TOWARDS ROBUST ONLINE DIALOGUE RESPONSE GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

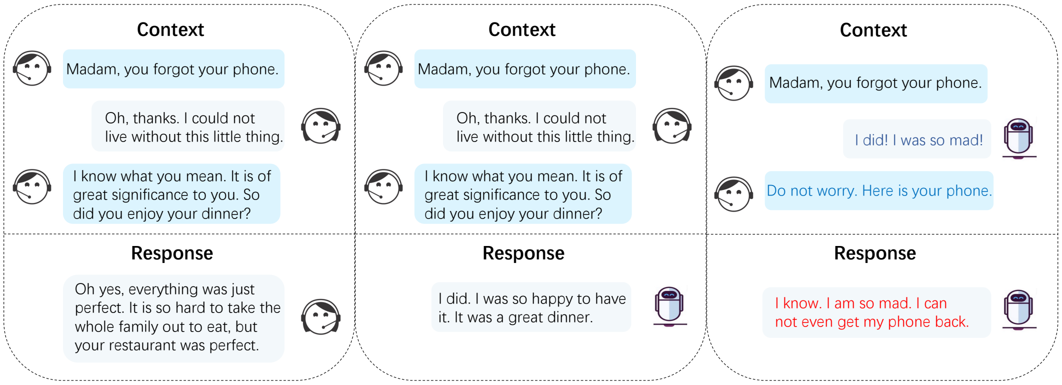
Although pre-trained sequence-to-sequence models have achieved great success in dialogue response generation, chatbots still suffer from generating inconsistent responses in real-world applications, especially in multi-turn settings. We argue that this can be caused by a discrepancy between training and real-world testing. While the chatbot generates the response based on the gold context during training, it has to predict the next utterance based on the context consisting of both the user's and the bot's own utterances in real-world testing. With the growing number of utterances, this discrepancy becomes more severe 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 increases the dialogue coherence implicitly. 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 INTRODUCTION

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

With the advance of large-scale pre-training (Zhang et al., 2020a; Roller et al., 2020; Lewis et al., 2020) and high-quality conversational datasets (Li et al., 2017; Dinan et al., 2019b), dialogue response generation models are able to generate fluent and informative responses (Shum et al., 2018). However, despite achieving promising performance on the standard evaluation metrics (e.g., F-1, BLEU, PPL), models still suffer from unsatisfactory user experience in practice (Welleck et al., 2020; Ram et al., 2018). Previous work shows that chatbots give 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 gold context is used as input for each dialogue turn. However, *the gold context cannot be accessed in online settings*, where inappropriate outputs in one turn affect subsequent turns. Figure 1 (c) shows a human-bot conversation in practice. The gold 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 the previous chatbot generated utterance (r) and the human response (u), which is inconsistent with the training settings.

Such utterance-level discrepancy between *offline* training and *online* testing is analogous to the exposure bias problem (Bengio et al., 2015; Ranzato et al., 2016). Researchers alleviate 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 & Martins, 2019), which simulates the inference stage where gold target input tokens are replaced by the model predictions during training. However, the unique challenge in dialogue response generation is the existence of both the utterance-level and token-level discrepancy in a hierarchical manner under the multi-turn settings, which is more severe compared to above generation tasks.



Madam, you forgot your phone.

Oh, thanks. I could not live without this little thing.

I know what you mean. It is of great significance to you. So did you enjoy your dinner?

Oh yes, everything was just perfect. It is so hard to take the whole family out to eat, but your restaurant was perfect.

(a) Online training.

Madam, you forgot your phone.

Oh, thanks. I could not live without this little thing.

I know what you mean. It is of great significance to you. So did you enjoy your dinner?

I did. I was so happy to have it. It was a great dinner.

(b) Offline test.

Madam, you forgot your phone.

I did! I was so mad!

Do not worry. Here is your phone.

I know. I am so mad. I can not even get my phone back.

(c) Online test.

