Shutong Feng

Heinrich Heine University Düsseldorf
Universitätsstraße 1
40225 Düsseldorf
Germany

shutong.feng@hhu.de
https://shutongfeng.github.io/

1 Research interests

My research interests lie in the area of **modelling affective behaviours of interlocutors in conversations**. In particular, I look at emotion perception, expression, and management in information-retrieval task-oriented dialogue (ToD) systems. Traditionally, ToD systems focus primarily on fulfilling the user's goal by requesting and providing appropriate information. Yet, in real life, the user's emotional experience also contributes to the overall satisfaction. This requires the system's ability to recognise, manage, and express emotions. To this end, I incorporated emotion in the entire ToD system pipeline (Feng et al., 2024). In addition, in the era of large language models (LLMs), emotion recognition and generation have been made easy even under a zero-shot set-up (Feng et al., 2023b; Stricker and Paroubek, 2024). Therefore, I am also interested in building ToD systems with LLMs and examining various types of affect in other ToD set-ups such as depression detection in clinical consultations and user confidence estimation in tutoring systems (Litman et al., 2009).

1.1 Emotion-aware ToD System

While existing works have explored user emotions or similar concepts in various ToD modelling tasks (Lukin et al., 2017; Guo et al., 2024), none has so far combined these emotional aspects into a fully-fledged dialogue system nor conducted interaction with human or simulated users. Therefore, I propose to incorporate emotion into the complete ToD interaction process, involving understanding, management, and generation.

To achieve this, I first extended the EmoWOZ dataset (Feng et al., 2022) with system emotion labels. With this ToD dataset containing both user and system emotion labels, I could train a both emotionally and semantically conditioned natural language generator, as well as an emotional user simulator (Lin et al., 2023) that both reacts to system emotion and expresses user emotions. Leveraging off-the-shelf dialogue state tracker (van Niekerk et al., 2021) and user emotion recogniser (Feng et al., 2023a), I set up the system around a dialogue policy (Geishauser et al., 2022), which takes dialogue state extended with user emotion as input and outputs action including system emotions. The policy was optimised via reinforcement learning (RL) with the emotional user simulator on the language level. For the reward signal, the policy considered both task success and user sentiment level.

In addition to the above-mentioned modular ToD system, I also took the inspiration from an existing LLM-based end-to-end system (Stricker and Paroubek, 2024). I extended the system to output emotional actions and trained it with the newly collected dataset.

With both systems, I conducted corpus-level evaluation and interactive evaluation with both simulated and real users. Our results show that incorporating emotion into the full ToD pipeline can effectively enhance the user's emotional experience and task success at the same time. This aligns with our hypothesis and intuition that emotion is crucial in ToD systems. I believe this points to a promising direction on improving ToD systems.

The future work would be to combine the advantages of modular systems and end-to-end systems, specifically by incorporating RL with human feedback (RLHF) to LLM-based end-to-end systems. Modular systems are usually centred around a dialogue policy optimised via RL for long-term task success. Yet, they are prone to errors from each small modules. End-to-end models, on the other hand, can leverage the capacity of large pre-trained models but existing models are trained on the corpus with supervised learning. This usually leads to sub-optimal performance in interactive evaluation. Incorporating RLHF in the training could potentially be a solution and further boost the performance of end-to-end ToD systems. Efficient acquirement of response preference labels and RL training will be my next research efforts.

1.2 Recognising Affect using LLMs

I am also interested in how LLMs can be used to recognise user affects in conversations. My goal was not to build state-of-the-art affect recognition models with LLMs but rather to understand the potential of current LLMs under vanilla set-ups for such a purpose. Specifically, I conducted experiments with a set of LLMs on different types of datasets under an array of prompt-based training set-ups. For datasets, I examined three differ-

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 81–83
September 16–17, 2024.

ent types of affects: emotions in ToDs, emotions in chit-chat, and depression. For training set-ups, I looked at zero-shot learning, few-shot in-context learning, and supervised learning with different amount of data. I also considered LLMs as a text-processing back-end in SDS by investigating how automatic speech recognition errors could influence model prediction. With experimental results, I draw insights on LLMs' zero and few-shot ICL ability, data efficiency in task-specific fine-tuning, ability to handle long input sequence, ability to recognise different types of affects, robustness to ASR errors, and so on.

