Towards Robust Online Dialogue Response Generation

Anonymous ACL submission

Abstract

Although pre-trained sequence-to-sequence models have achieved great success in dialogue response generation, chatbots still suffer from generating inconsistent responses in real-world practice, especially in multi-turn settings. We argue that this can be caused by a discrepancy between training and real-world testing. At training time, chatbot generates response with the golden context, while it has to generate based on the context consisting of both user utterances and the model predicted utterances during real-world testing. With the growth of the number of utterances, this discrepancy becomes more serious in the multi-turn settings. In this paper, we propose a hierarchical sampling-based method consisting of both utterance-level sampling and semi-utterance-level sampling, to alleviate the discrepancy, which implicitly increases the dialogue coherence. We further adopt reinforcement learning and re-ranking methods to explicitly optimize the dialogue coherence during training and inference, respectively. Empirical experiments show the effectiveness of the proposed methods for improving the robustness of chatbots in real practice [1].

1 Introduction

Sequence-to-sequence neural models (Vinyals and Le, 2015) serve as a foundation for dialogue response generation (Roller et al., 2020; Zhang et al., 2020b), where typical models adopt the autoregressive framework (Sutskever et al., 2014). During training, models are optimized to maximize the token-level likelihood of the golden response given the golden dialogue history context as input; during inference, the dialogue response generation model is required to predict the response token by token based on the golden multi-turn dialogue context.

With advance in large-scale pre-training (Zhang et al., 2020a; Roller et al., 2020; Lewis et al., 2020)

[1]Codes are attached to the supplementary material and will be publicly available once accepted.

Golden Context
u_1: Ma'am, you forgot your phone.
u_2: Oh, thanks, I couldn't live without this little thing.
u_1: I know what you mean. It is of great significance to you. So did you enjoy your dinner?
Response
r: Oh yes, everything was just perfect. It's so hard to take the whole family out to eat, but your restaurant was perfect.

(a) Training.

Golden Context
u_1: Ma'am, you forgot your phone.
u_2: Oh, thanks, I couldn't live without this little thing.
u_1: I know what you mean. It is of great significance to you. So did you enjoy your dinner?
Prediction
r: I did. I was so happy to have it. It was a great dinner.

(b) Offline Test.

Predicted Context
u : Ma'am, you forgot your phone.
r : I did! I was so mad!
u : Do not worry. Here is your phone.
Prediction
r : I know. I am so mad. I can not even get my phone back.

(c) Online Test.

Figure 1: The illustration of how Blender-bot generates responses in different settings. Blender-bot uses golden context in both training and offline test settings. The blue part indicates the discrepancy utterances in the context of real-world testing (online test). Blender-bot generates an incoherent response in human-bot conversation (Red utterance in Figure 1(c)).

and the availability of high-quality conversational datasets (Li et al., 2017; Dinan et al., 2019b), models are able to generate fluent and informative responses (Shum et al., 2018). On the other hand, despite achieving promising performance on the standard evaluation metrics (e.g., F-1, BLEU, PPL), dialogue response generation models still suffer from unsatisfactory user experience in practice (Welleck et al., 2020; Ram et al., 2018). Previous work shows that chatbots generate repetition (Li et al., 2020a) and contradictory responses (Nie et al., 2021; Li et al., 2021a). One possible reason is that current research focuses on the *offline* evaluation settings, where the golden context is used as input. However, the golden context cannot be accessed in *online* settings. Figure 1(c) shows a human-bot conversation in practice. The golden context in Figure 1(a) and Figure 1(b) is replaced with a system-generated context in Figure 1(c). In this real-world setting, the multi-turn context consists of both previous chatbot generated utterance (r) and human response (u), which is inconsistent with the training settings.

Such utterance-level discrepancy between *offline* training and *online* testing is reminiscent of the exposure bias problem (Bengio et al., 2015; Ranzato et al., 2016). Recent research has made solid strides towards alleviating the exposure bias problem in various generation tasks, such as image captioning (Bengio et al., 2015), speech recognition (Bengio et al., 2015), and neural machine translation (Zhang et al., 2019; Mihaylova and Martins, 2019). They simulate the inference stage by replacing golden target input tokens with the model predictions during training. Intuitively, it can be applied to dialogue generation also. However, the unique challenge in multi-turn dialogue response generation is the existence of both the utterance-level and token-level discrepancy in a hierarchical manner, which is more severe compared to the above tasks. Given the golden context, 93.3% of generated utterances are coherent with the context after 10 turns in our experiments. However, when it comes to the predicted context, the coherence rate drops to less than 30% (Figure 2).

