Multi-Turn Dialogue Generation in E-Commerce Platform with the Context of Historical Dialogue

Weisheng Zhang[*1], **Kaisong Song**[*2], **Yangyang Kang**[2], **Zhongqing Wang**[†1],
Changlong Sun[2], **Xiaozhong Liu**[2], **Shoushan Li**[1], **Min Zhang**[1], **Luo Si**[2]
[1]School of Computer Science and Technology, Soochow University, China
[2]Alibaba Group, Hangzhou, Zhejiang, China
wszhang0@stu.suda.edu.cn, {kaisong.sks, yangyang.kangyy,
luo.si}@alibaba-inc.com, {wangzq, lishoushan, minzhang}@
suda.edu.cn, changlong.scl@taobao.com, liu237@indiana.edu

Abstract

As an important research topic, customer service dialogue generation tends to generate generic seller responses by leveraging current dialogue information. In this study, we propose a novel and extensible dialogue generation method by leveraging sellers' historical dialogue information, which can be both accessible and informative. By utilizing innovative historical dialogue representation learning and historical dialogue selection mechanism, the proposed model is capable of detecting most related responses from sellers' historical dialogues, which can further enhance the current dialogue generation quality. Unlike prior dialogue generation efforts, we treat each seller's historical dialogues as a list of Customer-Seller utterance pairs and allow the model to measure their different importance, and copy words directly from most relevant pairs. Extensive experimental results show that the proposed approach can generate high-quality responses that cater to specific sellers' characteristics and exhibit consistent superiority over baselines on a real-world multi-turn customer service dialogue dataset.

Current Dialogue	
Customer	Hello.
Seller	I'm grad to service you, dear.
Seller	What can I do for you?
Customer	1.65 meters tall and weigh 48 kg, which size should I buy?
Seller	In my experience, you may fit the M size.
Customer	Aright, when could you send it off?
Seller	As soon as we can.
Customer	Will you give me some discount?
Seller's Historical Dialogues	
C_1	Hello.
C_2	I'm looking for some help.
S_1	Welcome to our store.
S_2	What can I do for you?
C_3	I see and is there any coupons?
S_3	You can find it on our main page.
S_4	Click the link to get it.
C_4	OK, I find it, thank you.
S_5	I'm grad that I can help you.
Seller's Response	
HRED	I'm sorry not.
Ground Truth	You can open the main page of our store and draw a coupon.

Table 1: The example of customer server dialogue between the Seller (S) and the customer (C) plus the generative results. The above block is the current dialogue context, the middle one is the historical dialogue of the server, and the below one is the generated response.

1 Introduction

Over the past years, online shopping has experienced incredible growth. In e-commerce platforms, e.g., *Amazon* and *Taobao*, brilliant customer service is becoming increasingly important because of significantly reducing the workload of shop sellers. Ideally, sellers should provide high-quality responses to address the personal needs of the customers. However, such cost can be prohibitive for most small businesses, which inspires us to be concerned with the multi-turn dialogue generation task, which is critical in many natural language processing applications, such as customer services, intelligent assistants, and chatbot.

Despite most existing research works on single-turn dialogue generation (Zhao et al., 2019), multi-turn dialogue generation has gained increasing attention from both academia and industry. One reason is that it is more accordant with the real application scenario, such as chatbot and customer services. More importantly, the generation process is more difficult since there are more context information and constraints to consider. Serban et al. (2016) proposed HRED, which uses the hierarchical encoder-decoder framework to model

[*] Both authors contributed equally to this research.
[†] Corresponding Author: Zhongqing Wang.

Findings of the Association for Computational Linguistics: EMNLP 2020, pages 1981–1990
November 16 - 20, 2020. ©2020 Association for Computational Linguistics

all the context sentences. Since then, the HRED based models have been widely used in different multi-turn dialogue generation tasks, and many variants have been proposed. However, the standard HRED can not adapt easily to our customer service scenario well because of two reasons: simply treating all contexts indiscriminately is not proper since the response is only usually related to a few previous contexts; deliberately ignoring dialogue background knowledge is problematic since the response also has a close relationship with specific products, service mode and even seller characteristics. Table 1 illustrates an example in which standard HRED trained on massive data tends to generate generic responses and cannot simulate such unique seller specific responses without using any external knowledge (e.g., $\mathbf{S}_3$).

Recent studies have noticed the problem and focused on generating appropriate seller responses by integrating external information, e.g., product attributes and titles, into single-turn dialogue generation (Zhao et al., 2019; Chen et al., 2019; Gao et al., 2019). However, they are difficult to generalize in reality because of limited materials on hand and different scenarios. Intuitively, sellers' historical dialogues contain richer reply clues, e.g., similar topics or even the same responses happened previously. Ideally, incorporating historical dialogues into our task should further improve response quality. However, such dialogues may be filled with noises or relevant content, which poses a huge challenge to the automatic selection of helpful context. The sellers' historical dialogues mentioned above are multi-turn dialogues pre-selected from the same sellers in our study. In this paper, we propose a novel and extensible Conditional Historical Generation model to generate high-quality seller responses. The main contributions are summarized as below:

- We propose an extensible model which first studies the effectiveness of incorporating historical dialogue contexts into generation.

- We propose a novel dialogue selection mechanism to locate the most relevant historical customer utterances and seller utterances, and then produce their context representations.

