On Controlling Fallback Responses for Grounded Dialogue Generation

Hongyuan Lu, Wai Lam, Hong Cheng, Helen M. Meng
The Chinese University of Hong Kong
{hylu,wlam,hcheng,hmmeng}@se.cuhk.edu.hk

Abstract

Dialogue agents can leverage external textual knowledge to generate responses of a higher quality. To our best knowledge, most existing works on knowledge grounded dialogue settings assume that the user intention is always answerable. Unfortunately, this is impractical as there is no guarantee that the knowledge retrievers could always retrieve the desired knowledge. Therefore, this is crucial to incorporate fallback responses to respond to unanswerable contexts appropriately while responding to the answerable contexts in an informative manner. We propose a novel framework that automatically generates a control token with the generator to bias the succeeding response towards informativeness for answerable contexts and fallback for unanswerable contexts in an end-to-end manner. Since no existing knowledge grounded dialogue dataset considers this aim, we augment the existing dataset with unanswerable contexts to conduct our experiments. Automatic and human evaluation results indicate that naively incorporating fallback responses with controlled text generation still hurts informativeness for answerable context. In contrast, our proposed framework effectively mitigates this problem while still appropriately presenting fallback responses to unanswerable contexts. Such a framework also reduces the extra burden of the additional classifier and the overheads introduced in the previous works, which operates in a pipeline manner.[1]

1 Introduction

Building knowledge grounded dialogue agents has been an important research line (Bordes et al., 2016; Young et al., 2017; Zhou et al., 2018; Chaudhuri et al., 2019; Moon et al., 2019; Dziri et al., 2021). Such incorporation of real-world knowledge (Young et al., 2017; Zhou et al., 2018) gives rise to consistent, informative and engaging response generation. Unfortunately, even with a high-quality knowledge retriever, there is no guarantee that the desired knowledge can always be retrieved. There is indeed even no guarantee for the existence of the desired knowledge in the knowledge database. Hence, presenting fallback responses is an essential ability for grounded dialogue agents. We make use of the notion of answerability that represents whether a dialogue context is answerable or not conditioned on the knowledge retrieved. Figure 1 depicts an example to illustrate the importance of answerability in grounded dialogue response generation. As in the unanswerable dialogue context, a fallback response is desirable. Conversely, as in the answerable dialogue context, the response should be as informative as possible.

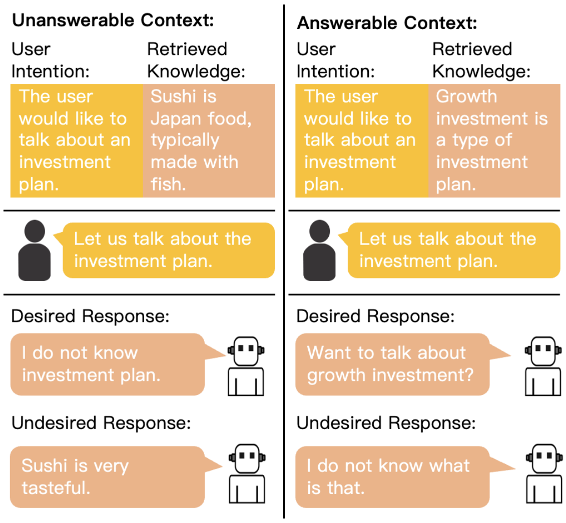

Unanswerable Context:
User Intention:
Retrieved Knowledge:
The user would like to talk about an investment plan.
Sushi is Japan food, typically made with fish.
Let us talk about the investment plan.
Desired Response:
I do not know investment plan.
Undesired Response:
Sushi is very tasteful.
Answerable Context:
User Intention:
Retrieved Knowledge:
The user would like to talk about an investment plan.
Growth investment is a type of investment plan.
Let us talk about the investment plan.
Desired Response:
Want to talk about growth investment?
Undesired Response:
I do not know what is that.

Figure 1: An illustrated example for answerable and unanswerable context conditioned on retrieved knowledge, along with corresponding desired and undesired responses. We demonstrate an easy single-turn conversation for simplicity. Better viewed in colour.

Although the concept of answerability has been well explored in other NLP areas such as Question Answering (Rajpurkar et al., 2018), it is underex-

[1] Related resources can be found at https://github.com/HongyuanLuke/OCFR.

Findings of the Association for Computational Linguistics: ACL 2022, pages 2591 - 2601
May 22-27, 2022 ©2022 Association for Computational Linguistics

plored in the dialogue community. Most existing knowledge grounded dialogue agents (Young et al., 2017; Chaudhuri et al., 2019; Prabhumoye et al., 2021) and knowledge grounded dialogue datasets (Zhou et al., 2018) ignore the fallback issue. However, this is almost impractical, and it is unlikely to happen in the real world that all the contexts are answerable. One recent approach has been proposed to calibrate responses with the appropriate linguistic confidence (Mielke et al., 2020); however, it overlooks informativeness, or diversity, (Li et al., 2016a; Vijayakumar et al., 2016; Fan et al., 2018; Holtzman et al., 2020; Tang et al., 2021; Wang et al., 2021), which is an important quality metric for a dialogue system. Though the previous work mentioned above (Mielke et al., 2020) employs an additional classifier for answerability, or in their case, linguistic confidence level, we demonstrate that our proposed method can achieve higher accuracy with the response generator.