Figure 1: Illustration of how Blender-bot Roller et al. (2020) generates responses in different settings. The gold context in Figure 1 (a) is sampled from MuTual (Cui et al., 2020). Blender-bot uses gold 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)).

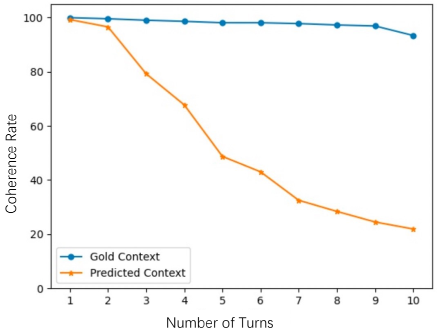


Figure 2: We fine-tune BART-large on Wizard of Wikipedia (Dinan et al., 2019b) and report the coherence rate against number of utterances on test set. Coherence rate (Equation 2) measures the percentage of responses is coherence with the corresponding contexts.

In the real-world scene, the nature of utterance predictions quickly accumulates errors along with the number of turns increased, which yields lower coherence rates for a longer context. Our experiment (Figure 2) reveals that when given the gold context, 93.3% of generated utterances are coherent with the context after 10 turns (Offline test). Whereas only less than 30% of generated utterances are coherent when given the predicted context (Online test).

To alleviate the inconsistency between offline training and online testing, we propose both utterance-level and semi-utterance-level sampling-based methods to improve the performance of the online setting. In particular, we sample the whole utterances with a scheduled probability and use model-generated utterances to replace gold utterances. We schedule our sampling in a hierarchical way. The Utterance-level sampling method generates the utterance based on the previous context to simulate the online-testing scenario during training. Semi-utterance-level sampling method generates the utterance by using both the previous context and the first few tokens in the sampled utterance, to keep semantic similarity between the generated utterance and the gold utterance. To further boost the performance, we adopt reinforcement learning and re-ranking to directly optimize 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), DailyDialogue (Li et al., 2017), 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 find 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-world testing. We release our code and models at `https://anonymous`.

2 Related Work

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

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

3 Settings

3.1 Task

Given a dialogue context $\mathbf{U} = \{\mathbf{u}_1, \ldots, \mathbf{u}_{l-1}\}$, where $\mathbf{u}_i = \{x_1^{\mathbf{u}_i}, \ldots, x_{|\mathbf{u}_i|}^{\mathbf{u}_i}\}$ represents the i-th utterance. $\mathbf{U}$ can also be formed as $\mathbf{U} = \{x_1, \ldots, x_T\}$ by concatenating all utterances as a flat token sequence, where 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 estimated by:

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

which can be calculated by a sequence-to-sequence neural network (i.e., transformers) with parameters θ. Our goal is to learn the 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 Evaluation

Offline Evaluation. A conventional practice Li et al. (2017); Dinan et al. (2019b) for evaluating dialogue generation 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 gold dialogue context $\mathbf{U}$. And then the lexical similarity (i.e., F1, BLEU) between the gold response $\mathbf{r}$ and the generated response $\hat{\mathbf{r}}$ is calculated to measure the performance.

Online Evaluation. In practice, a chatbot is used to communicate with human users online. Taking the l-th turn human-bot conversation as an example, 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 the previous context $\hat{\mathbf{U}}_1^{i-1}$. In this setting, the gold 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.

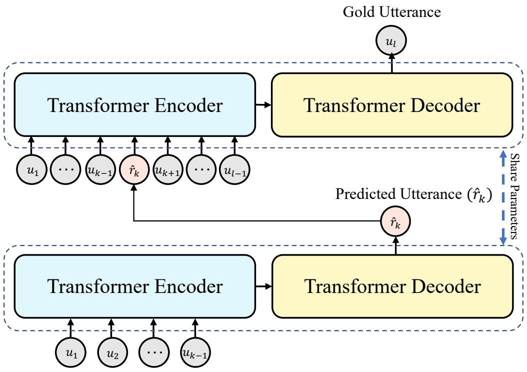


Figure 3: Training with proposed sampling-based methods.

