COMP5214 Proposal: 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 Proposal

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 in 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. To do that, we will craft the few-shot examples carefully by using a mix of one extracted unanswerable question and two generated answerable questions according to the given context.

We will use T0_3B (Sanh et al., 2021), a version of T5 (Raffel et al., 2020) with 3 billion parameters which has been trained to do zero-shot predictions. We will use examples from QA datasets containing unanswerable questions to craft the prompts, such as SQuAD 2.0 (Rajpurkar et al., 2018), TidyQA (Clark et al., 2020) or QuAC (Choi et al., 2018). As T0 has been pre-trained on the first two datasets, we might focus more on QuAC to avoid any biases the model might have picked up during pre-training. In our evaluation, we will focus on two points: if the answer is correct with BLEU score (Papineni et al., 2002), as we want to ensure that our method does not damage the performance of the model; and if the answer is more grounded in knowledge by using the Q^2 metric (Honovich et al., 2021), a recently proposed metric using question generation and question answering to evaluate the factual faithfulness of a given answer. If we have enough time, we would like to try with one more generative model such as GPT-Neo (Black et al., 2021), and add human evaluation.

*Not final.

References

Sid Black, Gao Leo, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau 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.

Jonathan H Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. Tydi qa: A benchmark for information-seeking question answering in typologically diverse languages. *Transactions of the Association for Computational Linguistics*, 8:454–470.

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*.

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*, pages 311–318.

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.

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.

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*.