Our proposed model employs controlled text generation (CTG, Niu and Bansal 2018; Mielke et al. 2020; Gehman et al. 2020; Xu et al. 2020; Baheti et al. 2021). Its central idea is to bias the generation towards a specific style by placing a control token in the input context. This control token has been investigated via two strategies: manually placed (Baheti et al., 2021) or model classified (Mielke et al., 2020). One can manually place a control token with low offensiveness to prevent the dialogue response generator from generating an offensive context (Baheti et al., 2021). One can also use a classifier to determine the linguistic confidence that the generator should present in its response generation (Mielke et al., 2020). In contrast to these works, one of our characteristics is that *while these works focus on the classification task only, our work turns the classification task into a generative manner and then exploits the classification result for the succeeding generation task within a single autoregressive generator.*

Since no existing dataset is suitable for our task, we derived a dataset by augmenting an existing dialogue dataset with unanswerable tuples of the dialogue context and the knowledge retrieved, and we conducted our experiments on the derived dataset. Our experimental results indicate that incorporating controlled text generation (Mielke et al., 2020) can capture answerability and rigorously replies with a fallback response to unanswerable contexts. However, it still undesirably hurts informativeness for answerable contexts by frequently responding with fallback responses to answerable contexts. Our method can achieve higher accuracy in classifying answerability than the traditional controlled text generation. This reduces the chance of responding with fallback to answerable contexts and thus improves the informativeness for responses to answerable contexts while still responding appropriately with fallback to unanswerable contexts.

2 Related Work

2.1 Grounded Dialogue Generation

Augmenting the dialogue agents with either table-formatted knowledge base (Bordes et al., 2016) or graph-formatted knowledge base (Moon et al., 2019) enables the dialogue agents to leverage real-world facts. This is crucial in both task-oriented dialogue (Moon et al., 2019) and chitchat dialogue (Chaudhuri et al., 2019). Dialogue agents grounded with common sense tends to be more engaging as well (Young et al., 2017). Furthermore, it also has been pointed out that using a knowledge base could reduce the problem of hallucinations (Dziri et al., 2021). Another research line tends to compress knowledge into model parameters, either by training set augmentation with template-based method (Madotto et al., 2020) or using neural architectures as domain-specific adapters (Xu et al., 2021).

2.2 Fallback Response in Dialogue Generation

Fallback response, or even answerability, remains under-explored for grounded dialogue agents. One recent close work calibrates responses with appropriate linguistic confidence (Mielke et al., 2020). Another close work parapharses fallback responses with contextualization (Shrivastava et al., 2021).

2.3 Informative Dialogue Generation

Informativeness, or diversity, plays an important role in engaging response generation. Modified decoding strategy with a dedicated objective improves diversity (Vijayakumar et al., 2016). Maximum mutual information (Li et al., 2016a) improves diversity with a diversity-promoting objective function for reranking. More recently, top-k sampling (Fan et al., 2018) and nucleus sampling (Holtzman et al., 2020) improve diversity by truncating the vocabularies or probability density to be sampled from and has shown their superiority over the traditional beam search for diverse dialogue generation.

3 Methodology

3.1 Background

We focus on the task of dialogue generation that is capable of recognizing unanswerable dialogue contexts and generating fallback response generation in an end-to-end manner. We adopt an end-to-end autoregressive language model (Zhang et al., 2020) as our neural dialogue generator. We denote this model as $\mathcal{M}$. By further denoting the knowledge retrieved as $\boldsymbol{k}$, dialogue context as $\boldsymbol{c}$ and dialogue response as $\boldsymbol{r}$, this generation task can be formulated as a mapping function that generates the dialogue response conditioned on the dialogue context and the knowledge retrieved:

$$\mathcal{M} : \boldsymbol{k}, \boldsymbol{c} \rightarrow \boldsymbol{r}$$

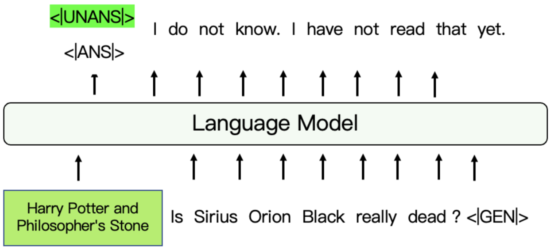
Unfortunately, naively approximating this function with maximum likelihood estimation might confuse the generator as the responses for the unanswerable contexts typically confess ignorance. This type of fallback response then becomes universally likely. Without an appropriate control on generating fallback responses, our generator can even give an answerable context a response that confesses ignorance. For example, a response that confesses ignorance could be templated as 'I do not know, I have not ...' where the contextualization follows. However, simply training on this instance will make 'I' to be universally likely followed by 'do'. Therefore, even for answerable user intention, the generator could fail into producing a fallback response immediately after decoding an 'I'.

