Sicong Huang

University of California, Santa Cruz
Silicon Valley Campus
Santa Clara, CA
United States

shuan213@ucsc.edu

1 Research interests

My primary research interests lie in **evaluating and improving the faithfulness of language model-based text generation systems**. Recent advances in large language models (LLMs) such as GPT-4 (OpenAI et al., 2024) and Llama (Touvron et al., 2023) have enabled the wide adoption of LLMs in various aspects of natural language processing (NLP). Despite their widespread use, LLMs still suffer from the problem of hallucination (Huang et al., 2023), limiting the practicality of deploying such systems in use cases where being factual and faithful is of critical importance. My research specifically aims to evaluate and improve the faithfulness, i.e. *the factual alignment between the generated text and a given context*, of text generation systems. By developing techniques to reliably **evaluate**, **label**, and **improve** generation faithfulness, we can enable wider adoption of dialog systems that need to converse with human users using accurate information.

1.1 Evaluating the Faithfulness of Dialog Summarization Systems

Evaluating generated text is often considered a task that is as difficult as generating text per se. Besides gold-standard human evaluation, long-standing automatic metrics such as ROUGE (Lin, 2004) and BLEU (Papineni et al., 2002) have been shown to poorly correlate with human judgements in evaluating faithfulness (Maynez et al., 2020). More recent LM-based metrics such as CTC (Deng et al., 2021) and BARTScore (Yuan et al., 2021) that are designed to target faithfulness exhibit higher correlation with human judgements but often do not account for the types of errors dialog summaries tend to make, resulting in lower performance when evaluating summaries in the dialog domain. To address this issue, I developed technique to improve upon the existing BARTScore metric, tailoring them specifically to account for the unique challenges of dialog summarization, such as colloquial speech, ellipses, and coreference errors.

In-domain Fine-tuning As highlighted in Huang et al. (2022), metrics that perform well on news summarization often fail to transfer effectively to dialog summarization. I investigated techniques to adapt BARTScore to the dialog domain by (1) fine-tuning on other dialog data along with automatically generated summaries (2) fine-tuning on the training set of the evaluation data directly. Results show both approaches can improve metric performance on dialog summaries with fine-tuning directly on the evaluation data being the most effective.

Learning from Synthetic Negative Samples To capture the common types of errors that come with dialog summaries, I investigated methods to generate negative sample summaries that reflect common error types in dialog summaries (Tang et al., 2022), such as entity swapping, mask and regenerate, and totally irrelevant hallucination. Then I applied unlikelihood training to the negative samples, which minimized the probability of generating negative tokens, thus assigning lower score to unfaithful summaries. Results show learning from negative samples can further improve BARTScore's correlation with human judgements, and using all error types yields the highest performance gain.

Through this research, I developed techniques to improve the performance of BARTScore (a high performing metric at the time) at evaluating dialog summaries, enabling more reliable assessment of dialog summarization systems.

1.2 Improving the Faithfulness of Abstractive Summarization Systems

In this research, I attempted at improving summarization faithfulness by investigating methods to properly leverage span-level hallucination information.

Span-level Hallucination Labeling To identify the spans of text that contain hallucinated information, I leveraged GPT-4 as an automatic labeler. I created a dataset of summaries with span-level hallucination annotations by prompting GPT-4 to label information in generated summaries that is inconsistent to the source document.

Comparison of Training Methods This research compared different training approaches that can leverage negative samples to reduce unfaithfulness, including *gra-*

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 37–39

dient ascent (Yao et al., 2024), *unlikelihood training* (Welleck et al., 2020), and *task vector negation* (Ilharco et al., 2023). The results indicate that unlikelihood training is particularly effective in reducing unfaithful information in LLM-generated summaries. The reduction of hallucinated content is also confirmed by human annotations on a subset of generated summaries.

Through this research, I found an effective method to improve summary faithfulness, i.e. span-level annotation and unlikelihood training, and the improvement is consistent across both news and dialog domain. These findings pave ways to reduce hallucinations in text generation more generally.

1.3 Fine-grained Annotation of Generated Text

As Section 1.2 has shown that span-level hallucination annotation can provide valuable information that can be leveraged to improve summary faithfulness, obtaining reliable span-level annotation becomes a critical step in improving faithfulness. Moreover, most text generation metrics only provide scalar value scores, revealing no information on the reasoning and the part of the text that resulted in such scores. Due to the uninterpretability of these metrics, they provide little guidance on how to improve the generated text. Thus, I am currently looking into developing a metric that is based on span-labeling, providing not only scores but also the reasons that resulted in the scores. I believe that interpretability is a crucial feature in the next generation of evaluation metrics.

