Toward User-adaptive, Accountable and Privacy-Preserving Dialog Systems

My goal is to build dialog systems that can responsibly engage users without privacy leakage. To achieve this goal, I propose to equip the dialog systems with *user adaptation*, *accountable dialog generation*, and *privacy protection mechanisms*.

- **User Adaptation**. Conversation is not a solo dance. To create more engaging user experiences, I design dialog systems that can adapt to individual users by incorporating multi-modal information such as sentiment [4], and personalized conversational strategies [5]. Our study laid the ground for developing a personalized persuasive dialog system and won the *best paper nomination* at ACL 2019 [5].
- **Accountable Dialog Generation**. Current dialog systems still suffer from long-standing generation errors such as repetition, contradiction, and other types of nonsense. My work applied RL [3] to refine the language model without user simulators and distill sentence-level information about repetition, inconsistency, and task relevance through rewards. Moreover, to better accomplish complex tasks like persuasion, the model learns from human demonstration to imitate human persuasion behaviors and selects the most persuasive responses.
- **Privacy Protection**. With more dialog agents such as Apple Siri and Amazon Alexa deployed in daily life, how to protect user privacy while building high-utility models becomes increasingly important. To tackle this challenge, we proposed a new differential privacy notion–*selective differential privacy* (SDP)–to protect the sensitive portion of the data [1, 2], and developed effective mechanisms to build SDP-protected NLP models. With SDP, we march one step closer toward more privacy-preserving and high-performing language models and hope to inspire more research in this direction in the NLP community.

To conclude, I build effective dialog systems to address real-world problems. Meanwhile, many of the proposed methods can be generalized to tasks beyond NLP. As described above, my research is highly interdisciplinary and connects different fields such as social science, human-computer interaction, machine learning, and computer security. I am a strong advocator of interdisciplinary research and am always excited to collaborate with researchers from different fields. For future works, I am enthusiastic about the following topics:

Learning through Interactions. One key property of dialog systems is that they are not static: the deployed systems need to interact with users and then self-evolve based on the feedback in a timely fashion. This involves two steps: 1) updating the model offline with the collected feedback; and 2) adjusting the model online during interactions. I plan to utilize offline RL to update models efficiently with limited feedback data, and online RL to guide the dialog trajectory in real-time, and interleave these two steps toward a system with self-evolving abilities.

Ethical Dialog systems. It has always been my biggest passion to build dialog agents for social good, e.g., to educate the next generations, offer counseling for patients, provide companions for elderly people, and so on. To build such systems, we need to first understand the human-computer interaction pattern of different groups of people. However, in our persuasion study, only 2.1% (42 out of 2034) participants are senior citizens (aged 65+) as they are less exposed to computers and technologies. Systems developed with a biased population will only reinforce this phenomenon. As a researcher, I sincerely hope to take on social responsibility to break the technology barriers and develop conversational agents beneficial for everyone in society.

References

[1] Weiyan Shi, Si Chen, Chiyuan Zhang, Ruoxi Jia, and Zhou Yu. Just fine-tune twice: Selective differential privacy for large language models. *arXiv preprint arXiv:2204.07667*, 2022. URL: https://arxiv.org/abs/2204.07667.

[2] Weiyan Shi, Aiqi Cui, Evan Li, Ruoxi Jia, and Zhou Yu. Selective differential privacy for language modeling. In *NAACL*, 2022. URL: https://arxiv.org/abs/2108.12944.

[3] Weiyan Shi, Yu Li, Saurav Sahay, and Zhou Yu. Refine and imitate: Reducing repetition and inconsistency in persuasion dialogues via reinforcement learning and human demonstration. *EMNLP Findings*, 2021. URL: https://aclanthology.org/2021.findings-emnlp.295/.

[4] Weiyan Shi and Zhou Yu. Sentiment adaptive end-to-end dialog systems. *NAACL*, 2018. URL: https://aclanthology.org/P18-1140/.

[5] Xuewei Wang*, Weiyan Shi*, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. *ACL*, 2019. URL: https://aclanthology.org/P19-1566/.