- We use a gated strategy to generate the final response by comprehensively considering the different importance of current dialogue and historical dialogues under a hybrid network.

- Empirical results show that our proposed approach outperforms state-of-the-art competitors significantly on a real-world multi-turn customer service dialogue dataset with both automatic and manual evaluation.

2 Related Work

Previous research on multi-turn dialogue generation (Chaudhuri et al., 2018; Zhou et al., 2018; Olabiyi et al., 2018) has drawn a huge amount of attention from academia and industry, which has broader usage scenario than single-turn dialogue generation (Zhang et al., 2018; Li et al., 2017). Serban et al. (2016); Chen et al. (2018); Wu et al. (2016) proposed a hierarchical encoder-decoder framework to model all the context utterances which can better grasp the overall information of the dialogues. However, these models are difficult to generalize, and their results are unsatisfied since responses maybe vary a lot for the same question towards different occasions and speakers.

Recent studies have noticed the problem and try to alleviate it by incorporating helpful external information into response generation, e.g., speakers' emotional information. (Zhang et al., 2019a,b; Wang et al., 2020). Zhao et al. (2019) proposed a review response generation model in the E-commerce platform, which used the reinforcement learning and copy mechanism to fuse external product information, thereby generating informative and diverse responses. Zheng et al. (2019) proposed a dialogue generation model considering personality traits such as age, name, and gender. Meng et al. (2019) proposed RefNet, which used background descriptions about the target dialogue and used a copy mechanism to copy tokens or semantic units. However, all these models are difficult to generalize in reality because of using different materials, which are not always accessible.

Different from previous studies, which either simply ignore or selectively consider limited external information, we propose a novel and extensible model which integrates sellers' historical dialogues into a multi-turn dialogue generating process and avoids interference from background noise. To our best knowledge, this is the first attempt to incorporate helpful historical dialogues into multi-turn customer service dialogue generation.

3 Conditional Historical Generation

Given current dialogue D and its R relevant historical dialogues participated by the same seller, i.e., $H = \{D_1, D_2, ..., D_R\}$, our task aims to generate a high quality response Y based on the current dialogue D and its historical dialogues H. In this section, we propose a novel **C**onditional **H**istorical **G**eneration (**CHG**) model and display its architecture in Figure 1, which consists of four main modules: *Current Dialogue Encoder*, *Historical Dialogue Encoder*, *Response Representation Encoder* and *Context-Response Attention Decoder*.

3.1 Current Dialogue Encoder

Let a dialogue D containing L utterances as $D = [u_1, u_2, ..., u_L]$, where $u_i = [w_{i1}, w_{i2}, ..., w_{il}]$ is the i-th utterance posted by a customer or a seller. The encoder represents the hierarchical information in the dialogue D, which consists of two layers: *Utterance Layer* and *Dialogue Layer*.

Utterance Layer transforms an utterance u_i into a sequence of low-dimensional dense vectors $\boldsymbol{u}_i = [\boldsymbol{e}_{i1}, \boldsymbol{e}_{i2}, ..., \boldsymbol{e}_{il}]$ via a look-up table $\boldsymbol{E} \in \mathbb{R}^{V \times K}$, where V is the vocabulary size and K is the dimension of word embeddings. Each word embedding $\boldsymbol{e}_i$ is then fed into a bidirectional-GRU, and produces hidden state $\boldsymbol{h}_{ij} \in \mathbb{R}^Z$ according to the formula as below:

$$\boldsymbol{h}_{ij} = \left[\overrightarrow{\text{GRU}}(\boldsymbol{e}_{ij}); \overleftarrow{\text{GRU}}(\boldsymbol{e}_{ij})\right], j \in [1, l] \quad (1)$$

Actually, there are various ways to produce utterance representation, and the simplest one is to use the last $\boldsymbol{h}_{il}$ as the final utterance representation $\boldsymbol{u}_i$.

Dialogue Layer can represent the global context in the dialogue via a N-layer Transformer-Block. One critical advantage of the block is that it has the ability to capture long distant dependencies among utterances. Specifically, we first parameterize position embeddings $\{\boldsymbol{c}_i | i \in [1, L]\}$ for all the consisted utterances. The position embeddings are then simply concatenated to the utterance representations $\{\boldsymbol{u}_i | i \in [1, L]\}$. Finally, we obtain a sequence of utterance representations: $\boldsymbol{U} = \left[\overline{\boldsymbol{u}}_1, \overline{\boldsymbol{u}}_2, ..., \overline{\boldsymbol{u}}_L\right]$ and $\overline{\boldsymbol{u}}_i = \boldsymbol{u}_i \oplus \boldsymbol{c}_i$, and "$\oplus$" denotes the element-wise summation operation.

After that, we feed a matrix of n queries $\boldsymbol{Q} \in \mathbb{R}^{n \times d}$, keys $\boldsymbol{K} \in \mathbb{R}^{n \times d}$ and values $\boldsymbol{V} \in \mathbb{R}^{n \times d}$ into the Transformer-Block, the output representation $\boldsymbol{O} \in \mathbb{R}^{n \times d}$ can be represented by the formula:

$$\boldsymbol{O} = \text{Transformer}^N(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}) \quad (2)$$

To obtain the context representation of dialogue D, the Transformer-Block feed the $\boldsymbol{U}$ as queries, keys, and values in equation 2, and finally output the dialogue context representation $\boldsymbol{O}^D$.