2 Spoken dialogue system (SDS) research

The field of Spoken Dialog Systems (SDS) research is poised for significant advancements in the coming years, driven by the rapid progress in large language models and the increasing demand for more natural and reliable human-computer interactions. In my opinion, developing more context-aware and factually consistent systems along with reliable evaluation metrics should be important themes of SDS research for the next 5 to 10 years. Advancements in these areas will enable wider adoption of SDS in various scenarios:

- Healthcare: Assisting in patient triage, mental health support, and chronic disease management.
- Education: Providing personalized tutoring and language learning assistance.
- Customer Service: Handling complex queries and providing more empathetic interactions.

Our generation of young researchers has the potential to make significant contributions in several areas:

- Developing highly faithful and contextually appropriate response generation techniques.
- Creating more robust and interpretable evaluation metrics that can reliably assess system outputs of desired quality, such as coherence, engagingness, and faithfulness.
- Improving the handling of ambiguity and implicit information in conversations.

To achieve these goals, we may need to answer some key questions:

- How can we effectively combine the strengths of rule-based systems with the flexibility of neural approaches to achieve faithful responses?
- How can we make SDS and its evaluation metrics more interpretable?
- What are the best ways to incorporate real-world knowledge and common sense reasoning into SDS?

As we advance in this field, it will also be important to address challenges related to privacy, bias mitigation, and maintaining the balance between automation and human oversight. The goal should be to create SDS that not only understand and respond accurately but also enhance human capabilities and improve quality of life across diverse user groups and applications.

3 Suggested topics for discussion

- Are autoregressive LLMs limited by their token-by-token nature, thus unable to plan their outputs and produce fully faithful generations?
- What does it mean for a language processing system to "understand" language?
- Bisk et al. (2020); Bender and Koller (2020) have suggested that training on predicting the next word alone is unable to capture meaning which requires grounding. Are vision-language models (or LLMs trained on more modalities of data) enough to capture meaning? Or are symbolic representations required?

References

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, pages 5185–5198. https://doi.org/10.18653/v1/2020.acl-main.463.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, pages 8718–8735. https://doi.org/10.18653/v1/2020.emnlp-main.703.

Mingkai Deng, Bowen Tan, Zhengzhong Liu, Eric Xing, and Zhiting Hu. 2021. Compression, transduction, and creation: A unified framework for evaluating natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pages 7580–7605. https://doi.org/10.18653/v1/2021.emnlp-main.599.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions.

Sicong Huang, Asli Celikyilmaz, and Haoran Li. 2022. Ed-faith: Evaluating dialogue summarization on faithfulness. https://arxiv.org/abs/2211.08464.

Gabriel Ilharco, Marco Tulio Ribeiro, Mitchell Wortsman, Ludwig Schmidt, Hannaneh Hajishirzi, and Ali Farhadi. 2023. Editing models with task arithmetic. In *The Eleventh International Conference on Learning Representations*. https://openreview.net/forum?id=6t0Kwf8-jrj.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, pages 74–81. https://aclanthology.org/W04-1013.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, pages 1906–1919. https://doi.org/10.18653/v1/2020.acl-main.173.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, and Ilge Akkaya et al. 2024. Gpt-4 technical report.

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*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, pages 311–318. https://doi.org/10.3115/1073083.1073135.

Xiangru Tang, Arjun Nair, Borui Wang, Bingyao Wang, Jai Desai, Aaron Wade, Haoran Li, Asli Celikyilmaz, Yashar Mehdad, and Dragomir Radev. 2022. CONFIT: Toward faithful dialogue summarization with linguistically-informed contrastive fine-tuning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, Seattle, United States, pages 5657–5668. https://doi.org/10.18653/v1/2022.naacl-main.415.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, and Yasmine Babaei et al. 2023. Llama 2: Open foundation and fine-tuned chat models.

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

Yuanshun Yao, Xiaojun Xu, and Yang Liu. 2024. Large language model unlearning.

Weizhe Yuan, Graham Neubig, and Pengfei Liu. 2021. BARTScore: Evaluating generated text as text generation. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*. https://openreview.net/forum?id=5Ya8PbvpZ9.

Biographical sketch

Sicong Huang is a first-year Ph.D. student at University of California, Santa Cruz advised by Prof. Ian Lane. He completed an AI Residency at Meta where he devised ways to improve summarization metrics' reliability. Prior to this, he did a Master of Science in Computational Linguistics at the University of Washington, during which he interned at a startup company, Seasalt AI, where he built and deployed a meeting summarization service in production. Before this, he completed a Bachelors in Electrical and Computer Engineering also at University of Washington. During his undergraduate study, for 6 months, he was an exchange student at Tokyo Institute of Technology where he started learning and researching about natural language processing under Prof. Manabu Okumura.