COMP5214 Project Milestone: Improving factual faithfulness for prompt-based knowledge-grounded dialog in few-shot settings[*]

Willy Chung, Cheuk Tung Shadow Yiu
(whcchung, ctyiuac)@connect.ust.hk

1 Introduction

Knowledge-grounded dialogue agents are conversational systems that leverages external information provided by the user, such as a supporting document or a webpage, to generate a discussion on the given topic. All state-of-the-art and best performing agents are based on very large language models (LMs) pretrained on considerable amount of text data (Gopalakrishnan et al. (2019), Roller et al. (2021)), that are then further finetuned for the downstream task. By leveraging the internal representation of natural language that those models have learned during pre-training, their ability to generate near-fluent responses became much better. While those dialogue agents seem more natural and human-like in their response, previous work has shown that the generated text is often not faithful to the provided knowledge: in other words those large LMs often tend to hallucinate (Ji et al. (2022), Dziri et al. (2021b)). This raises serious concerns over safety and controllability of such agents as they may answer to questions counterfactually but in very convincing ways.

Previous research has been done for improving the factual faithfulness of such models mainly through a more controlled training setup (Rashkin et al., 2021), or using external knowledge graphs (Dziri et al., 2021a). However, those methods are very expensive both computationally and timewise. In this work, we explore a way to improve goundedness without any further training for the models, but by leveraging the few-shot capability of the larger LMs (Madotto et al., 2021). Recent work has shown that simply tuning the prompt or providing a few in-context examples can generate as good -if not better- predictions (Lester et al. (2021), Reynolds and McDonell (2021)). More specifically, we propose a framework for few-shot learning in knowledge-grounded dialog that aims at improving the factual faithfulness of the answers by automatically crafting prompts that will guide the model in better understanding the concept of what an answer grounded in knowledge is. Our proposed method consists of crafting the few-shot examples using a mix of extracted unanswerable questions and generated answerable questions given a context. The motivation is to provide the model, which already understand natural language to a certain extent, more in-context knowledge about whether a certain information is grounded or not, independently from the fact that the question is well-formulated or not.

For example, given the context "This banana is very sweet" and the question "What color is the banana?", a zero-shot language model will most likely still try to answer the question with "yellow", even though this question is not answerable given the provided context. This happens because the pre-trained model has learned from the large web-crawled data that "bananas are yellow" most of the time (Zhou et al., 2020). Objectively, the language model should not be able to answer this question given the context, and answer "None". This is what we aim to provide as in-context knowledge in the few-shot setting to guide the model to understand knowledge-grounded answers: provide extracted unanswerable questions that are still relevant to the given context, and telling the model objectively that he should not be able to answer them, while contrasting this with answerable questions.

2 Problem Statement

2.1 Dataset

We need two datasets for this project, one for evaluation and the other one for crafting the few-shot examples. To be coherent with our motivation claim, we have to exclude datasets that have been used during the pre-training of T0 to avoid any bias.

[*] Not final.

Topical-Chat (Gopalakrishnan et al., 2019) Topical-chat is a knowledge-grounded dialogue dataset with approximately 10k human to human conversations of 20 turns on average. This dataset has been originally built to provide both knowledge-grounded and open-domain conversation within the same framework, as most of the knowledge-grounded dialogue dataset usually set explicit roles for the speakers. Because of the input length soft limit of seq2seq models such as T0 (512 input token), we choose to use the shorter conversations of the dataset so that we can provide enough few-shot examples in the input. We will directly test our method on the test set of topical-chat in the few-shot settings.

QuAC (Choi et al., 2018) For the QA dataset from which we will extract unanswerable question and generate answerable questions, we chose QuAC as it has not been used for pre-training T0, and also have a good amount of unanswerable question with provided context. As for Topical-Chat, we will extract the shortest context-question pairs to not overload the input length. We are currently looking if there is a better suited dataset for this part.

2.2 Evaluation

Baseline We aim to run multiple baseline to compare the performance of our proposed method depending on the time and resources at disposition. At the very least we will compare our result with the zero-shot setting to show that the few-shot example given are impactful. We aim to run multiple experiments while modifying two parameters: the ratio of unanswerable/answerable question to provide for each example in the few-shot, and the number of total examples to use for the few-shot as long as it is supported by the input length limit. If time allows, we will also test our method in the BEGIN benchmark (Dziri et al., 2021b), but we are skeptical because this benchmark leverages a classification task to evaluate groundedness, which might not be suitable for seq2seq models such as T0.

Metrics As for the evaluation metrics, we choose to use perplexity, F1 score, BLEU and Q^2. As the knowledge grounded dialogue task is still a natural language generation task, we decided to include perplexity in the evaluation metric. We highly doubt that the few-shot settings will actually improve perplexity as it does not fundamentally impact the model's internal representation of natural language, but only the sequence output of the given model. F1 score and BLEU are chosen because we want to assure that our proposed method does not impact the performance of the generation itself, while providing an improvement in factual faithfulness evaluated through the Q^2 metric. Q^2 (Honovich et al., 2021) is a recently proposed metric used to evaluate factual consistency in knowledge-grounded dialogues using question generation, question answering and natural language inference that shows high correlation with human annotations of such conversations.

2.3 Expected results

If this method work as intended, we expect to see a lift of performance in Q^2 score that correlates with the number of few-shot example we provide in the input sequence. From zero-shot to three-shot for example, we expect the Q^2 score to gradually increase. However, we do not expect the perplexity to change much as explained in 2.2, and also hope to not decrease the BLEU and F1 score while we provide more few-shot example which is a possible downside when increasing the sparsity of the information in the input prompt (Reynolds and McDonell, 2021).