3.2 Historical Dialogue Encoder

For the same question initiated by a customer, different sellers may respond differently, depending on various scenarios. It is observed that historical dialogues contain lots of unique seller-specific words which can not be generated easily. This encoder can represent relevant customer questions and seller responses, respectively. It includes two layers: *Utterance Layer* and *Dialogue Selection*.

Utterance Layer: In a historical dialogue, each customer utterance (i.e., question) usually matches one or more seller utterances (i.e., responses). For example, in Figure 2, u_{C_1} is responded by closely followed u_{S_1} and u_{S_2}, u_{C_2} is responded by closely followed u_{S_3} and u_{S_4}. With the same utterance encoder, each customer/seller utterance is represented as $\{\boldsymbol{u}_{C_i}\}/\{\boldsymbol{u}_{S_j}\}$. For any specific $\boldsymbol{u}_{C_i}$, there are N_{C_i} related seller utterances $\{\boldsymbol{u}_{S_j}\}_{N_{C_i}}$. Note that the processing method is similar for multiple historical dialogues via simple concatenation.

Dialogue Selection: Different historical utterances contribute differently to the target response generation. On the one hand, only a few historical customer utterances are semantically similar to the latest customer question. On the other hand, not all the historical seller utterances respond to the historical customer utterances nearby. In Figure 2, we employ a dialogue selection strategy which contains two layers: *customer attention layer* selects relevant customer utterances $\{\boldsymbol{u}_{C_i}\}$ for the customer question $\boldsymbol{u}_i$; *seller attention layer* finds relevant utterances from $\{\boldsymbol{u}_{S_j}\}_{N_{C_i}}$ for each $\boldsymbol{u}_{C_i}$.

(1) Customer Attention Layer: Given the latest customer question $\boldsymbol{u}_L$ in the current dialogue, we use it to find similar customer utterances from historical dialogues. Specifically, we opt for an attention mechanism which is formulated by:

$$\begin{gathered}\mathrm{p}_i^C = \boldsymbol{v}^{\mathrm{T}} \tanh\left(\overline{\boldsymbol{W}}^C \boldsymbol{u}_{C_i} + \boldsymbol{W}^C \boldsymbol{u}_L + \boldsymbol{b}^C\right) \\ \boldsymbol{\alpha}^C = \text{softmax}\left(\mathbf{p}^C\right), \; \boldsymbol{o}^C = \sum_{i=1}^{N_C} \alpha_i^C \boldsymbol{u}_{C_i}\end{gathered} \quad (3)$$

where $\overline{\boldsymbol{W}}^C$, $\boldsymbol{W}^C$, $\boldsymbol{v}$ and $\boldsymbol{b}^C$ are trainable model parameters, α_i^C is the attention weight, $\mathcal{N}_C$ is the number of historical customer questions and $\boldsymbol{o}^C$ is the representation of all the related questions.

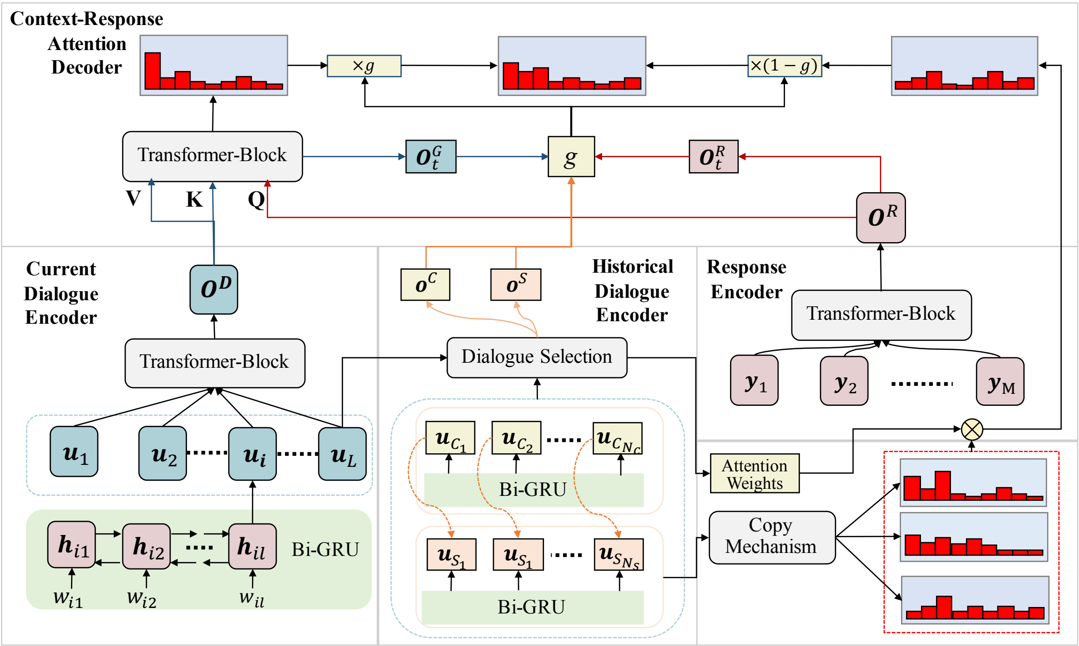

Figure 1: The architecture of our proposed model.