To alleviate the inconsistency between training and real-world testing, we propose both utterance-level and semi-utterance-level sampling-based methods to improve the performance for bot-bot conversation (self-talk). In particular, we sample whole utterances with a scheduled probability and use model generated utterances to replace golden utterances. We schedule our sampling in a hierarchy way. Utterance-level sampling method generates the utterance based on the previous context, which simulates the online-testing scene during training. Semi-utterance-level sampling generates an utterance by using both the previous context and the first few tokens in the sampled utterance, for keeping the semantic similarity between the generated utterance and the golden utterance. To further boost the performance, we adopt reinforcement learning and re-ranking to directly optimize

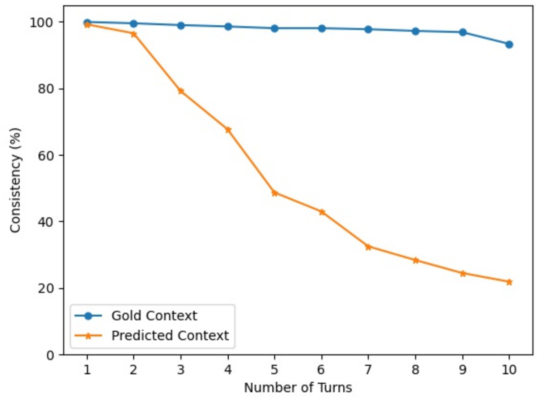


Figure 2: Coherence rate against number of utterances in the context. Coherence rate (Eq 6) measures the percentage of responses is coherence with the corresponding contexts.

the dialogue coherence between the context and the response in the simulated online setting, by consulting an external natural language inference (NLI) based coherence classifier during training and inference, respectively.

We conduct our experiments on Wizard of Wikipedia (Dinan et al., 2019b) and human-bot conversation. Empirical results show that our hierarchical sampling approach improves the abilities of dialogue models on generating coherent and less repetitive responses without introducing external training signals. We further demonstrate that an external coherence classifier can be used in both training and inference to help models produce more coherent responses. Finally, we demonstrate that these methods make chatbots more robust in real-word testing. We release our code and models at `https://anonymous`.

2 Related Work

Alleviating Discrepancy. To bridge the gap between training and inference in auto-regressive models, Bengio et al. (2015) first attempted to randomly sample the previous generated token to replace the ground-truth token during training. Zhang et al. (2019) extended the work of Bengio et al. (2015) by sampling candidates using beam search. Mihaylova and Martins (2019) considered scheduled sampling for transformer-based model. Liu et al. (2021a) and Liu et al. (2021b) further designed sampling strategy based on the model confidence and decode steps, respectively. Xu et al. (2021) introduced scheduled sampling in the one-to-many generation scenario. All these method are designed for mitigating the token-level exposure bias problem. To our knowledge, we are the first to improve the utterance-level discrepancy between training and real-world testing.

Dialogue Coherence. Welleck et al. (2019) modeled dialogue coherence as natural language inference and released the dialogue NLI dataset based on persona (Zhang et al., 2018). Li et al. (2020b) leveraged NLI as supervision to reduce incoherent and repetition response via unlikelihood training. Nie et al. (2021) extended dialogue NLI by releasing a human-written multi-domain dataset. Qin et al. (2021) further introduced dialogue NLI in task-oriented dialogue system. Khandelwal (2021) used reinforcement learning to optimize semantic coherence and consistent flow. Li et al. (2021b) proposed a dynamic flow mechanism to model the context flow. We use coherence as a measure of online dialogue quality. In contrast, existing work all consider the offline setting where the input is a golden history.

3 Definition

3.1 Task

Given a dialogue context $\mathbf{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_{l-1}\}$, where $\mathbf{u}_i = \{\mathbf{x}_1^{\mathbf{u}_i}, \ldots, \mathbf{x}_{|\mathbf{u}_i|}^{\mathbf{u}_i}\}$ represents the i-th utterance. $\mathbf{U}$ can be formed as $\mathbf{U} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\}$ by concatenating all utterances as a token sequence, where $\mathbf{x}_i$ denotes the i-th token in $\mathbf{U}$. The corresponding response can be denoted as $\mathbf{r} = \mathbf{u}_l = \{y_1, y_2, \ldots, y_{T'}\}$. Given a training context-response pair $\{\mathbf{U}, \mathbf{r}\}$, the probability $P(\mathbf{r}|\mathbf{U})$ can be computed by:

$$p(\mathbf{r}|\mathbf{U}) = \prod_{t=1}^{T'} p(y_t|\mathbf{U}, y_{1:t-1}) \quad (1)$$

which can be estimated by a sequence-to-sequence neural network (i.e., transformers) with parameters θ. Our goal is to learn a dialogue generation model $P_\theta(\mathbf{r}|\mathbf{U})$, which is able to generate response $\mathbf{r}$ based on the context $\mathbf{U}$.

3.2 Model

We adopt a standard Transformer (Vaswani et al., 2017) seq2seq model in a dialogue response generation setting.

The dialogue context $\mathbf{U}$ is first fed into the transformer encoder, yielding a sequence of hidden representations.

$$\mathbf{h}^{enc} = \text{TRANSFORMER\_ENCODER}(\mathbf{U}) \quad (2)$$

At the t th step of the decoder, $\mathbf{h}^{enc}$ and the previous output tokens $y_{1:t-1}$ are then used as inputs, yielding an output representation

$$\mathbf{h}_t^{dec} = \text{TRANSFORMER\_DECODER}(\mathbf{h}^{enc}, y_{1:t-1}) \quad (3)$$

The generative probability distribution of y_t is given by a linear projection of the hidden vector $\mathbf{h}_t^{dec}$ followed by a softmax transformation

$$p(y_t|\mathbf{U}, y_{1:t-1}) = softmax(\mathbf{W}^o\mathbf{h}_t^{dec} + \mathbf{b}^o) \quad (4)$$

where $\mathbf{W}^o$ and $\mathbf{b}^o$ are trainable parameters.

