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

Weiyan Shi, Columbia University

I aim to build ML-based dialog systems that can engage users reliably without privacy leakage: imagine Alice talking to her personal trainer bot "I exercised on Monday at ABCD. too tired for the gym today", the bot should reply "I know you've worked hard lately but you are not a quitter!" (✓), instead of saying "you didn't exercise on Monday so you should today" (✗), or memorizing the unnecessary detail on Alice's location on Monday (✗). This requires the system to produce sensible responses, and adapt to the user over time while protecting their privacy.

- **Accountable Generation**. Conversation has to first make sense. But existing chatbots still repeat or contradict themselves. My work proposed DialGAIL [3] to let the system first learn from its own mistakes through RL, and then imitate human persuasion strategies via imitation learning. Such systems achieved an average donation of $0.62 out of $2, an 70% increase over human persuaders, and raised $2000+ donations to a children's charity.
- **User Adaptation**. Besides, conversation is not a solo dance. To create engaging user experiences, I design dialog agents that can adapt to individual users by incorporating multi-modal information such as sentiment, and personality. Our work laid the ground for a personalized persuasive dialog system and won the *best paper nomination* at ACL 2019 [4].
- **Privacy Protection**. Conversation is also private. To train privacy-preserving dialog models and mitigate the low-utility issue caused by traditional privacy notions, we proposed a new differential privacy notion–*selective differential privacy* (SDP)–to protect sensitive information specified by a secret detector in the data, and developed effective privacy mechanisms to achieve SDP-protected models [2]. It is the first work to introduce a suitable privacy notion to the NLP community. We are also the first to propose techniques to mitigate the leakage caused by imperfect secret detectors and theoretically analyze the noise needed for imperfect detectors [1].

To conclude, I develop effective and safe dialog systems. As described above, my research is highly interdisciplinary and connects ML with various fields such as social science, HCI, and computer security. I am always excited to collaborate with researchers from different fields. For future work, I am passionate about these topics:

- **Ethical Dialog Systems.** First, conversational agents have to be well-intentioned. My biggest passion is to build systems for marginalized groups in this technology-driven world, e.g., to accompany the elderly, educate the young, and counsel the ones in need. In our study, only 2.1% participants are senior citizens. As a researcher, my first step toward ethical dialog systems is to invite different underrepresented groups for studies and understand their HCI patterns to break the technology barriers for them.
- **Learning through Interactions.** Conversational agents are also dynamic: they should interact with users and evolve given human feedback in a timely fashion. This involves three steps: 1) updating the model offline with collected feedback via offline RL; 2) adjusting the dialogue trajectory online given real-time reactions via offline RL; 3) interleaving these two steps toward a system that can continuously improve itself.

References

[1] Weiyan Shi, S. Chen, C. Zhang, R. Jia, and Z. Yu. Just fine-tune twice: Selective differential privacy for large language models. *arXiv preprint arXiv:2204.07667*, 2022.

[2] Weiyan Shi, A. Cui, E. Li, R. Jia, and Z. Yu. Selective differential privacy for language modeling. In *NAACL*, 2022.

[3] Weiyan Shi, Y. Li, S. Sahay, and Z. Yu. Refine and imitate: Reducing repetition and inconsistency in persuasion dialogues via reinforcement learning and human demonstration. *EMNLP Findings*, 2021.

[4] X. Wang*, Weiyan Shi*, R. Kim, Y. Oh, S. Yang, J. Zhang, and Z. Yu. Persuasion for good: Towards a personalized persuasive dialogue system for social good. *ACL*, 2019.