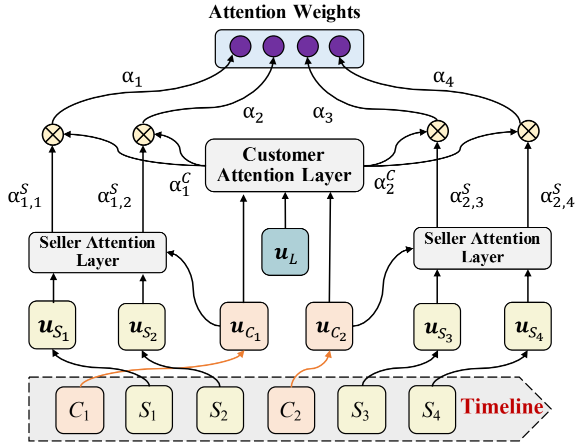

Figure 2: The historical dialogue selection module.

(2) Answer Attention Layer: Given the representation of any historical customer question $\boldsymbol{u}_{C_i}$, we use it to match most relevant answers from the historical dialogue $\{\boldsymbol{u}_{S_j}\}_{N_{C_i}}$. Specifically, we use another attention mechanism to calculate the different importance of seller utterances as below:

$$\begin{aligned} \mathrm{p}_{ij}^S = \overline{\boldsymbol{v}}^{\mathrm{T}} \tanh\left(\overline{\boldsymbol{W}}^S \boldsymbol{u}_{S_j} + \boldsymbol{W}^S \boldsymbol{u}_{C_i} + \boldsymbol{b}^S\right) \\ \boldsymbol{\alpha}_i^S = \mathrm{softmax}\left(\mathrm{p}_i^S\right) \end{aligned} \tag{4}$$

where $\overline{\boldsymbol{W}}^S$, $\boldsymbol{W}^S$, $\overline{\boldsymbol{v}}$ and $\boldsymbol{b}^S$ are learnable model parameters, α_{ij}^S is the attention weight of $\boldsymbol{u}_{S_j}$ read by $\boldsymbol{u}_{C_i}$. In order to obtain the final attention weight for each $\boldsymbol{u}_{S_i}$, we use a cascading attention multiplication operation, which is formulated by:

$$\alpha_j = \alpha_{ij}^S \alpha_i^C, \ \boldsymbol{o}^S = \sum_{j=1}^{N_S} \alpha_j \boldsymbol{u}_{S_j} \tag{5}$$

where α_j is the compound attention weight, N_S is the number of seller utterances and $\boldsymbol{o}^S$ is the representation of all the historical seller's utterances.

3.3 Response Representation Encoder

Given the response $Y = \{y_1, ..., y_M\}$ as the input, the same utterance encoder is used to transform Y into a sequence of low-dimensional dense vectors $\boldsymbol{Y} = \left[\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_M\right]$. Then, We can also parameterize position embeddings $\left\{\boldsymbol{c}_t^Y | t \in [1, M]\right\}$. Another Transformer-Block feed the input $\overline{\boldsymbol{U}} = \left[\overline{\boldsymbol{y}}_1, \overline{\boldsymbol{y}}_2, ..., \overline{\boldsymbol{y}}_M\right]$ and output the response representation $\boldsymbol{O}^R$, where $\overline{\boldsymbol{y}}_t = [\boldsymbol{y}; \boldsymbol{c}_i^Y]$. Note that we also use the mask operator on the response for the training, i.e., we mask $\{y_{t+1}, ..., y_M\}$ and only see $\{y_1, ..., y_{t-1}\}$ if y_t is expected to be generated.

3.4 Context-Response Attention Decoder

The Decoder is a hybrid between a dialogue generation network and a dialogue copy network, as it allows both directly copying words from historical dialogues through copy mechanism and generating words from a fixed vocabulary.

Dialogue Generation: The third Transformer-Block component feeds the output of the Current Dialogue Encoder $\boldsymbol{O}^D$ as keys and values, and the output of the Response Representation Encoder $\boldsymbol{O}^R$ as queries, and finally outputs $\boldsymbol{O}^G$. Then, we utilize a softmax layer to obtain the word probability for the generation process as below:

$$\mathbf{p}^G = \mathrm{softmax}\big(\boldsymbol{W}^G \boldsymbol{O}^G + \boldsymbol{b}^G\big) \tag{6}$$

where $\boldsymbol{W}^G$ and $\boldsymbol{b}^G$ are trainable parameters, $\mathbf{p}^G$ is the probabilities of all the words in the vocabulary.

Dialogue Copy: Inspired by the copy mechanism used in (Vinyals et al., 2015), we allow the decoder to copy words from historical dialogues directly. For each seller utterance u_{S_j}, we use the word vector $\boldsymbol{O}^R_{t-1}$ to find the most important words by attention mechanism. For any word w_i in u_{S_j}, we obtain its attention weight α^w_{ij}. Finally, we sum all the attention weights $\{\alpha^w_{ij}\}$ after multiplying the answer weights $\{\alpha_j\}$ (calculated in Equation 5), and we can obtain the probability of copying any word y_t. The calculation process can be formulated as below:

$$\begin{aligned} \mathrm{p}^w_{ij} = \tilde{\boldsymbol{v}}^{\mathrm{T}}\tanh\big(\overline{\boldsymbol{W}}^R \boldsymbol{O}^R_{t-1} + \boldsymbol{W}^R \boldsymbol{h}_{S_j i} + \boldsymbol{b}^R\big) \\ \boldsymbol{\alpha}^w_j = \mathrm{softmax}(\mathbf{p}^w_j) \\ \mathbf{p}^C_t = \sum_{j=1}^{N_S}\sum_{i=1}^{l}\alpha^w_{ij}\alpha_j \boldsymbol{I}_{w_i=y_t} \end{aligned} \tag{7}$$

where $\overline{\boldsymbol{W}}^R$, $\boldsymbol{W}^R$, $\tilde{\boldsymbol{v}}$ and $\boldsymbol{b}^R$ are trainable model parameters, $\boldsymbol{h}_{S_j i}$ denotes the representation of w_i in u_{S_j}, indicator function $\boldsymbol{I}_{w_i=y_t}$ equals one only when $w_i = y_t$, otherwise zero.