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:

$$c_k = \sum_{i=1}^{D} \frac{\mathbb{1}(f_c(\hat{\mathbf{U}}_1^{i-1}, \hat{\mathbf{r}}_i) = 1)}{D} \tag{2}$$

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

4 Method

We improve the dialogue coherence from two directions: 1) We design hierarchical sampling method to bridge the online-offline discrepancy gap which implicitly improves the coherence in Section 4.1; 2) We introduce explicit coherence optimization methods in 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 gold 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 gold context $\mathbf{U}_1^{l-1}$, we sample an utterance $\mathbf{u}_i,\ i \in [1, l-1]$ using geometric distribution $\sim$ $Geo(p)$ (with $p = 0.2$ and max clip $i_{max} = 10$), 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 $\mathbf{U}_1^{'i-1}$, and then we use the predicted utterance $\hat{\mathbf{r}}_\mathbf{i}$ to replace the gold utterance $\mathbf{u}_i$ in the context $\mathbf{U}_1^{l-1} = \{\mathbf{u}_1, \ldots, \mathbf{u}_i, \ldots, \mathbf{u}_{l-1}\}$, yielding the mixed context $\mathbf{U}_1^{'l-1} = \{\mathbf{u}_1, \ldots, \hat{\mathbf{r}}_i, \ldots, \mathbf{u}_{l-1}\}$. Finally, $\mathbf{U}_1^{'l-1}$ is fed into the encoder. Given a training pair $(\mathbf{U}_1^{l-1}, \mathbf{r})$, the objective is to minimize:

$$\mathcal{L}_{dialogue} = -\sum_{t=1}^{T'} \log p(y_t | \mathbf{U}_1^{'l-1}, y_{1:t-1}) \tag{3}$$

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 $\hat{\mathbf{r}}'_i$ and $\mathbf{u}_i$. While a smaller j to simulate more accumulate errors along with the inference steps. The same as utterance level sampling, $\hat{\mathbf{r}}'_i$ is used to replace $\mathbf{u}_i$.

4.2 Explicit Coherence Optimization

Hierarchical sampling implicitly (indirectly) optimizes online settings by bridging the gap between online settings and offline settings. In this section, we show the possibility of explicitly (directly) optimizing online dialogue response generation. To this end, we consider reinforcement learning (RL), which has been widely used in the dialogue literature (Li et al., 2016b; Saleh et al., 2019; Zhao et al., 2020). Most existing work use RL to optimize offline metrics, however, we use RL to brdige the gap between offline training and online testing by directly optimizing online multi-turn objectives on the training set.

Training. The dialogue model is fine-tuned to optimize the reward model. In particular, we first ask the model to generate a response $\hat{\mathbf{r}}$ based on the context $\mathbf{U}$. Then an external classifier f_c is used to justify whether the response is coherent with the context $\mathbf{U}$. We adopt the logits of f_c corresponding to the coherent label as the reward. The inference stage can be the same as the baseline methods. More details of training with reinforcement learning can be found in Appendix A.2.

5 Experiments

We train our model based on the gold context - response pair on two chit-chat dialogue benchmarks, Wizard of Wikipedia (Dinan et al., 2019b) and DailyDialog (Li et al., 2017). For building Wizard of Wikipedia, two annotators are employed to chat based on an initial topic. The dataset contains 18,430 training dialogues with 1,365 topics. DailyDialog is obtained from educational websites that help English learners to practice English speaking, which consist of 13,118 dialogues.

5.1 Metrics

Following Dinan et al. (2019b) and Kim et al. (2020), we take the perplexity (PPL) of the ground-truth response, given the gold context as input 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 applications, 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. We use c_n to denote the coherence rate between the first n turn and the generate response, and use $avg_n = \frac{\sum_{i=1}^{n} c_i}{n}$ to define the average coherence rate.