3 Technical Approach

3.1 Language Model

We chose to use T0 (Sanh et al., 2021) as the pre-trained language model to run our experiments. T0 is seq2seq language model based on T5 (Raffel et al., 2020) to assess zero-shot generalization capability of very large language models. During the pre-training of T0, a considerable amount of natural language tasks have been mapped to a prompted form. Both the prompting and the zero-shot generalization makes T0 a good model to use in the context of this work. However, due to resource limitation, we will use the smaller version of T0 with 3 billion parameters instead of the original one with 11 billion parameters.

3.2 Prompt adaptation

From the QA dataset used to craft the prompt, we define our k-shot examples data $\mathcal{F} = \{(C^k_{QA}, Q^k_{unans}, Q^k_{ans,i}, A^k_i)\}^3_{k=1}$ with C^k_{QA} the context, Q^k_{unans} the extracted unanswerable question, $Q^k_{ans,i}, A^k_i$ the i^{th} generated answerable

<table>
<tr><th colspan="3">Prompt</th></tr>
<tr><td colspan="3">Given the context "Bananas are very sweet, but they can be the cause of tropical diseases ." and the conversation:
" Amy: I like bananas . Chris: Why? Amy: Because they are very sweet! Chris: I don’t like bananas. Amy: Why?" **<TASK>**.</td></tr>
<tr><th><TASK></th><th>Generated output</th><th>Factual Faithfulness</th></tr>
<tr><td>write Chris’ answer.</td><td>Amy doesn’t like bananas because they can be the cause of tropical diseases.</td><td>Hallucination</td></tr>
<tr><td>write a follow up conversation.</td><td>Amy likes bananas because they are very sweet. Chris doesn’t like bananas.</td><td>Neutral</td></tr>
<tr><td>what did Chris say?</td><td>Chris doesn’t like bananas because they can be the cause of tropical diseases.</td><td>Knowledge-Grounded</td></tr>
</table>

Table 1: Case illustration of various output using T0_3B given the same prompt, but with different task formulation. Just by changing the task formulation, the output can greatly vary in groundedness

question-answer pair for example k. We also define our knowledge-grounded dialogue data $\mathcal{D} = \{C_d, D_d, G_d\}$ where C_d is the context of the dialogue d, D_d the dialogue and G_d the generated dialogue turn. Our prompt adaptation consists of building $\mathcal{F}$ from the QA dataset, and provide the final input sequence for the model as $\{\mathcal{F}, \mathcal{D}\}$.

3.3 Question generation

We are still working on this part, and exploring a few pre-built pipeline for question generation.

3.4 Preliminary result

Prompt formulation importance The overall prompt formulation follows the prompt design that T0 (Sanh et al., 2021) has been pre-trained with. However, we can see from some examples that the task formulation is crucial for factual faithfulness. Depending on how we formulate the task, the output changes drastically while retaining its fluency in natural language. As shown in table 1, while all three task formulations make sense to a human, we can see that the first one leads to a counterfactual output, the second one leads to a neutral generation and the last one is a knowledge-grounded answer, where the knowledge here consists of the context and the past conversation. This strengthen our claim that the prompt design can easily impact the factual faithfulness of knowledge-grounded dialogue, and provides more motivation for our proposed method. We are currently working on the pipeline to craft the whole prompt instead of showing a few hand-crafted examples, and testing the Q^2 metric as it is not very straight-forward to use.

References

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wentau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. Quac: Question answering in context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2174–2184.

Nouha Dziri, Andrea Madotto, Osmar Zaiane, and Avishek Joey Bose. 2021a. Neural path hunter: Reducing hallucination in dialogue systems via path grounding. *arXiv preprint arXiv:2104.08455*.

Nouha Dziri, Hannah Rashkin, Tal Linzen, and David Reitter. 2021b. Evaluating groundedness in dialogue systems: The begin benchmark. *arXiv preprint arXiv:2105.00071*.

Karthik Gopalakrishnan, Behnam Hedayatnia, Qinlang Chen, Anna Gottardi, Sanjeev Kwatra, Anu Venkatesh, Raefer Gabriel, and Dilek Hakkani-Tür. 2019. Topical-chat: Towards knowledge-grounded open-domain conversations. *Proc. Interspeech 2019*, pages 1891–1895.

Or Honovich, Leshem Choshen, Roee Aharoni, Ella Neeman, Idan Szpektor, and Omri Abend. 2021. Q2:: Evaluating factual consistency in knowledge-grounded dialogues via question generation and question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7856–7870.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Yejin Bang, Andrea Madotto, and Pascale Fung. 2022. Survey of hallucination in natural language generation. *arXiv preprint arXiv:2202.03629*.

Brian Lester, Rami Al-Rfou, and Noah Constant. 2021. The power of scale for parameter-efficient prompt tuning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3045–3059.

Andrea Madotto, Zhaojiang Lin, Genta Indra Winata, and Pascale Fung. 2021. Few-shot bot: Prompt-based learning for dialogue systems. *arXiv preprint arXiv:2110.08118*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67.

Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. 2021. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 704–718.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems*, pages 1–7.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, et al. 2021. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325.

Victor Sanh, Albert Webson, Colin Raffel, Stephen H Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Teven Le Scao, Arun Raja, et al. 2021. Multitask prompted training enables zero-shot task generalization. *arXiv preprint arXiv:2110.08207*.

Xuhui Zhou, Yue Zhang, Leyang Cui, and Dandan Huang. 2020. Evaluating commonsense in pre-trained language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9733–9740.