Hybrid Network uses a flexible gated mechanism to decide the degree of copying historical information automatically. Given any word y_t, we combine both $\mathbf{p}^G_t$ and $\mathbf{p}^C_t$ together into the final probability $\mathbf{p}_t$. Note that if the word never appears in any seller utterance, $\mathbf{p}^C_t$ should be zero.

$$\mathbf{p}_t = g\mathbf{p}^G_t + (1-g)\mathbf{p}^C_t \tag{8}$$

where $g \in (0, 1)$ is calculated by the gated mechanism as below:

$$g = \sigma\big(\boldsymbol{W}^G[\boldsymbol{O}^R_t; \boldsymbol{O}^G_t; \boldsymbol{o}^C; \boldsymbol{o}^S] + b^G\big) \tag{9}$$

where $\boldsymbol{W}^G$ and b^G are learnable model parameters, $[;]$ denotes the vector concatenation operation, and $\sigma(\cdot) = \frac{1}{1+e^{-x}}$ is the sigmoid function.

3.5 Training

Our model is optimized in an end-to-end manner. Let θ denote all the model parameters. Given any input $\mathcal{C} = (D, H)$, the log-likelihood of the response $Y = \{y_1, ..., y_M\}$ can be formulated as:

$$\log \mathbf{p}(Y|\mathcal{C};\theta) = \sum_{t=1}^{M}\log \mathbf{p}(y_t|\mathcal{C}, y_1, .., y_{t-1};\theta) \tag{10}$$

We use back propagation to calculate the gradients of all the model parameters, and update them with Adam Optimizer (Kingma and Ba, 2014).

Statistical Results	Num
total number of dialogues	60,000
average length of utterances	27
average length of dialogues	9
average number of historical dialogues	3

Table 2: The detail statistical results of our dataset.

Hyper-parameter	Num	Hyper-parameter	Num
vocab size	3,470	learning rate	1e-4
embedding size	256	dropout rate	0.2
hidden size	512	gradient clipping	10
batch size	32	transformer layer	2
attention head	8		

Table 3: The settings of our model hyper-parameters.

4 Experiments

In this section, we conduct extensive experiments to study the effectiveness of our approach with both automatic and human evaluation metrics.

4.1 Dataset Construction

As far as we know, existing public dialogue datasets do not contain enough sellers' historical dialogues, so we construct a real-world dataset from a top online shopping website in China. Though our experiments are based on a Chinese dataset, our approach can be easily adapted to other languages, such as English and Japanese.

Specifically, we collect 60K multi-turn service dialogues in the clothing domain. For each dialogue, we randomly sample 1-5 latest historical dialogues with the same seller, product, and service topic. According to the statistics, the average utterance number for each dialogue is 9, and each utterance contains 27 Chinese characters on average. We partition the dataset into train/validation/testing set by an 80/10/10 split. The statistical results of our dataset are displayed in Table 2. All the related resources will be publicly available[1].

[1] https://sites.google.com/view/nlp-chg

Model	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	Distinct-1	Distinct-2
Without Historical Dialogues						
Seq2Seq+Att (Sutskever et al., 2014)	30.1	14.8	29.2	11.5	0.018	0.064
HRED (Serban et al., 2016)	32.6	18.8	31.6	16.8	0.017	0.075
ReCoSa (Zhang et al., 2019c)	33.8	20.5	32.9	20.1	0.019	0.099
With Historical Dialogues						
HRED + HD	40.2	27.6	39.4	24.6	0.023	0.108
ReCoSa + HD	41.0	28.1	39.9	25.7	0.026	0.137
CHG (50% HD)	34.1	21.3	33.2	19.8	0.026	0.151
CHG (Our Model)	**41.4**	**29.9**	**40.7**	**30.0**	**0.029**	**0.178**

Table 4: Comparison among different dialogue generation models using various automatic evaluation metrics. HD denotes history dialogues. The best results are highlighted for easier reading.

Model	3	2	1	0	Score
Seq2Seq + Att	8%	59%	22%	11%	1.64
HRED	11%	57%	31%	1%	1.78
ReCoSa	14%	54%	32%	0%	1.86
HRED + HD	21%	56%	22%	1%	1.97
ReCoSa + HD	22%	53%	24%	1%	1.96
CHG	24%	57%	18%	1%	**2.04**

Table 5: Comparison among different dialogue generation models using human evaluation metric.

4.2 Experimental Settings

All the learnable model parameters are initialized by sampling values from a uniform distribution $\mathcal{U}(-0.01, 0.01)$. The hyper-parameters are tuned on the validation set. The best settings of all the hyper-parameters are summarized in Table 3.