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

$$\text{distinct} - n = \frac{\textsc{count}(\textsc{unique}_{n\text{-gram}_i \in \hat{\mathbf{U}}}(n\text{-gram}_i))}{\textsc{count}(\textsc{total}_{n\text{-gram}_i \in \hat{\mathbf{U}}}(n\text{-gram}))} \tag{4}$$

where COUNT(), UNIQUE() and TOTAL() denote counting the item of a list, unique items in a list and enumeration of 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 for 10 turns. The final human-bot test set we derive contains 50

	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/ gold 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/ Noise	99.2	95.4	76.5	58.7	47.1	35.4	31.4	22.1	23.1	12.4	75.4	50.1	18.1
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. Noted that "BART w/ Golden context" represents the offline test result, which can be consider the ceiling performance on the online setting (the model generate the gold response at each turn).

Model	avg_5	avg_10	PPL
Multi-turn BART	66.3	49.8	8.6
BART w/ Hierarchial	70.8	58.7	**8.1**
Multi-turn Blender	68.7	66.1	11.6
Blender w/ Hierarchial	**73.2**	**70.1**	11.6

Table 2: Test performance on DailyDialogue test set.

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.

5.2 Baselines

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

BART w/ gold context. We fine-tune BART on the Wizard of Wikipedia training set. During inference at turn k, the gold context $\mathbf{U}_1^{k-1}$ is used to produce the response $\hat{\mathbf{r}}_k$. Because the gold 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 gold context-response pair. Different from BART w/ gold 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.

w/ Noise After sample an utterance $\mathbf{u}_i$, we use a random noise $\mathbf{u}_{random}$ randomly sampled from the training set to replace $\mathbf{u}_i$.

5.3 Results

Table 1 and Table 2 report the performance of coherence rate as well as PPL for various methods on Wizard of Wikipedia and DailyDialog. Table 3 shows the distinct-n for the predicted context generated by these methods.

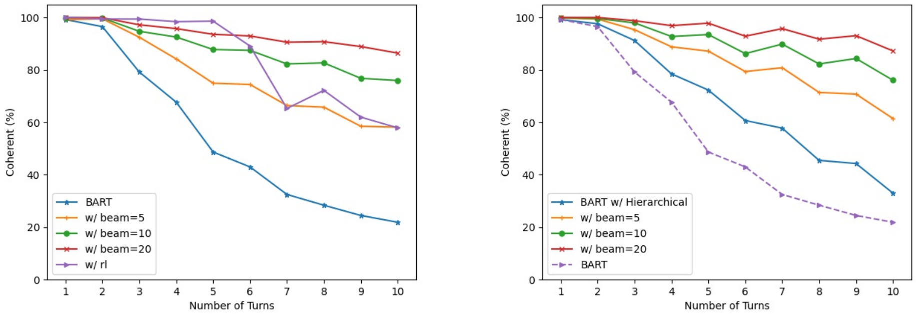
Predicted Context vs gold Context. We first compare whether the dialogue generation model is able to generate coherence response based on the gold context and the predicted context. As shown on the top of Table 1, the coherence rate of BART w/ gold context does not decrease significantly with the number of turns increases. The performance drops by only 5.6 points coherence rate from 2

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

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



(a) Multi-turn BART.

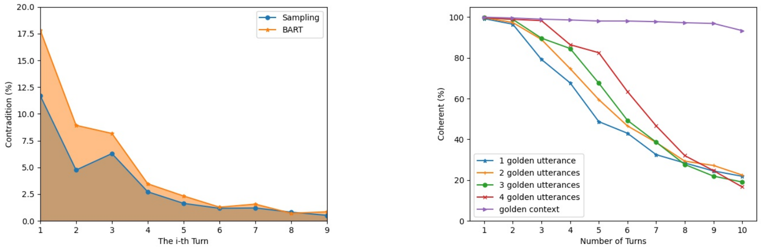
(b) Multi-turn BART w/ Hierarchical Sampling.

Figure 4: Coherence rate with explicit optimization.

turns to 10 turns. However, given the predicted context, the coherence rate decreases sharply as the number of turns increases, with only 21.9 c_{10}. This shows the severity of the discrepancy problem in real-world multi-turn dialogue generation.