The standard cross-entropy loss is used to optimize the parameters θ. Given a training pair $(\mathbf{U}, \mathbf{r})$, the objective is to minimize:

$$\mathcal{L}_{dialogue} = -\sum_{t=1}^{T'} \log p(y_t|\mathbf{U}, y_{1:t-1}) \quad (5)$$

During inference, models auto-regressive generate the response $\hat{\mathbf{r}}$ based on the context $\mathbf{U}$.

3.3 Evaluation

Offline Evaluation. A conventional practice for evaluating dialogue generation model is formed as a lexical similarity task. In particular, the dialogue generation model is first required to generate response $\hat{\mathbf{r}}$ based on the golden dialogue context $\mathbf{U}$. And then the lexical similarity (i.e., F1, BLEU) between the golden response $\mathbf{r}$ and the generated response $\hat{\mathbf{r}}$ is calculated to measure the performance.

Online Evaluation. In real practice, chatbot is used to communicate with human users online. As an example for the l-th turn, the dialogue context consists of both human utterances and chatbot utterances generated in previous turns, formed as $\hat{\mathbf{U}} = \{\mathbf{u}_1, \hat{\mathbf{r}}_2, \mathbf{u}_3, \hat{\mathbf{r}}_4, \ldots, \mathbf{u}_{l-1}\}$, where $\mathbf{u}_i$ represents the i-th user utterances and $\hat{\mathbf{r}}_i$ represents the chatbot prediction based on $\hat{\mathbf{U}}_1^{i-1}$. In this setting, the golden context $\mathbf{U}$ does not exist, because the context has been dynamically generated. An intuitive method for online evaluation is to employ a human to talk with chatbot naturally. However this evaluation method is high-cost (Li et al., 2021a) and relative subjective (Dinan et al., 2019a), which cannot be adopted in large-scale evaluation. Following Deriu et al. (2020), we use bot-bot conversations (self-talk) to simulate human-bot conversation, and conduct a NLI-based classifier $f_c(\hat{\mathbf{U}}, \hat{\mathbf{r}})$ to estimate whether the generated response is in line with the context. In particular, given a prompt utterance $\mathbf{u}_1$, we conduct K turns self-talk conversations, yielding a list of utterances $\hat{\mathbf{U}} = \{\mathbf{u}_1, \hat{\mathbf{r}}_2, \hat{\mathbf{r}}_3, \ldots, \hat{\mathbf{r}}_K\}$. At turn $k \in [1, K]$, the coherence rate c_k is calculated by:

$$c_k = \sum_{i=1}^{D} \frac{\mathbb{1}(f_c(\hat{\mathbf{U}}_1^{i-1}, \hat{\mathbf{r}}_i) = 1)}{D} \tag{6}$$

where D represents the number of instances for evaluation, $\mathbb{1}(\cdot)$ returns 1 if $\cdot$ is true and 0 otherwise.

4 Method

We take sampling-based methods to simulate online consentaneous (Section 4.1), and introduce a reinforcement learning method and a re-ranking method to optimize the dialogue coherence explicitly (Section 4.2).

4.1 Hierarchical Sampling

The main difference between training and inference in real world practice when generating $\hat{\mathbf{r}}$ is whether we use the golden context $\mathbf{U}$ or the predicted context $\hat{\mathbf{U}}$ partly predicted by the model. We address this by introducing the hierarchical sampling to optimize dialogue coherence implicitly.

Utterance Level Sampling. Our utterance-level sampling mechanism is shown in Figure 3. Given a golden context $\mathbf{U}_1^{l-1}$, we sample an utterance $\mathbf{u}_i,\ i \in [1, l-1]$ from a distribution (i.e., geometric distribution $\sim Geo(p)$), which tends to sample previous utterance to be replaced. After obtaining the utterance $\mathbf{u}_i$, we first ask the model to predict the response $\hat{\mathbf{r}}_i$ based on the previous context $\mathbf{U}_1^{\prime i-1}$, and then we use the predicted utterance $\hat{\mathbf{r}}_\mathbf{i}$ to replace the golden utterance $\mathbf{u}_i$ in the golden context $\mathbf{U}_1^{l-1} = \{\mathbf{u}_1, \ldots, \mathbf{u}_i, \ldots, \mathbf{u}_{l-1}\}$, yielding the mixed context $\mathbf{U}_1^{\prime l-1} = \{\mathbf{u}_1, \ldots, \hat{\mathbf{r}}_i, \ldots, \mathbf{u}_{l-1}\}$. Finally, $\mathbf{U}_1^{\prime l-1}$ are fed into the encoder. Accordingly, equation 5 is modified as below:

$$\mathcal{L}_{dialogue} = -\sum_{t=1}^{T'} \log p(y_t | \mathbf{U}_1^{\prime l-1}, y_{1:t-1}) \tag{7}$$