3.2 Controlling Fallback Response

To effectively bias generation towards confessing ignorance for unanswerable dialogue as well as bias generation towards expressing informativeness for answerable contexts, we leverage controlled text generation. The task can be expressed as:

$$p(\boldsymbol{r} \mid \boldsymbol{k}, \boldsymbol{c}) \propto p(\boldsymbol{a} \mid \boldsymbol{k}, \boldsymbol{c})\, p(\boldsymbol{r} \mid \boldsymbol{a}, \boldsymbol{k}, \boldsymbol{c}),$$

where the answerability $\boldsymbol{a}$ in the above formula is a binary control token that is either <|ANS|> or <|UNANS|>. The former biases the succeeding dialogue response generation towards informativeness, and the latter biases the succeeding generation towards fallback. In the previous work done by Mielke et al. (2020), this control token is predicted by employing an extra classifier that outputs whether the dialogue context is answerable:



<|UNANS|>
<|ANS|>
I do not know. I have not read that yet.
Language Model
Harry Potter and Philosopher's Stone
Is Sirius Orion Black really dead ? <|GEN|>

Figure 2: An illustration for the inferencing stage for our proposed framework. This dialogue is not answerable since the retrieved document does not contain the discussed user intention. Therefore, our proposed framework automatically selects a binary control token, which controls the succeeding response generation towards expressing ignorance.

$$p(\boldsymbol{r} \mid \boldsymbol{k}, \boldsymbol{c}) \propto p_{\text{classifier}}(\boldsymbol{a} \mid \boldsymbol{k}, \boldsymbol{c})\, p(\boldsymbol{r} \mid \boldsymbol{a}, \boldsymbol{k}, \boldsymbol{c})$$

This introduces extra parameters from the classifier and extra overheads for the inference. Indeed, this work has primarily focused on rephrasing responses with appropriate linguistic confidence, and their methodology requires two generators and one classifier. Our method differs as we augment the dialogue agent with the unstructured textual knowledge while theirs tests the knowledge inherently encoded in the model.[2] Their proposed method operates in a pipeline fashion that first generates a response, then obtains the control token with the classifier, and finally rephrases the generation with the second generator. An important observation is that the question or the dialogue context already contains enough information to judge the appropriate linguistic confidence level (Mielke et al., 2020). In addition, our primary goal is to directly control the fallback generation rather than maintain the semantics while calibrating the linguistic confidence. Therefore, we exclude the use of the first generator throughout our experiments.

3.3 Control Token Generation

Since a confidence level, or in our words, answerability, can be appropriately obtained even before generation, we could exploit this and remove the rephrasing generator. Furthermore, if we can further reduce the need for an answerability classifier, we can build an end-to-end system that replies with

[2] Previous work has employed a closed-book QA dataset as their testbed (Mielke et al., 2020).

fallback answers to unanswerable contexts. To this end, we propose a framework that incorporates the classification of control tokens into the response generation by leveraging the power of pre-trained language models to formulate language understanding tasks into a generative manner (Raffel et al., 2019; Liu et al., 2021a,b; Zhang et al., 2021). We illustrate the overall idea of our proposed framework in Figure 2. Our framework incorporates a notion called control token generation, where the control token could be automatically generated by the dialogue generator in an end-to-end manner. Firstly, we place a token of <|GEN|> as a prompt to signal the model to generate a binary control token, either <|ANS|> or <|UNANS|>. The former indicates the dialogue context as answerable, and the latter indicates the dialogue context as unanswerable. This then continues in an autoregressive manner for the model to complete the remaining response generation. For the control token of <|ANS|>, it follows a search space that is diverse and informative. In contrast, the control token <|UNANS|> guides into a semantical search space for fallback responses, which typically confesses ignorance, or low linguistic confidence level (Mielke et al., 2020). We thus formulate the problem as:

$$p(\boldsymbol{r} \mid \boldsymbol{k}, \boldsymbol{c}) \propto p_{\text{generator}}(\boldsymbol{a} \mid \boldsymbol{k}, \boldsymbol{c})\, p(\boldsymbol{r} \mid \boldsymbol{a}, \boldsymbol{k}, \boldsymbol{c})$$

Although previous works have formulated fallback response generation in a pipeline manner where the original response should attend (Mielke et al., 2020; Shrivastava et al., 2021), our proposed framework leverages control token to directly guide the response into either informative response or fallback response that confesses ignorance. Furthermore, our framework leverages the understanding power of large-scaled pre-trained language model (Liu et al., 2021b) and reduces the need for an extra answerability classifier by incorporating control token generation. As a result, this turns the whole system from a pipeline manner into an end-to-end manner, which drastically reduces the model size and the inference overheads.

3.4 Sequence-to-Sequence Learning

We adopt a single autoregressive Seq2Seq generator (Zhang et al., 2020) as both our control token generator as well as our dialogue response generator. Precisely, our network accepts an input concatenation of text knowledge $\boldsymbol{k}$ and dialogue context $\boldsymbol{c}$, and outputs an answerability control token $\boldsymbol{a}$ first, and then outputs the remaining dialogue response $\boldsymbol{r}$ one by one and left to right.

At the i-th timestep, the generator picks the next token r_i to be presented in the output that maximises the conditional probability:

$$r_i = \underset{r_i \in \mathcal{V}}{\operatorname{argmax}}\; p(r_i \mid r_1, ..., r_{i-1}, \boldsymbol{a}, \boldsymbol{k}, \boldsymbol{c})$$

Note that $\mathcal{V}$ in the equation above represents the vocabulary space to be decoded from.