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 gold 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 method 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 (p<0.01), 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. Furthermore, multi-turn BART w/ Noise do not work well, since sampled noises are difficult to accurately simulate errors of the inference scene during training. Experiments on DailyDialogue (Table 2) show that, our method also achieves 8.9% and 2.8% avg_10 improvement using BART and Blender as backbone, respectively.



(a) Contradiction rate across different turn. Contradiction rate is defined by (100% − coherence rate).

(b) Coherence rate across the number of gold utterances at the beginning.

Figure 5: Analysis.

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 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 3 reports the non-repetition rate of the proposed 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, with a large margin of 11.92, 17.47 and 18.94, respectively, 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 4 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 Coherence Optimization. Figure 4 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 try to enhance dialogue coherence explicitly during inference by using 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. We 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. Our hierarchical sampling-based methods also consistently perform better than multi-turn BART by introducing coherence re-ranking.

6 Analysis

The Number of Gold Turns. We investigate whether a larger number of gold turns at the start is able to help model to produce more coherent responses during inference. Figure 5(b) shows the

$\mathbf{u}_1$(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?
$\mathbf{u}_1$(Prompt)	Hi I love to draw cartoons.
Multi-turn BART	
$\hat{\mathbf{r}}_2$	I love to draw cartoons too! I like to draw animals, plants, birds, etc.
$\hat{\mathbf{r}}_3$	I like to draw animals too.
$\hat{\mathbf{r}}_4$	Do you know how long cartoons have been around?
$\hat{\mathbf{r}}_5$	I'm not sure, but I do know that cartoons have been around since the 17th century!
$\hat{\mathbf{r}}_6$	Wow that is a long time ago. Do you know if cartoons are still popular today?
Multi-turn BART w/ Hierarchical sampling	
$\hat{\mathbf{r}}_2$	I love to draw cartoons too. I like to use graphite pencils, pen and ink.
$\hat{\mathbf{r}}_3$	I like to use graphite pencils as well. What is your favorite cartoon?
$\hat{\mathbf{r}}_4$	My favorite cartoon is The Simpsons.
$\hat{\mathbf{r}}_5$	I love The Simpsons as well. What is your favorite episode of the Simpsons?
$\hat{\mathbf{r}}_6$	My favorite episode of The Simpsons is the one that aired in 2011.

Table 5: Examples of generated responses given a prompt utterance on the Wizard of Wikipedia Test Seen.

coherence rate against the number of gold utterances at the beginning during the self-talk, drawing using the gold context as a reference. It can be seen that a larger number of gold 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 gold turns cannot help to alleviate the discrepancy between training and real-world testing.

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 5(a), 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 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.

Case Study. We present examples for a better understanding of multi-turn BART and our model in Table 5. 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 (gold 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. 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, pp. 1171–1179, Cambridge, MA, USA, 2015. MIT Press.

Leyang Cui, Yu Wu, Shujie Liu, Yue Zhang, and Ming Zhou. MuTual: A dataset for multi-turn dialogue reasoning. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 1406–1416, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.130. URL https://aclanthology.org/2020.acl-main.130.

Leyang Cui, Yu Wu, Shujie Liu, and Yue Zhang. Knowledge enhanced fine-tuning for better handling unseen entities in dialogue generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 2328–2337, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.179. URL https://aclanthology.org/2021.emnlp-main.179.

Jan Deriu, Don Tuggener, Pius von Däniken, Jon Ander Campos, Alvaro Rodrigo, Thiziri Belkacem, Aitor Soroa, Eneko Agirre, and Mark Cieliebak. 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)*, pp. 3971–3984, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.326. URL https://aclanthology.org/2020.emnlp-main.326.