Semi-utterance Level Sampling. Our semi-utterance-level sampling method generates the response based on both the previous context and the first few tokens in the sampled utterance. In particular, after obtaining the sampled utterance $\mathbf{u}_i$, we further keep the first j tokens in $\mathbf{u}_i$ as additional cues to generate $\hat{\mathbf{r}}_i'$. Intuitively, a larger j increase both semantic-level and lexical-level overlap between the $\hat{\mathbf{r}}_i'$ and $\mathbf{u}_i$. A smaller j to simulate more accumulate errors along with the inference steps. The same as utterance level sampling in Section 4.1, $\hat{\mathbf{r}}_i'$ is used to replace $\mathbf{u}_i$.

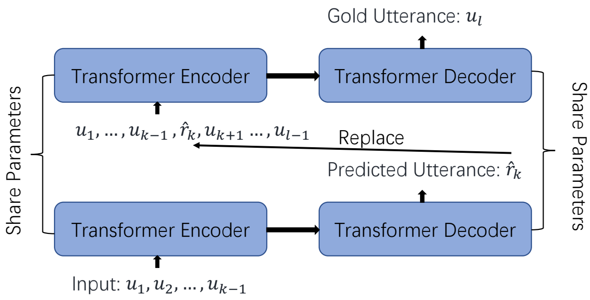


Figure 3: Training with proposed sampling-based methods.

4.2 Explicit Coherence Optimization

Training. We introduce a reinforcement learning method, which explicitly optimizes the coherence between the context and the generated response. We fine-tune the dialogue model P_θ to optimize the reward model P_θ^{RL}.

As shown in Figure 4(a), we first ask the model to generate a response $\hat{\mathbf{r}}$ based on the context $\mathbf{U}$. Then an external coherence classifier f_c is used to justify whether the response is coherent with the context. We adopt the logits of f_c corresponding to the coherent label as the reward. In particular, the input of f_c is a context-response pair $(\mathbf{U}, \mathbf{r})$ and the output is whether the response is coherent with the context. For training f_c, we turn context-response pair $(\mathbf{U}, \mathbf{r})$ to `[CLS]` $\mathbf{U}$ `[SEP]` $\mathbf{r}$ `[SEP]`, and feed it into the RoBERTa model. The hidden state of the `[CLS]` token is used for MLP followed by a softmax scoring function to obtain the coherence score. We train f_c on **D**ialogu**E** **CO**ntradiction **DE**tection (DECODE) (Nie et al., 2021), which is a human annotated corpus labeled with "contradiction (non-coherent)" and "non-contradiction (coherent)". The classifier achieves 94.24 on DECODE dev.

Following Ziegler et al. (2019) and Jaques et al. (2020), we additionally introduce a Kullback–Leibler (KL) divergence term to prevent P_θ^{RL} from drifting too far from P_θ (Figure 4(b)). Formally, given the context $\mathbf{U}$, we calculate the KL-divergence between two models' output probabilities

$$KL(\mathbf{U}) = \sum_{t=1}^{T'} \log \frac{p_\theta^{RL}(\mathbf{x}_t | \mathbf{U}, \mathbf{x}_{1:t-1})}{p_\theta(\mathbf{x}_t | \mathbf{U}, \mathbf{x}_{1:t-1})} \tag{8}$$

$KL(\mathbf{U})$ can be considered as a KL-divergence for the language model task.

Finally, we optimize P_{θ}^{RL} using Proximal Policy Optimization (PPO) (Schulman et al., 2017) with the clipped reward:

$$Reward(\mathbf{U}, \mathbf{r}) = f_c(\mathbf{U}, \hat{\mathbf{r}}) - \beta KL(\mathbf{U}) \quad (9)$$

where β is a hyper-parameter to control the contribution of the KL term. Intuitively, we use the classifier to encourage the model to generate coherent responses, and rely on the KL term to ensure fluency. The inference stage can be the same as the baseline methods in Section 3.2.

Inference with Re-ranking. Another method to enhance dialogue coherence explicitly is inference with re-ranking. In particular, we first adopt beam search to produce multiple candidates responses, and then re-rank the utterances using the coherence classifier f_c. At each turn, the candidate with the highest coherence score is used as the response.

5 Experiments

We train our model based on the golden context - response pair on Wizard of Wikipedia (Dinan et al., 2019b), a chit-chat dialogue benchmark. Two annotators are employed to chat based on an initial topic. The dataset contains 18,430 training dialogues with 1,365 topics.

5.1 Metrics

Following Dinan et al. (2019b) and Kim et al. (2020), the perplexity (PPL) of the ground-truth response, given the golden context as input is taken as one automatic metric. Additionally, coherence rate and non-repetition rate are used as automatic metrics, and human evaluation is conducted.