Training To train our language model, we preprocess the original training instances to incorporate control token generation. The original training instance is the concatenation of knowledge, dialogue context, and response:

$$[\boldsymbol{k}; \boldsymbol{c}; \boldsymbol{r}]$$

We derive our new training instances as the concatenation of knowledge, dialogue context, control token, and response:

$$[\boldsymbol{k}; \boldsymbol{c}; \texttt{<|GEN|>}; \boldsymbol{a}; \boldsymbol{r}]$$

Note that <|GEN|> is a prompt token to signal the model to generate the succeeding answerability control token, and $\boldsymbol{a}$ is the binary control token that guides the subsequent dialogue generation.

Inferencing While our dialogue generation follows the traditional scheme where we adopt the nucleus sampling, we found in our early experiments that greedy decoding can be effective for the task of control token generation, which improves classification accuracy. We thus propose two decoding strategies:

- *Unhindered Sampling* uses nucleus sampling for both control token generation or answerability classification and dialogue response generation throughout the decoding stage.
- *Bottleneck Sampling*[3] uses greedy decoding for control token generation and nucleus sampling for dialogue response generation.

Although the former is straightforward and easy to implement, we demonstrate that the latter variant can remarkably improve the answerability classification accuracy and hence improve the succeeding response generation. Both of them can improve the response quality for the answerable contexts.

[3] The name is due to the fact that it has a shrunk distribution like a neck with greedy decoding at the place of control token before a flatten probability distribution for dialogue sampling.

4 Experimental Setup

Dataset Preparation Since no existing dataset is suitable for our aim, we derive our dataset based on the CMU DOCUMENT GROUNDED CONVERSATIONS DATASET (CMU DOG) dataset (Zhou et al., 2018). CMU DOG is a multi-turn dyadic dialogue dataset in which two crowdsource workers converse and find out more about a specific movie based on that particular film profile. While most of the dialogue datasets focus only on either chitchat (Zhang et al., 2018) or task-oriented dialogue (Budzianowski et al., 2018), CMU DOG interleaves chitchat and task-oriented dialogue (Zhou et al., 2018). It thus requires the agent to be both informative and knowledge grounded. Such knowledge grounded dialogue agents should appropriately respond with fallbacks to the unanswerable contexts without hurting informativeness on the responses to the answerable contexts. Therefore, CMU DOG is a suitable dataset to validate the effectiveness of our proposed framework.

We label all of the original instances as answerable conditioned on the ground truth knowledge. Indeed, the crowdsource workers converse based on the ground truth knowledge (Zhou et al., 2018). We then augment with unanswerable dialogues by sampling two instances $[\boldsymbol{k_1}; \boldsymbol{c_1}; \boldsymbol{r_1}]$ and $[\boldsymbol{k_2}; \boldsymbol{c_2}; \boldsymbol{r_2}]$ from the original dataset where $\boldsymbol{k_1} \neq \boldsymbol{k_2}$. This results into two unanswerable instances $[\boldsymbol{k_1}; \boldsymbol{c_2}; \boldsymbol{f}]$ and $[\boldsymbol{k_2}; \boldsymbol{c_1}; \boldsymbol{f}]$, where $\boldsymbol{f}$ represents the fallback responses that typically confess ignorance. This operation derives into a training / development / testing partition with 100,497 / 6,677 / 18,921 instances respectively for the CMU DOG dataset.[4]

Unlike chitchat dialogue datasets (Zhang et al., 2018) which consist of several dialogue topics that can be irrelevant to each other, the movie profiles from CMU DOG guarantees to be within the same domain. This is important as real-world retrievers can be competitive, meaning that irrelevant retrieved knowledge can make the task oversimplified into relevance classification. Fortunately, our augmentation strategy can still derive an answerability task with moderate difficulty in which the competitive classifiers report only about 82% test accuracy on the derived CMU DOG dataset.

Baseline and Comparison Model Our baseline adopts a vanilla Seq2Seq generator as a basic function mapper as described in Section 3.1 which maps the concatenation of knowledge retrieved $\boldsymbol{k}$ and dialogue context $\boldsymbol{c}$ to dialogue response $\boldsymbol{r}$ without any control over the fallback response as well as the notion of answerability. One comparison model is derived from the previous work done by Mielke et al. (2020) to employ an additional classifier to map the concatenation of knowledge retrieved $\boldsymbol{k}$ and dialogue context $\boldsymbol{c}$ to answerability control token $\boldsymbol{a}$. It then follows the classical controlled text generation procedure to feed the concatenation of the knowledge, context and control token into the generator for response generation.