Emily Dinan, Varvara Logacheva, Valentin Malykh, Alexander Miller, Kurt Shuster, Jack Urbanek, Douwe Kiela, Arthur Szlam, Iulian Serban, Ryan Lowe, and et al. The second conversational intelligence challenge (convai2). *The Springer Series on Challenges in Machine Learning*, pp. 187–208, Nov 2019a. ISSN 2520-1328. doi: 10.1007/978-3-030-29135-8_7. URL http://dx.doi.org/10.1007/978-3-030-29135-8_7.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*, 2019b. URL https://openreview.net/forum?id=r1l73iRqKm.

Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, Craig Ferguson, Agata Lapedriza, Noah Jones, Shixiang Gu, and Rosalind Picard. Way off-policy batch deep reinforcement learning of human preferences in dialog, 2020. URL https://openreview.net/forum?id=rJl5rRVFvH.

Anant Khandelwal. 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)*, pp. 69–80, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.dialdoc-1.10. URL https://aclanthology.org/2021.dialdoc-1.10.

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

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 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*, pp. 7871–7880, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.703. URL https://aclanthology.org/2020.acl-main.703.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 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*, pp. 110–119, San Diego, California, June 2016a. Association for Computational Linguistics. doi: 10.18653/v1/N16-1014. URL https://aclanthology.org/N16-1014.

Jiwei Li, Will Monroe, Alan Ritter, Dan Jurafsky, Michel Galley, and Jianfeng Gao. Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1192–1202, Austin, Texas, November 2016b. Association for Computational Linguistics. doi: 10.18653/v1/D16-1127. URL https://aclanthology.org/D16-1127.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4715–4728, Online, July 2020a. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.428. URL https://aclanthology.org/2020.acl-main.428.

Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4715–4728, Online, July 2020b. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.428. URL https://aclanthology.org/2020.acl-main.428.

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

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 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*, pp. 1057–1067, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.91. URL https://aclanthology.org/2021.findings-acl.91.

Zekang Li, Jinchao Zhang, Zhengcong Fei, Yang Feng, and Jie Zhou. 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)*, pp. 128–138, Online, August 2021b. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.11. URL https://aclanthology.org/2021.acl-long.11.

Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. Confidence-aware scheduled sampling for neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pp. 2327–2337, Online, August 2021a. Association for Computational Linguistics. doi: 10.18653/v1/2021.findings-acl.205. URL https://aclanthology.org/2021.findings-acl.205.

Yijin Liu, Fandong Meng, Yufeng Chen, Jinan Xu, and Jie Zhou. 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, 2021b.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. URL https://openreview.net/forum?id=Bkg6RiCqY7.

Tsvetomila Mihaylova and André F. T. Martins. Scheduled sampling for transformers. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pp. 351–356, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-2049. URL https://aclanthology.org/P19-2049.

Yixin Nie, Mary Williamson, Mohit Bansal, Douwe Kiela, and Jason Weston. 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)*, pp. 1699–1713, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.134. URL https://aclanthology.org/2021.acl-long.134.

Libo Qin, Tianbao Xie, Shijue Huang, Qiguang Chen, Xiao Xu, and Wanxiang Che. 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*, pp. 2357–2367, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.emnlp-main.182.

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. Conversational ai: The science behind the alexa prize, 2018.

Marc'Aurelio Ranzato, Sumit Chopra, Michael Auli, and Wojciech Zaremba. Sequence level training with recurrent neural networks. In Yoshua Bengio and Yann LeCun (eds.), *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016. URL http://arxiv.org/abs/1511.06732.

Alan Ritter, Colin Cherry, and William B. Dolan. Data-driven response generation in social media. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pp. 583–593, Edinburgh, Scotland, UK., July 2011. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D11-1054.

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. Recipes for building an open-domain chatbot, 2020.

Abdelrhman Saleh, Natasha Jaques, Asma Ghandeharioun, Judy Hanwen Shen, and Rosalind W. Picard. Hierarchical reinforcement learning for open-domain dialog. *CoRR*, abs/1909.07547, 2019. URL http://arxiv.org/abs/1909.07547.