Coherence Rate. To evaluate online performance in real-world practice, we conduct self-talk to simulate the human-bot conversation, and measure whether the generated response is coherent with the previous context as one automatic metric. The maximum interaction turn is set to 10. As model-based methods have been proved efficient and reliable (Nie et al., 2021; Cui et al., 2021; Li et al., 2021a), and we evaluate the dialogue coherence by consulting f_c in Section 4.2.

Non-Repetition Rate. Inspired by Li et al. (2016), we adopt non-repetition rate to quantify the diversity of the generated sequence during self-talk as a second automatic metric. We calculate

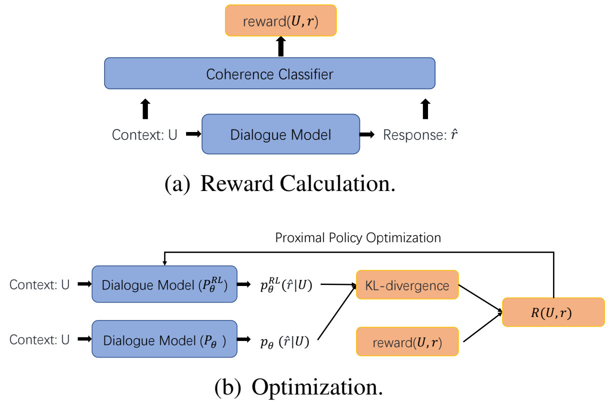


Figure 4: Coherence-Oriented Reinforcement Learning.

distinct-1, distinct-2 and distinct-3 by counting the diversity of uni-grams, bi-grams and tri-grams, respectively. For each context $\hat{\mathbf{U}}$, the distinct-n is calculated by:

$$\text{distinct}-n = \frac{\text{COUNT}(\text{UNIQUE}_{n\text{-gram}_i \in \hat{\mathbf{U}}}(n\text{-gram}_i))}{\text{COUNT}(\text{TOTAL}_{n\text{-gram}_i \in \hat{\mathbf{U}}}(n\text{-gram}_i))} \quad (10)$$

where COUNT(), UNIQUE() and TOTAL() denote count the item of a list, unique items in a list and enumeration a list, respectively. A higher distinct-n indicates a lower repetition rate during self-talk.

Human Evaluation. Following previous work (Ritter et al., 2011), we conduct human evaluation on self-talk to compare our hierarchical sampling-based methods with our baseline multi-turn BART by randomly sampling 50 instances (including 500 utterances). Following Wu et al. (2018), we employ three annotators to do a side-by-side human evaluation.

In order to pursue more authentic evaluation in real practice, we further adopt a human-bot conversation to online evaluate these two methods. In particular, given a prompt utterance, we ask an annotator to chat with chatbot 10 turns. The final human-bot test set we derive contains 50 dialogues (including 500 utterances) for each model. We define three metrics for human evaluation, including fluency, non-repetitive and coherence. Each aspect is scored into three grades (0, 1 and 2) representing "bad", "normal" and "good", respectively. We further calculate the Pearson correlation between the human annotated coherence rate and the model assigned coherence rate.

	Online Evaluation												Offline
	c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9	c_{10}	avg_5	avg_10	PPL
BART w/ Golden context	99.7	98.9	98.2	96.0	97.6	97.2	96.0	94.2	94.1	93.3	99.0	96.5	-
Single-turn BART	99.2	88.1	71.5	63.5	57.2	53.0	46.7	41.8	37.3	34.9	75.9	59.3	21.3
Multi-turn BART	99.2	96.5	79.2	67.7	48.7	43.0	32.5	28.4	24.5	21.9	78.3	54.2	17.8
w/ Utterance	98.4	97.0	89.3	76.7	71.6	59.1	**60.5**	**45.7**	**49.8**	**35.6**	86.6	**68.4**	17.2
w/ Semi-Utterance	98.1	97.2	85.7	69.2	64.0	50.5	52.1	36.4	43.6	29.1	82.9	62.6	**17.1**
w/ Hierarchical	**99.2**	**97.6**	**91.2**	**78.5**	**72.3**	**60.7**	57.8	45.5	44.3	33.0	**87.8**	68.0	17.4

Table 1: Test performance of self-talk given a prompt utterance on Wizard test set.

5.2 Baselines

We compare the proposed methods with the following BART-based baselines:

BART w/ Golden context. We fine-tune BART on the Wizard training set. During inference at turn k, the golden context $\mathbf{U}_1^{k-1}$ is used to produce the response $\hat{\mathbf{r}}_k$. Because the golden context is unavailable in practice, the performance can be considered as the ceiling performance for alleviating the discrepancy between training and real-world testing.

Multi-turn BART. During training, we fine-tune BART based on the golden context-response pair. Different from BART w/ Golden context, we use the context $\hat{\mathbf{U}}_1^{k-1}$ predicted by previous turns to generate the response $\hat{\mathbf{r}}_k$ during inference.

Single-turn BART. We fine-tune BART for the dialogue generation following the single-turn setting (Wang et al., 2013). Only the last predicted utterance $\hat{\mathbf{r}}_{k-1}$ is fed to the encoder to generate $\hat{\mathbf{r}}_\mathbf{k}$ for both training and inference. Single-turn BART ignores the history in previous utterances.