Implementation Details For all the generators implemented for the baseline, comparison model and our method, we employ the state-of-the-art GPT2-based (Radford et al., 2019) dialogue response generator DIALOGPT-SMALL (Zhang et al., 2020). We also attempted on DIALOGPT-MEDIUM and DIALOGPT-LARGE. We found all three of them tend to respond inappropriately with fallbacks to the answerable dialogue contexts, and they report similar diversity measurements. Therefore, we adopt DIALOGPT-SMALL for simplicity. We use a learning rate of $5e{-}4$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1e{-}8$. We adopt ROBERTA-BASE (Liu et al., 2019) as the answerability classifier to be used in our comparison model. We also experimented on BERT-BASE, BERT-LARGE (Devlin et al., 2019) and ROBERTA-LARGE, which led to a similar accuracy. Therefore, we adopt ROBERTA-BASE for simplicity. For the classifier, we use a learning rate of $5e{-}6$, $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 1e{-}8$. Since we are interested in diversity, or informativeness, we use nucleus sampling, or top-p sampling (Holtzman et al., 2020) as our decoding mechanism throughout our experiments for all our baseline, comparison model, and our method, in which we set $p = 0.95$ as the hyper-parameter as in the work done by Holtzman et al. (2020). We conduct our experiments with the TRANSFORMERS library (Wolf et al., 2020).

Evaluation Metrics In this work, we mainly focus on generation diversity for answerable contexts. We also report our investigation on fallback issues for unanswerable contexts as well as the classification accuracy. We followed previous works to adopt Distinct-n (Li et al., 2016b; Gao et al., 2019;

[4] All the partitions are individually processed to avoid data leakage. They contain answerable and unanswerable contexts half by half. We follow the train / dev / test split from the ParlAI platform (Miller et al., 2017). We drop some instances to match the maximum input length for BERT and ROBERTA.

Model	B^{+}	B-2^{+}	B-3^{+}	B-4^{+}	D-1^{+}	D-2^{+}	D-3^{+}	D-4^{+}	D-5^{+}	D-6^{+}
CMU Document Grounded Conversations Dataset										
E2E Baseline	0.062	0.620	0.121	0.037	0.034	0.185	0.340	0.424	0.461	0.477
Mielke et al. (2020)	0.037	0.562	0.083	0.021	0.037	0.207	0.390	0.493	0.540	0.560
Ours w/ Unhindered S.	0.095	0.533	0.143	0.055	0.041	0.237	0.459	0.591	0.654	0.681
Ours w/ Bottleneck S.	**0.123**	**0.682**	**0.164**	**0.075**	**0.041**	**0.239**	**0.465**	**0.601**	**0.666**	**0.694**

Table 1: Generation results on CMU DoG dataset. We report n-gram Distinct where n={1,2,3,4,5,6}. B-2^{+} denotes the metrics of BLEU-2 on the answerable contexts. D-2^{+} denotes the metrics of Distinct-2 on the answerable contexts. The same convention follows for the remaining metrics. The best results are highlighted in bold.

Model	FR^{+}	FR^{-}
E2E baseline	3957	7767
Mielke et al. (2020)	2640	**8945**
Ours w/ Unhindered Sampling	877	8699
Ours w/ Bottleneck Sampling	**744**	8719

Table 2: Quality measurements for fallback response generation reported on CMU DoG. FR^{+} and FR^{-} represents the number of fallback responses replied to answerable and unanswerable contexts respectively.

Cai et al., 2019; Lippe et al., 2020). It is the ratio of the number of unique n-grams against the total number of n-grams generated. We follow the work done by Gao et al. (2019) to calculate Distinct-n:

$$\text{Distinct-n} = \frac{\mid \bigcup_{i=1}^{N} \mathcal{R}_i \mid}{\sum_{i=1}^{N} \mid \mathcal{R}_i \mid},$$

where $\mathcal{R}_i$ represents the set of n-grams in the sample i and $\mid \mathcal{R}_i \mid$ represents the number of elements in the set. Gao et al. (2019) has employed n={1,2}, and they primarily focused on task-oriented dialogue. In contrast, we conducted our experiments on CMU DoG, which interleaves chit-chat and task-oriented dialogue. Since two tasks naturally differ, for our investigation on CMU DoG, we extend the unigrams and the bigrams to trigrams, four-grams, five-grams and six-grams, and we report Distinct-n where n={1,2,3,4,5,6} to measure phrase-level and sentence-level diversity. We also report BLEU score, which is a widely adopted sequence evaluation metrics (Papineni et al., 2002). To investigate fallback response generation, we report the number of fallback responses replied to answerable (FR^{+}) and unanswerable contexts (FR^{-}).[5] The former attains a better quality with lower values and the latter attains a better quality with higher values. For the control token generation, or answerability classification, we report the overall accuracy (Acc.), recall (Rec.), precision (Pre.), and F1-score (F1).

5 Results and Discussions

5.1 Main Results

Table 1 depicts the main results on dialogue generation. B represents BLEU scores, and D represents Distinct scores. We mainly report on the answerable dialogue contexts, i.e. the original dataset. As done in Mielke et al. (2020), we build a comparison model by incorporating the idea of controlled text generation to generate fallback responses. Incorporating controlled text generation does improve response diversity; however, it degrades the BLEU scores, which could be a side effect of naively incorporating controlled text generation. We postulate that placing an unanswerable control token makes the model more confident in outputting a fallback response even to answerable contexts. In contrast, a basic E2E model without controlled text generation can still escape from the fallback situation during the decoding phase. This leads to the conclusion that naively incorporating controlled text generation still hurts the response quality. In contrast, our proposed methods are not influenced by the side effect discussed above and report better BLEU scores than our baselines. In addition to the remarkable improvement in BLEU scores, our proposed method can improve word-level diversity (Li et al., 2016b; Gao et al., 2019; Cai et al., 2019; Lippe et al., 2020) as well as phrase-level and sentence-level diversity, which surpasses all our baseline and comparison models.