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

Heung-yeung Shum, Xiao-dong He, and Di Li. From eliza to xiaoice: challenges and opportunities with social chatbots. *Frontiers of Information Technology & Electronic Engineering*, 19(1):10–26, Jan 2018. ISSN 2095-9230. doi: 10.1631/FITEE.1700826. URL https://doi.org/10.1631/FITEE.1700826.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 27. Curran Associates, Inc., 2014. URL https://proceedings.neurips.cc/paper/2014/file/a14ac55a4f27472c5d894ec1c3c743d2-Paper.pdf.

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

Hao Wang, Zhengdong Lu, Hang Li, and Enhong Chen. A dataset for research on short-text conversations. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pp. 935–945, Seattle, Washington, USA, October 2013. Association for Computational Linguistics. URL https://aclanthology.org/D13-1096.

Sean Welleck, Jason Weston, Arthur Szlam, and Kyunghyun Cho. Dialogue natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 3731–3741, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1363. URL https://aclanthology.org/P19-1363.

Sean Welleck, Ilia Kulikov, Stephen Roller, Emily Dinan, Kyunghyun Cho, and Jason Weston. Neural text generation with unlikelihood training. In *International Conference on Learning Representations*, 2020. URL https://openreview.net/forum?id=SJeYe0NtvH.

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

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

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2204–2213, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1205. URL `https://aclanthology.org/P18-1205`.

Wen Zhang, Yang Feng, Fandong Meng, Di You, and Qun Liu. Bridging the gap between training and inference for neural machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pp. 4334–4343, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1426. URL `https://aclanthology.org/P19-1426`.

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

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. Dialogpt: Large-scale generative pre-training for conversational response generation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020b. doi: 10.18653/v1/2020.acl-demos.30. URL `http://dx.doi.org/10.18653/V1/2020.ACL-DEMOS.30`.

Xueliang Zhao, Wei Wu, Can Xu, Chongyang Tao, Dongyan Zhao, and Rui Yan. Knowledge-grounded dialogue generation with pre-trained language models. In *EMNLP*, 2020.

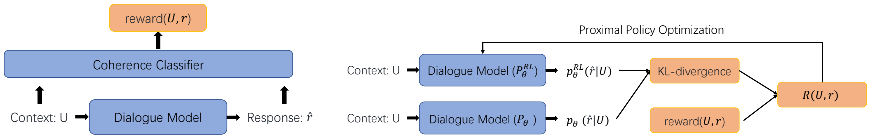
Daniel M. Ziegler, Nisan Stiennon, Jeffrey Wu, Tom B. Brown, Alec Radford, Dario Amodei, Paul Christiano, and Geoffrey Irving. Fine-tuning language models from human preferences, 2019.

A Appendix

A.1 Setup

We implement our methods with `transformers` and choose `bart-base` as the pre-trained transformer language model. AdamW (Loshchilov & 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. The training time of our methods is 0.6 times slower than the baseline method. Our inference time is the same as that of the baseline. For the coherence-oriented reinforcement learning method, we set β in Equation 6 as 0.2. For computational efficiency, we truncate the maximum decode length as 20 to calculate the KL-divergence.

A.2 RL methods



(a) Reward Calculation.

(b) Optimization.

Figure 6: Coherence-Oriented Reinforcement Learning.

As shown in Figure 6, we fine-tune the dialogue model P_θ to optimize the reward model P_θ^{RL}. 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 6(b)). Formally, given the context $\mathbf{U}$, we calculate the KL-divergence between two models' output probabilities

$$KL(\mathbf{U}) = \sum_{t=1}^{T'} \log \frac{p_\theta^{RL}(\mathbf{x}_t|\mathbf{U}, \mathbf{x}_{1:t-1})}{p_\theta(\mathbf{x}_t|\mathbf{U}, \mathbf{x}_{1:t-1})} \tag{5}$$

$KL(\mathbf{U})$ 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:

$$Reward(\mathbf{U}, \mathbf{r}) = f_c(\mathbf{U}, \hat{\mathbf{r}}) - \beta KL(\mathbf{U}) \tag{6}$$

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.