5.3 Results

Table 1 reports the performance of coherence rate as well as PPL for various methods, and Table 2 shows the distinct-n for the predicted context generated by these methods.

Predicted Context vs Golden Context. We first compare whether the dialogue generation model is able to generate coherence response based on the golden context and the predicted context. As shown on the top of Table 1, the coherence rate of BART w/ Golden context does not decrease significantly with the number of turns increasing. The performance drops by only 5.6 points coherence rate from 2 turns to 10 turns. However, given the predicted context, the coherence rate decreases sharply as the number of turns increase, with only 21.9 c_{10}. This shows the severity of the discrepancy problem in real-world multi-turn dialogue generation.

Model	Dis-1	Dis-2	Dis-3
Multi-turn BART	24.37	32.30	36.35
w/ Hierarchical sampling	36.29	49.77	55.29

Table 2: Non-Repetition Rate (%) for n-gram. 'Dis-n' means 'Distinct-n'.

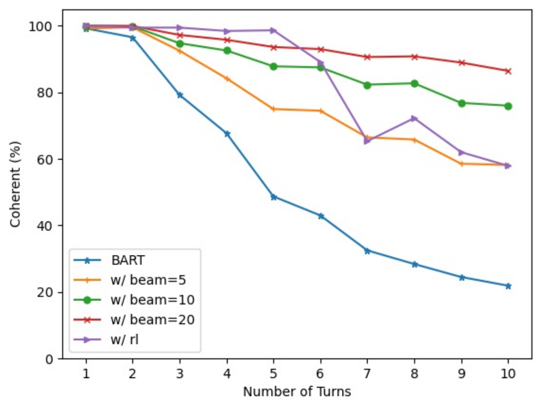
Single-turn vs Multi-turn. In *offline* evaluation, multi-turn BART achieves 17.8 PPL, which significantly outperforms single-turn BART. This indicates that context information is important for response generation. However, we have mixed results in *online* evaluation. For example, multi-turn BART outperforms single-turn BART when the number of utterances in the context is less than four in Table 1. When the number of utterances becomes larger, single-turn BART surprisingly gives better results compared with multi-turn BART. The reason can be that the mismatch between the golden context and the predicted context hinders the model performance as the number of utterances grows for multi-turn model.

Sampling vs w/o Sampling. In Table 1, the proposed sampling-based approach performs slightly better on PPL compared to the multi-turn BART, which shows our methods also work well in general offline settings. When it comes to online settings, our sampling-based methods outperform multi-turn BART significantly in all metrics, although there is no direct supervision signal on coherence. For example, when measured in context corresponding to 5 turns, multi-turn BART w/ hierarchical sampling gives a c_5 of 72.3%, as compared to 48.7% by multi-turn BART.

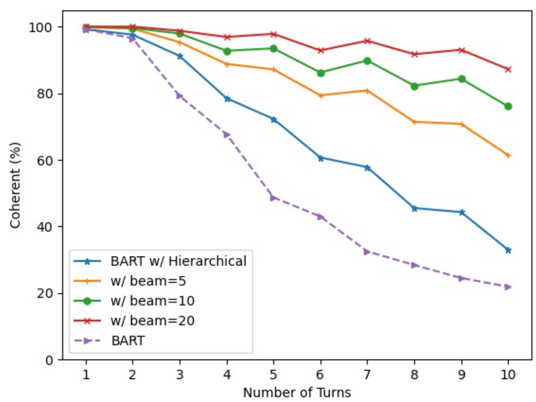
Utterance vs Hierarchical. In Table 1, semi-utterance level sampling underperforms utterance-level sampling in online evaluation. This is because semi-utterance level sampling cannot accurately simulate errors of the inference scene during

Model	Fluency	Rep	Coh
Self-talk			
Multi-turn BART	**1.93**	0.89	0.74
w/ Hierarchical sampling	1.91	**1.37**	**1.45**
Human-bot Conversation			
Multi-turn BART	1.89	0.96	0.63
w/ Hierarchical sampling	**1.90**	**1.53**	**1.32**

Table 3: Human Evaluation. 'Rep' and 'Coh' indicate non-repetition and coherence, respectively.



(a) Multi-turn BART.



(b) Multi-turn BART w/ Hierarchical Sampling.

Figure 5: Coherence rate with explicit optimization.

training. For instance, the dialogue model tends to generate the response beginning with the word "*I*". While semi-utterance level sampling keeps the first few tokens in the sampled utterance. When integrating utterance-level and semi-utterance level sampling, hierarchical sampling gives the best coherence rate when context less than six turns, which achieves 87.8% on avg_5. This shows the effectiveness of sampling in a hierarchy way, which simulates the errors on both utterance-level and token-level.

Repetition. Table 2 reports the non-repetition rate of our sampling-based methods, drawing multi-turn BART as a reference. We find that our methods give higher distinct-n measured by uni-gram, bi-gram and tri-gram, which shows the effect of introducing hierarchical sampling to reduce copying and repetition in model generated context. This also provides support for the effectiveness of sampling-based methods to increase the robustness of multi-turn models.