To evaluate our approach, we adopt widely used BLEU, ROUGE, and Distinct as automatic evaluation metrics. BLEU (Papineni et al., 2002) is widely used in neural machine translation, which measures word overlap between the generated text and the ground-truth. BLEU score is calculated using the NLTK[2] package, in which the score is an average of BLEU-1˜4. ROUGE[3] (Lin, 2005) is another popular automatic evaluation metric in text summarization. The ROUGE score is obtained through the Rouge package. We report ROUGE-1, ROUGE-2, and ROUGE-L in this work. Distinct is recently proposed by Li et al. (2015), which evaluates the diversity degree of the generated responses by calculating the number of distinct unigrams and bigrams in the generated responses.

All the methods are implemented by ourselves with PyTorch and run on a server configured with a Tesla V100GPU, 2 CPU, and 32G memory.

4.3 Comparison with Baselines

We compare the proposed approach with the following advanced baseline methods, including:

1) Seq2Seq+Att is the standard Seq2Seq model with attention mechanism (Sutskever et al., 2014).

2) HRED uses a hierarchical encoder-decoder framework to model all the context utterances, which has been widely used in different multi-turn dialogue generation tasks (Serban et al., 2016).

3) HRED+HD augments HRED with the historical dialogues. We simply treat the historical dialogues as the context of the current dialogue.

4) ReCoSa uses the self-attention mechanism to measure the relevance between the response and each context, which is the "state-of-the-art" multi-turn dialogue generation model and closely related to our work. (Zhang et al., 2019c)

5) ReCoSa+HD uses the same merge method as that used in "HRED + HD".

Results and Analysis: The results of comparison are reported in Table 4. All the experiments are repeated 10 times, and a t-test proves the improvement of our model is significant (i.e., t <0.005). ReCoSa is the "state-of-the-art" method, which performs better than traditional "Seq2Seq+Att" and "HRED" because of using a self-attention mechanism. However, all these methods can not compete with the methods considering historical information. It is observed that "ReCoSa+HD" and "HRED+HD" achieve further improvements on all the metrics, which proves that their generated responses can be borrowed from sellers' historical dialogue information, which contains product attributes, seller characteristics, and even similar responses. The results illustrate the effectiveness of using historical information.

Our model performs better than "ReCoSa+HD" and "HRED+HD" consistently on all the metrics. This is because the competitors do not especially

[2] http://www.nltk.org/

[3] https://pypi.org/project/rouge/

Configuration	ROUGE-1	ROUGE-2	ROUGE-L	BLEU	Distinct-1	Distinct-2
-(C-S)	41.0	29.2	39.9	28.8	0.027	0.171
-gate	39.0	27.7	38.2	27.6	**0.030**	**0.189**
-copy	40.8	27.9	39.4	28.4	0.026	0.154
CHG	**41.4**	**29.9**	**40.7**	**30.0**	0.029	0.178

Table 6: Comparison among different configured dialogue generation models using automatic evaluation metrics. Notation "-" denotes removing of a specific component used in our model. The best results are highlighted.

Current Dialogue

C$_1$: 你好。(Hi.)
C$_2$: 身高一米八三，体重65千克。(1.83m tall and weight is 65kg.)
S$_1$: 您好，很高兴为您服务。(Hello, i'm happy to serve you.)
S$_2$: 建议您穿2xl码。(I suggest 2xl size.)
C$_3$: 你们提供运费险吗？ (Do you provide freight insurance?)

Historical Dialogue

C$_1$: 这条裙子怎么样。 (I want to know more about this skirt.)
S$_1$: 欢迎来到我们店铺。 (Welcome to our store.)
S$_2$: 麻烦提供一下你的身高和体重，我好给您推荐尺码。 (Please provide your height and weight, and I can recommend you the size.)
C$_2$: 有色差吗。 (Is there any color difference?)
S$_3$: 因为使用不同的显示器，图片和实物看起来会有轻微不同。 (Because different monitors are used, the picture and the actual product may look slightly different.)
S$_4$: 如果到手颜色不对可以随时退货。 (If the color is wrong, you can return it at any time.)
C$_3$: 有运费险吗？ (Is there a freight insurance?)
S$_5$: 我们现在不提供运费险，但是如果您不喜欢，我们会退给您6元运费。 (We don't provide freight insurance now, but if you don't like it, we will refund you 6 CNY freight.)
C$_4$: 好的，我下单了。 (OK, I'll take the order.)
S$_6$: 感谢您的购买。 (Thank you for your purchase.)
S$_7$: 我们会尽快发货的。(We will deliver the goods as soon as possible.)

Responses

ReCoSa: 有运费险的哦。 (Yes, we provide freight insurance.)
ReCoSa+HD: 很抱歉，我们并不提供运费险，如果有需要您可以自行购买。
(I'm sorry not but you can buy it by yourself.)
Our Model: 我们现在不提供运费险，但是如果您想退货，我们会支付6元运费。
(We don't provide freight insurance now, but if you want to return it, we will pay for 6 CNY freight.)
Ground Truth: 现在不提供运费险，如果你不喜欢，我们会承担6元运费退货或者换货。
(Freight insurance is not provided now. If you don't like it, we will pay 6 CNY freight return or exchange.)