5.2 Decoding Methods

Non-deterministic sampling can improve the diversity or surprisingness of the response generation (Fan et al., 2018; Holtzman et al., 2020). One should be curious about whether such a case applies to the control token generation as well. Our

[5] The scores are obtained by keyword detection.

Model	Rec.$^+$	Pre.$^+$	Rec.$^-$	Pre.$^-$	F1$^+$	F1$^-$	Acc.
BERT-BASE (Devlin et al., 2019)	83.0	83.1	83.5	83.4	83.0	83.4	83.2
BERT-LARGE (Devlin et al., 2019)	82.5	84.2	84.9	83.2	83.3	84.0	83.7
RoBERTa-BASE (Liu et al., 2019)	71.9	**91.5**	**93.5**	77.3	80.5	84.6	82.8
RoBERTa-LARGE (Liu et al., 2019)	73.7	88.4	90.6	77.9	80.4	83.8	82.2
Ours w/ Unhindered Sampling	90.4	90.7	90.9	90.5	90.5	90.7	90.6
Ours w/ Bottleneck Sampling	**92.3**	91.1	91.2	**90.9**	**91.7**	**91.1**	**91.6**

Table 3: Classification performance reported on the CMU DoG dataset with the competitive classifiers and our proposed method. Rec.$^+$ represents the recall rate for answerable dialogue contexts, and Rec.$^-$ represents the recall rate for unanswerable dialogue contexts. The precision rate (Pre.), and F1 scores follows the same convention. Acc. represents the overall accuracy.

results indicate that it is not the case. Our method with *Bottleneck Sampling* reports better diversity measurements and BLEU scores than *Unhindered Sampling* on CMU DoG. Indeed, we observe that decoding greedily on the answerability control token gives better accuracy than sampling, which could be the reason for the improved response generation. Still, *Unhindered Sampling* is straightforward to implement, and it reports a better quality in almost all of the metrics than our baselines, and the improvements with *Bottleneck Sampling* are less significant than the improvements in comparing *Unhindered Sampling* with our baselines.

5.3 Fallback Response Generation

Table 2 reports the number of fallbacks generated for answerable and unanswerable dialogue contexts on CMU DoG. As mentioned in Section 3.1, our observation is that the basic E2E model without controlled generation fails to capture the notion of answerability. Our model has a much better FR$^+$ score than our E2E baseline. For the baseline, such a failure in determining the answerability drastically affects the informativeness for answerable dialogue contexts by responding undesirably frequently with fallback. A similar phenomenon can be observed for our comparison model, though the problem is reduced, and the comparison model is better than the E2E baseline at responding with fallback to unanswerable contexts. However, our comparison model still suffers from responding with fallback to answerable contexts, which is undesirable for informative response generation for answerable contexts. In contrast, our method can reduce this problem more effectively and appropriately reply with fallback to unanswerable contexts. Note that the number reported here is not strictly the answerability classification accuracy, as we observed that a fallback response could be generated even with an answerable control token. This aligns with the fact reported in Baheti et al. (2021) that the model can generate an offensive response even with an offensiveness control token.

5.4 Answerability Classification

By prompting the dialogue response generator, our proposed methods can achieve better classification results than an external classifier that introduces extra model parameters as well as the extra classification overheads. As mentioned in Section 4, we are particularly interested in the dataset of CMU DoG, where all the knowledge for negative samples are in-domain movie profiles. This is important, as real-world retrievers are competitive, and we do not want the task to be oversimplified. Fortunately, our competitive classifiers achieve an accuracy of about 82% on the derived dataset. This fact validates that the derived task is with moderate difficulty as there was still space for improvements for the classical classification models.

Table 3 reports the classification results on CMU DoG dataset. Our proposed method has a better score on Rec.$^+$, Pre.$^-$, F1$^+$, F1$^-$ and Acc., which remarkably surpasses all the competitive classifiers. Our model also reports an on-par performance on Rec.$^-$ and Pre.$^+$ with RoBERTa-BASE. This aligns with the fact reported in Section 5.3 that our method can capture more answerable contexts and prevent the model from responding with fallback to them. Consequently, as we report in the main result in Section 5.1, it improves informativeness to the succeeding response generation to answerable contexts. *Unhindered Sampling* reports a bit lower accuracy than *Bottleneck Sampling*. This means that employing greedy decoding is desirable for classification and can improve the answerability classification accuracy. As in Table 1, such an improvement in classification accuracy

Criteria	E2E baseline	*Ours*
Appropriateness	29	71‡
Informativeness	30	70‡
Engagingness	29	71‡
Human-likeness	30	70‡

Table 4: Human evaluation results in winning percentages on CMU DoG. ‡ indicate the results as passing a two-tailed binomial significance test with $p < 0.01$.