Human Evaluation. Table 3 compares the hierarchical sampling-based method with multi-turn BART using human evaluation. All models are able to produce fluent responses due to the power of pre-training, where fluency exceeds 1.89 for all models. Measured in non-repetition and coherence, our hierarchical sampling method significantly outperforms the baselines (p<0.01) on both self-talk and human-bot conversation. In human-bot conversation, the coherence increases largely from 0.96 to 1.53, showing that sampling enhances the robustness of online multi-turn conversation. For self-talk, the pearson correlation between the human annotated and the model assigned coherence rate is 0.78, which also demonstrates the effectiveness of the model-based evaluation methods.

Explicit Objectives. Figure 5 shows the effect of the explicit coherence optimization method. Training model with reinforcement learning outperforms with MLE measured by coherence rate, showing the usefulness of optimizing the dialogue coherence directly. We also find that the coherence rate improves significantly after re-ranking in the inference scene for both multi-turn BART and multi-turn BART w/ hierarchical sampling. Furthermore, as the number of candidate utterances increases, the coherence rate increases. Multi-turn BART w/ beam=20 even achieves 86.42 c_{10} compared with 21.9 c_{10} for multi-turn BART. This indicates that the dialogue model can give coherent response candidates, which can be re-ranked by an external coherence classifier to produce a coherent response. Our hierarchical sampling-based methods also consistently perform better than multi-turn BART by introducing coherence re-ranking.

6 Analysis

6.1 The Number of Golden Turns

We investigate whether a larger number of golden turns at the start is able to help model to produce

$\mathbf{u}_1$(Prompt)	My favorite video game is Quake. Have you ever played it?
	Multi-turn BART
$\hat{\mathbf{r}}_2$	I have not played it, but I know it was developed by the Quake team.
$\hat{\mathbf{r}}_3$	Oh, Quake was developed by the Quake team. It's a great game!
$\hat{\mathbf{r}}_4$	I know it was developed by the Quake team. It was the first video game to be released commercially.
	Multi-turn BART w/ Hierarchical sampling
$\hat{\mathbf{r}}_2$	I have not played it, but I have heard it is a very good game.
$\hat{\mathbf{r}}_3$	Yes it is. It was developed by the Quake team. It's a team-based game.
$\hat{\mathbf{r}}_4$	That sounds like a fun game to play with friends. What other games do you like?

Table 4: Examples of generated responses given a prompt utterance on the Wizard of Wikipedia Test Seen.

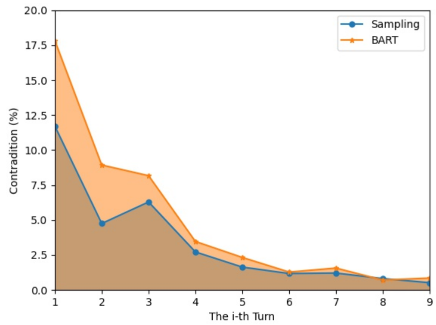


Figure 6: Contradiction rate across different turn. Contradiction rate defined by (1 − coherence rate) ×100%.

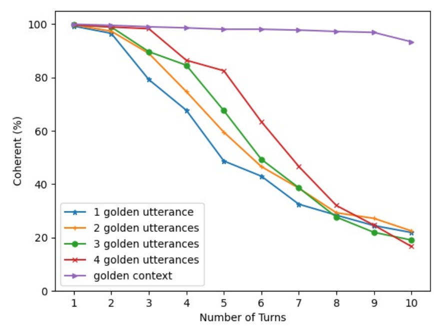


Figure 7: Coherence rate across the number of golden utterances at the beginning.

more coherent responses during inference. Figure 7 shows the coherence rate against the number of golden utterances at the beginning during the self-talk, drawing using the golden context as a reference. It can be seen that a larger number of golden utterance at the beginning yields a larger coherence rate in the first few turns. However, the coherence rate decreases sharply with the number of turns increasing, which shows that simply increasing beginning golden turns cannot help to alleviate the discrepancy between training and real-world testing.

6.2 Utterance-level Contradiction

To understand which turns in the context leads to an incoherence response, we introduce an utterance-based classifier to probe different utterances during generating the response at 10-th turn in self-talk. As shown in Figure 6, both models tend to generate response that contradict with the early turns. This shows that current models do not take full advantage of the long-range dialogue context. Compared with the multi-turn BART, the proposed sampling-based methods significantly decrease the contradiction rate in the early turns, and achieves the similar results in the later turns, which shows our hierarchical sampling-based methods are able to improve robustness of multi-turn models by alleviating the error accumulation.

6.3 Case Study

We present an example to better understanding of multi-turn BART and our model in Table 4. We observe that both models are able to generate reasonable response $\hat{\mathbf{r}_2}$. Because the context for generating $\hat{\mathbf{r}_2}$ contains prompt utterance (golden context) $\mathbf{u}_1$ only. However, when the model encounters the predicted utterance as context, multi-turn BART tends to generate response with repetition and contradiction. With hierarchical sampling, our model produces coherence responses during self-talk.

