2437–2447, 2018.

[Zhang *et al.*, 2019] Hainan Zhang, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. Recosa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3721–3730, 2019.

[Zhou *et al.*, 2016] Xiangyang Zhou, Daxiang Dong, Hua Wu, Shiqi Zhao, Dianhai Yu, Hao Tian, Xuan Liu, and Rui Yan. Multi-view response selection for human-computer conversation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 372–381, 2016.