Criteria	Mielke et al. (2020)	*Ours*
Appropriateness	43	57†
Informativeness	40	60‡
Engagingness	42	58‡
Human-likeness	43	57†

Table 5: Human evaluation results in winning percentages on CMU DoG. † and ‡ indicate the results as passing a two-tailed binomial significance test with $p < 0.05$ and $p < 0.01$ respectively.

correlates well with the improvements in response generation. In addition, this also aligns with the FR scores reported in Table 2, where *Bottleneck Sampling* has better FR scores than *Unhindered Sampling*. We conclude that our method is better at capturing answerable contexts than our baseline models while still achieving on-par performance on recalling unanswerable contexts and generating fallbacks to them.

5.5 Human Evaluation

We hired three experienced annotators who have degrees relevant to English Linguistics. We present 400 questions with 100 sampled answerable testing instances and ask them to conduct A/B testing. We conduct two sets of the experiment. The first set compares the baseline with our model, and the second set compares the comparison model we built as done in Mielke et al. (2020) and our model. By following previous work (Li et al., 2019; Zou et al., 2021), we adopt the following criteria:

- **(Appropriateness)**: *"Which one is more appropriate given the dialogue context?"*
- **(Informativeness)**: *"Which one presents a more informative and diverse answer?"*
- **(Engagingness)**: *"Which one would you prefer to talk with for a long talk?"*
- **(Human-likeness)**: *"Which one do you think sounds like a real person?"*

Table 4 and Table 5 report the human evaluation results. Our proposed method significantly surpasses our baseline and our comparison model in all of the four quality metrics. This phenomenon is expected and aligns with the fact presented in Section 5.1 which states that the automatic evaluation reports better diversity measurements on the response generation. This also aligns with the fact reported in Section 5.3 and Section 5.4 that the E2E baseline is unaware of the notion of answerability, and our competitive classifier employed for our comparison model has a low Rec.$^+$ on answerable contexts. In contrast, our method solidly improves the overall response quality by appropriately incorporating controlled fallback response generation in an end-to-end manner. Note that we conduct both sets of human evaluation based on our proposed method with *Bottleneck Sampling*.

6 Conclusion

Building a grounded dialogue agent is an important research line. However, most previous works have overlooked the situation when the retrieved knowledge cannot help the agent answer the dialogue. Under such a situation, fallback answers should be appropriately presented, and such incorporation should not degrade the informativeness in responses to answerable contexts. We demonstrate that a standard language model fails to handle this situation well and degrades the informativeness of responses to answerable dialogue contexts. Controlled text generation can be a solution that rigorously replies with fallback to unanswerable contexts. However, naively incorporating controlled text generation still hurts informativeness for the answerable contexts. We propose a novel end-to-end framework that leverages the understanding power of language models for answerability classification that steps into controlled response generation naturally in an autoregressive manner. Our experimental results from both automatic and human evaluation demonstrate that our method achieves higher accuracy on dialogue answerability classification than the competitive models specially designed for language understanding. This improves the informativeness for answerable dialogue contexts while still maintaining the ability to reply with fallback to unanswerable dialogue contexts.

Acknowledgments

This research/paper was supported by the Center for Perceptual and Interactive Intelligence (CPII) Ltd under the Innovation and Technology Commission's InnoHK scheme.

Ethics Statement

This work conducts experiments on the well-known dialogue datasets, and the dataset pre-processing does not make use of any external textual resource. Pre-trained end-to-end dialogue generators using large-scale text corpus are also employed, which might be subjected to offensive contexts and demographic or historical biases buried in the training data. Although the model releasers have attempted their efforts to reduce offensiveness contexts and biases in their training data, the model retains the potential to generate output that triggers offensive replies and might express agreement towards offensive or unethical contexts. The reverse situation also applies, and the model might express disagreement towards ethical contexts. However, due to the fact that current state-of-the-art end-to-end pre-trained dialogue generators or pre-trained language models are mostly trained on large corpus or conversations that naturally occur, the above-mentioned issues are widely known to commonly exist for these models. Either heuristics or neural-based methods are suggested to be employed to post-process the outputs to eliminate any potential ethical issues presented by the models. Finally, we declare that any biases or offensive contexts generated from the model do not reflect the views or values of the authors.

References

Ashutosh Baheti, Maarten Sap, Alan Ritter, and Mark Riedl. 2021. Just say no: Analyzing the stance of neural dialogue generation in offensive contexts. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 4846–4862, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Antoine Bordes, Y-Lan Boureau, and Jason Weston. 2016. Learning End-to-End Goal-Oriented Dialog. *CoRR*, abs/1605.07683.

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 5016–5026, Brussels, Belgium. Association for Computational Linguistics.

Deng Cai, Yan Wang, Wei Bi, Zhaopeng Tu, Xiaojiang Liu, Wai Lam, and Shuming Shi. 2019. Skeleton-to-response: Dialogue generation guided by retrieval memory. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 1219–1228, Minneapolis, Minnesota. Association for Computational Linguistics.

Debanjan Chaudhuri, Md Rashad Al Hasan Rony, Simon Jordan, and Jens Lehmann. 2019. Using a KG-Copy Network for Non-Goal Oriented Dialogues. *CoRR*, abs/1910.07834.