7 Conclusion

We quantified online dialogue generation in practice, and proposed the hierarchical sampling-based methods to alleviate the discrepancy between training and real-world testing. We further introduce an external coherence classifier on both training and inference to boost the performance. Experiments demonstrate the effectiveness of our methods for generating robust online response on both self-talk and human-bot conversation.

References

Samy Bengio, Oriol Vinyals, Navdeep Jaitly, and Noam Shazeer. 2015. Scheduled sampling for sequence prediction with recurrent neural networks. In *Proceedings of the 28th International Conference on Neural Information Processing Systems - Volume 1*, NIPS'15, page 1171–1179, Cambridge, MA, USA. MIT Press.

Leyang Cui, Yu Wu, Shujie Liu, and Yue Zhang. 2021. Knowledge enhanced fine-tuning for better handling unseen entities in dialogue generation. In *EMNLP*.

Jan Deriu, Don Tuggener, Pius von Däniken, Jon Ander Campos, Alvaro Rodrigo, Thiziri Belkacem, Aitor Soroa, Eneko Agirre, and Mark Cieliebak. 2020. Spot the bot: A robust and efficient framework for the evaluation of conversational dialogue systems. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3971–3984, Online. Association for Computational Linguistics.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, and et al. 2019a. The second conversational intelligence challenge (convai2). *The Springer Series on Challenges in Machine Learning*, page 187–208.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019b. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. 2020. Way off-policy batch deep reinforcement learning of human preferences in dialog.

Anant Khandelwal. 2021. WeaSuL: Weakly supervised dialogue policy learning: Reward estimation for multi-turn dialogue. In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*, pages 69–80, Online. Association for Computational Linguistics.

Byeongchang Kim, Jaewoo Ahn, and Gunhee Kim. 2020. Sequential Latent Knowledge Selection for Knowledge-Grounded Dialogue. In *ICLR*.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119, San Diego, California. Association for Computational Linguistics.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020a. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728, Online. Association for Computational Linguistics.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. 2020b. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728, Online. Association for Computational Linguistics.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021a. Addressing inquiries about history: An efficient and practical framework for evaluating open-domain chatbot consistency. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1057–1067, Online. Association for Computational Linguistics.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 2021b. Conversations are not flat: Modeling the dynamic information flow across dialogue utterances. 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 128–138, Online. Association for Computational Linguistics.

Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021a. Confidence-aware scheduled sampling for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2327–2337, Online. Association for Computational Linguistics.

Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 2021b. Scheduled sampling based on decoding steps for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Online.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *International Conference on Learning Representations*.

Tsvetomila Mihaylova and André F. T. Martins. 2019. Scheduled sampling for transformers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 351–356, Florence, Italy. Association for Computational Linguistics.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 2021. I like fish, especially dolphins: Addressing contradictions in dialogue modeling. 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 1699–1713, Online. Association for Computational Linguistics.

Libo Qin, Tianbao Xie, Shijue Huang, Qiguang Chen, Xiao Xu, and Wanxiang Che. 2021. Don't be contradicted with anything! CI-ToD: Towards benchmarking consistency for task-oriented dialogue system. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2357–2367, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. 2018. Conversational ai: The science behind the alexa prize.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. 2016. Sequence level training with recurrent neural networks. In *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*.

Alan Ritter, Colin Cherry, and William B. Dolan. 2011. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 583–593, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

Heung-yeung Shum, Xiao-dong He, and Di Li. 2018. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1):10–26.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*, volume 30, pages 5998–6008. Curran Associates, Inc.

Oriol Vinyals and Quoc Le. 2015. A neural conversational model.

Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. 2013. A dataset for research on short-text conversations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 935–945, Seattle, Washington, USA. Association for Computational Linguistics.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. 2020. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. 2019. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3731–3741, Florence, Italy. Association for Computational Linguistics.

Yu Wu, Furu Wei, Shaohan Huang, Yunli Wang, Zhoujun Li, and Ming Zhou. 2018. Response generation by context-aware prototype editing.

Haoran Xu, Hainan Zhang, Yanyan Zou, Hongshen Chen, Zhuoye Ding, and Yanyan Lan. 2021. Adaptive bridge between training and inference for dialogue.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2204–2213, Melbourne, Australia. Association for Computational Linguistics.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. 2019. Bridging the gap between training and inference for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4334–4343, Florence, Italy. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020a. Dialogpt: Large-scale generative pre-training for conversational response generation.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020b. Dialogpt: Large-scale generative pre-training for conversational response generation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. 2019. Fine-tuning language models from human preferences.

A Appendix

A.1 Setup

We implement our methods with `transformers` and choose `bart-base` as the pre-trained transformer language model. AdamW (Loshchilov and Hutter, 2019) with a batch size of 32 is used to optimize parameters. The initial learning is set as 5e-5, which will be halved in each training iteration. Following Lewis et al. (2020), we set the maximum input tokens as 512. For the coherence-oriented reinforcement learning method, we set β in Equation 9 as 0.2. For computational efficiency, we truncate the maximum decode length as 20 to calculate the KL-divergence.