Table 7: An example dialogue with generation results. Relevant phrases and words are colored in Red.

model the historical context information, and they are sensitive to irrelevant dialogue noises. Different from theirs, our model uses a dialogue selection module to pinpoint the most relevant responses in historical responses. Meanwhile, our model uses a gated mechanism to balance historical information copying and dialogue generation.

The amount of historical dialogues may influence model performance greatly. Therefore, we build a smaller historical dialogue dataset by halving each seller's historical dialogues. The results show that our model with 50% historical dialogues still performs better than ReCoSa on nearly all the metrics, but slightly worse than our model trained on full historical dialogues. This is reasonable because more historical dialogues will contain more similar responses, and our model is insensitive to the dialogue noises.

4.4 Human Evaluation

We randomly sampled 2,000 dialogues to conduct a manual evaluation and employ three annotators with professional background knowledge to rate the generated responses with 0-3 scores and label each response with the majority score (Zhao et al., 2019). The annotators cannot see the historical dialogues, and only the current dialogue, the model-generated responses, and the ground truth are available for them to make the quality judgments. **Score 0**: *unreadable responses.* **Score 1**: *incorrect or irrelevant responses.* **Score 2**: *partially relevant and correct responses.* **Score**

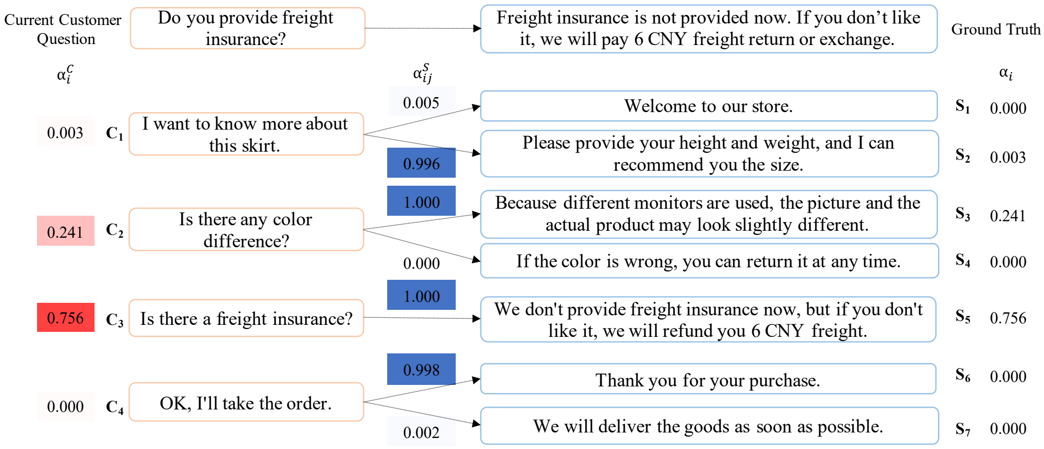
Current Customer Question
Do you provide freight insurance?
Freight insurance is not provided now. If you don't like it, we will pay 6 CNY freight return or exchange.
Ground Truth
α_i^C
α_{ij}^S
α_i
0.003
C1
I want to know more about this skirt.
0.005
0.996
Welcome to our store.
S1 0.000
Please provide your height and weight, and I can recommend you the size.
S2 0.003
0.241
C2
Is there any color difference?
1.000
0.000
Because different monitors are used, the picture and the actual product may look slightly different.
S3 0.241
If the color is wrong, you can return it at any time.
S4 0.000
0.756
C3
Is there a freight insurance?
1.000
We don't provide freight insurance now, but if you don't like it, we will refund you 6 CNY freight.
S5 0.756
0.000
C4
OK, I'll take the order.
0.998
0.002
Thank you for your purchase.
S6 0.000
We will deliver the goods as soon as possible.
S7 0.000

Figure 3: The pairwise interactive representation of the example in case study.

3: *correct and relevant responses.* **Score**: *the weighted sum of all the scores.* The distributions over scores for each model are displayed in Table 5.

From the results in Table 5, we can observe that the models using historical dialogues usually generate more high-quality responses than other competitors ignoring them. Our model obtains the highest weighted score among all the methods. This again proves that using historical dialogues indeed helps to generate high-quality responses, which are more consistent with the sellers' real responses in customer service scenario.

4.5 Ablation Study

Different model configurations may influence model performance greatly. Thus, we conduct an ablation study to validate the effectiveness of each model component used in this work. Table 6 shows the results of the ablation test based on various automatic evaluation metrics. We design several partially configured model variants, including: "**-(C-S)**" means the model doesn't distinguish between speakers and copies from all the historical utterances; "**-gate**" removes the gated mechanism; "**-copy**" removes the copy mechanism.

From Table 6, we can find all the partially configured models can not compete with our fully-configured model, and give in-depth analysis:

-(C-S): Customer and seller usually play different roles in historical dialogues, and seller utterances can provide more response clues compared with customer utterances. Without differentiating, speakers may cause the model to repeat customer questions rather than generate responses.

-copy: We find that the copy mechanism helps a lot in improving the Distinct metrics because it can directly copy some out-of-vocabulary words from the relevant historical dialogues, which tends to produce seller-specific responses rather than generic ones. This naturally achieves better performance on BLEU and ROUGE metrics.

-gate: The generation module and the copy module usually contribute differently to the generation at each time step. This is because the model prefers the generation module than the copy module, which leads to the generation of generic responses rather than a seller-specific response. Without the gating mechanism, P_t^G and P_t^C play equal importance, thus $P_t = \frac{1}{2}P_t^G + \frac{1}{2}P_t^C$.