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

Nouha Dziri, Andrea Madotto, Osmar Zaïane, and Avishek Joey Bose. 2021. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2197–2214, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Angela Fan, Mike Lewis, and Yann Dauphin. 2018. Hierarchical neural story generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 889–898, Melbourne, Australia. Association for Computational Linguistics.

Jun Gao, Wei Bi, Xiaojiang Liu, Junhui Li, and Shuming Shi. 2019. Generating multiple diverse responses for short-text conversation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6383–6390.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A. Smith. 2020. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Online. Association for Computational Linguistics.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *International Conference on Learning Representations*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. 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.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016b. 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, Jason Weston, and Stephen Roller. 2019. ACUTE-EVAL: Improved Dialogue Evaluation with Optimized Questions and Multi-turn Comparisons. *CoRR*, abs/1909.03087.

Phillip Lippe, Pengjie Ren, Hinda Haned, Bart Voorn, and Maarten de Rijke. 2020. Diversifying Task-oriented Dialogue Response Generation with Prototype Guided Paraphrasing. *CoRR*, abs/2008.03391.

Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig. 2021a. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *CoRR*, abs/2107.13586.

Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. 2021b. GPT Understands, Too. *CoRR*, abs/2103.10385.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Andrea Madotto, Samuel Cahyawijaya, Genta Indra Winata, Yan Xu, Zihan Liu, Zhaojiang Lin, and Pascale Fung. 2020. Learning knowledge bases with parameters for task-oriented dialogue systems. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2372–2394, Online. Association for Computational Linguistics.

Sabrina J. Mielke, Arthur Szlam, Y-Lan Boureau, and Emily Dinan. 2020. Linguistic calibration through metacognition: aligning dialogue agent responses with expected correctness. *CoRR*, abs/2012.14983.

Alexander Miller, Will Feng, Dhruv Batra, Antoine Bordes, Adam Fisch, Jiasen Lu, Devi Parikh, and Jason Weston. 2017. ParlAI: A dialog research software platform. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 79–84, Copenhagen, Denmark. Association for Computational Linguistics.

Seungwhan Moon, Pararth Shah, Anuj Kumar, and Rajen Subba. 2019. OpenDialKG: Explainable conversational reasoning with attention-based walks over knowledge graphs. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 845–854, Florence, Italy. Association for Computational Linguistics.

Tong Niu and Mohit Bansal. 2018. Polite dialogue generation without parallel data. *Transactions of the Association for Computational Linguistics*, 6(0):373–389.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Shrimai Prabhumoye, Kazuma Hashimoto, Yingbo Zhou, Alan W Black, and Ruslan Salakhutdinov. 2021. Focused attention improves document-grounded generation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4274–4287, Online. Association for Computational Linguistics.

Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *CoRR*, abs/1910.10683.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Ashish Shrivastava, Kaustubh Dhole, Abhinav Bhatt, and Sharvani Raghunath. 2021. Saying No is An Art: Contextualized Fallback Responses for Unanswerable Dialogue Queries. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 87–92, Online. Association for Computational Linguistics.

Zhiwen Tang, Hrishikesh Kulkarni, and Grace Hui Yang. 2021. High-quality dialogue diversification by intermittent short extension ensembles. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1861–1872, Online. Association for Computational Linguistics.

Ashwin K Vijayakumar, Michael Cogswell, Ramprasath R. Selvaraju, Qing Sun, Stefan Lee, David Crandall, and Dhruv Batra. 2016. Diverse Beam Search: Decoding Diverse Solutions from Neural Sequence Models. *CoRR*, abs/1610.02424.

Yida Wang, Yinhe Zheng, Yong Jiang, and Minlie Huang. 2021. Diversifying Dialog Generation via Adaptive Label Smoothing. *CoRR*, abs/2105.14556.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Jing Xu, Da Ju, Margaret Li, Y-Lan Boureau, Jason Weston, and Emily Dinan. 2020. Recipes for Safety in Open-domain Chatbots. *CoRR*, abs/2010.07079.

Yan Xu, Etsuko Ishii, Zihan Liu, Genta Indra Winata, Dan Su, Andrea Madotto, and Pascale Fung. 2021. Retrieval-Free Knowledge-Grounded Dialogue Response Generation with Adapters. *CoRR*, abs/2105.06232.

Tom Young, Erik Cambria, Iti Chaturvedi, Minlie Huang, Hao Zhou, and Subham Biswas. 2017. Augmenting End-to-End Dialog Systems with Commonsense Knowledge. *CoRR*, abs/1709.05453.

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)*, Melbourne, Australia. Association for Computational Linguistics.

Wenxuan Zhang, Yang Deng, Xin Li, Yifei Yuan, Lidong Bing, and Wai Lam. 2021. Aspect sentiment quad prediction as paraphrase generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9209–9219, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2020. DIALOGPT : Large-scale generative pre-training for conversational response generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 270–278, Online. Association for Computational Linguistics.

Kangyan Zhou, Shrimai Prabhumoye, and Alan W Black. 2018. A dataset for document grounded conversations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 708–713, Brussels, Belgium. Association for Computational Linguistics.

Yicheng Zou, Zhihua Liu, Xingwu Hu, and Qi Zhang. 2021. Thinking clearly, talking fast: Concept-guided non-autoregressive generation for open-domain dialogue systems. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2215–2226, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.