4.6 Case Study

To compare different models intuitively, we give a multi-turn dialogue example in Table 7, and the original Chinese text has been translated into English text. We compare our approach with ReCoSa ignoring/using historical information and display their generated results. From Table 7, we can find that when asking whether there is freight insurance, ReCoSa generates an inappropriate response (*I'm sorry not, but you can buy it by yourself.*). This is because ReCoSa can not learn seller-specific responses from massive data without considering any external information. Instead, "ReCoSa+HD" and our approach generate much better responses by using external information from the historical dialogue, which contains similar responses to the ground truth. Our approach performs the best because of allowing to copy more response details (e.g., "6 CNY") through our historical dialogue selection strategy.

We also give an example of calculating attention weights of historical seller utterances in Figure 3, where customer utterances are on the left and seller utterances are on the right, the edges denote Customer-Seller interactions, and the attention weights are listed aside. It is observed that S_5 has the largest attention weight through the formula of $0.756{*}1.000 = 0.756$, which again proves the effectiveness of our historical dialogue selection strategy on finding relevant seller responses.

5 Conclusion

In this paper, we propose a novel Conditional Historical Generation model for generating high-quality multi-turn dialogues in E-commerce scenario. Different from previous studies which utilize various external information limited to a specific scenario, our model incorporating historical dialogue information into generation is easy to generalize and applied to practical applications. Specifically, we introduce a novel historical dialogue selection strategy to find appropriate historical seller responses for the latest customer question. Finally, a gated mechanism is used to fuse the results from both the generation module and copy module. The experimental results on a real-world multi-turn dialogue dataset show the effectiveness of our approach.

In the future, we will consider using customer characteristics for generating personalized responses for different customers.

Acknowledgement

This work was supported by Alibaba Group through Alibaba Innovative Research Program. And it was also supported by the the National Natural Science Foundation of China (No. 61806137 and No. 61836007).

References

Debanjan Chaudhuri, Agustinus Kristiadi, Jens Lehmann, and Asja Fischer. 2018. Improving response selection in multi-turn dialogue systems by incorporating domain knowledge. *arXiv preprint arXiv:1809.03194*.

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

Shiqian Chen, Chenliang Li, Feng Ji, Wei Zhou, and Haiqing Chen. 2019. Driven answer generation for product-related questions in e-commerce. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 411–419.

Shen Gao, Zhaochun Ren, Yihong Zhao, Dongyan Zhao, Dawei Yin, and Rui Yan. 2019. Product-aware answer generation in e-commerce question-answering. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, pages 429–437.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*.

C Lin. 2005. Recall-oriented understudy for gisting evaluation (rouge). *Retrieved August*, 20:2005.

Chuan Meng, Pengjie Ren, Zhumin Chen, Christof Monz, Jun Ma, and Maarten de Rijke. 2019. Refnet: A reference-aware network for background based conversation. *arXiv preprint arXiv:1908.06449*.

Oluwatobi Olabiyi, Alan Salimov, Anish Khazane, and Erik T Mueller. 2018. Multi-turn dialogue response generation in an adversarial learning framework. *arXiv preprint arXiv:1805.11752*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

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

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in neural information processing systems*, pages 3104–3112.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in neural information processing systems*, pages 2692–2700.

Jiancheng Wang, Jingjing Wang, Changlong Sun, Shoushan Li, Xiaozhong Liu, Luo Si, Min Zhang, and Guodong Zhou. 2020. Sentiment classification in customer service dialogue with topic-aware multi-task learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9177–9184.

Yu Wu, Wei Wu, Chen Xing, Ming Zhou, and Zhoujun Li. 2016. Sequential matching network: A new architecture for multi-turn response selection in retrieval-based chatbots. *arXiv preprint arXiv:1612.01627*.

Dong Zhang, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019a. Effective sentiment-relevant word selection for multi-modal sentiment analysis in spoken language. In *Proceedings of the 27th ACM International Conference on Multimedia, MM 2019, Nice, France, October 21-25, 2019*, pages 148–156.

Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019b. Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5415–5421.

Hainan Zhang, Yanyan Lan, Jiafeng Guo, Jun Xu, and Xueqi Cheng. 2018. Reinforcing coherence for sequence to sequence model in dialogue generation. In *IJCAI*, pages 4567–4573.

Hainan Zhang, Yanyan Lan, Liang Pang, Jiafeng Guo, and Xueqi Cheng. 2019c. Recosa: Detecting the relevant contexts with self-attention for multi-turn dialogue generation. *arXiv preprint arXiv:1907.05339*.

Lujun Zhao, Kaisong Song, Changlong Sun, Qi Zhang, Xuanjing Huang, and Xiaozhong Liu. 2019. Review response generation in e-commerce platforms with external product information. In *The World Wide Web Conference*, pages 2425–2435.

Yinhe Zheng, Guanyi Chen, Minlie Huang, Song Liu, and Xuan Zhu. 2019. Personalized dialogue generation with diversified traits. *arXiv preprint arXiv:1901.09672*.

Xiangyang Zhou, Lu Li, Daxiang Dong, Yi Liu, Ying Chen, Wayne Xin Zhao, Dianhai Yu, and Hua Wu. 2018. Multi-turn response selection for chatbots with deep attention matching network. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1118–1127.