In the future, I will look at how affect recognition and generation can be improved under zero or few-shot setups. I will leverage existing resources such as annotator confusion and annotation schemes to elicit reliable reasoning and uncertainty estimation in LLMs.

2 Spoken dialogue system (SDS) research

The emergence of LLMs has great impact on approaches in spoken dialogue modelling. They also bring about opportunities in areas such as unsupervised ontology construction for system design (Vukovic et al., 2024). While LLMs have demonstrated promising abilities in general language modelling tasks and chat applications, smaller models and established modular system set-ups should not be overlooked. Therefore, instead of wishfully using LLMs to replace all SDSs, researchers will understand more about the limitations of LLMs so as to combine the strengths of LLMs and traditional methods.

There will also be more diverse requirements and evaluation criteria for SDSs. In the past, information-retrieval ToD systems focus primarily on task success and inform rate, and chit-chat systems focus on engagement, coherence, and naturalness. As we see more about what more powerful systems can achieve nowadays, we expect more from the system: safety, trust-worthiness, bias, emotion consistency, and many more. We may also expect our dialogue agents to be able to adapt to different challenging scenarios, from out-of-domain requests to cultural shifts. While we see more exciting research opportunities and directions, challenges such as the evaluation of more well-rounded SDSs emerge.

3 Suggested topics for discussion

- **Controllability of LLMs as Dialogue System Back-end:** The issue of hallucination can be especially detrimental in the domain of task-oriented dialogues and in the presence of an ontology and database. How should we make LLMs more controllable for SDS applications?
- **The Future of LLMs**: What ability would the next generation of LLMs have? What would be possible directions of the development in NLP?
- **Affective SDS:** What are risks of building SDSs for affect-related applications, such as emotion support, mental health counseling, more human-like personal assistant, etc.?

References

Shutong Feng, Hsien chin Lin, Christian Geishauser, Nurul Lubis, Carel van Niekerk, Michael Heck, Benjamin Ruppik, Renato Vukovic, and Milica Gašić. 2024. Infusing emotions into task-oriented dialogue systems: Understanding, management, and generation. https://arxiv.org/abs/2408.02417.

Shutong Feng, Nurul Lubis, Christian Geishauser, Hsien-chin Lin, Michael Heck, Carel van Niekerk, and Milica Gasic. 2022. EmoWOZ: A large-scale corpus and labelling scheme for emotion recognition in task-oriented dialogue systems. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, pages 4096–4113. https://aclanthology.org/2022.lrec-1.436.

Shutong Feng, Nurul Lubis, Benjamin Ruppik, Christian Geishauser, Michael Heck, Hsien-chin Lin, Carel van Niekerk, Renato Vukovic, and Milica Gasic. 2023a. From chatter to matter: Addressing critical steps of emotion recognition learning in task-oriented dialogue. In Svetlana Stoyanchev, Shafiq Joty, David Schlangen, Ondrej Dusek, Casey Kennington, and Malihe Alikhani, editors, *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, Prague, Czechia, pages 85–103. https://doi.org/10.18653/v1/2023.sigdial-1.8.

Shutong Feng, Guangzhi Sun, Nurul Lubis, Chao Zhang, and Milica Gašić. 2023b. Affect recognition in conversations using large language models. https://arxiv.org/abs/2309.12881.

Christian Geishauser, Carel van Niekerk, Hsien-chin Lin, Nurul Lubis, Michael Heck, Shutong Feng, and Milica Gašić. 2022. Dynamic dialogue policy for continual reinforcement learning. In Nicoletta Calzolari, Chu-Ren Huang, Hansaem Kim, James Pustejovsky, Leo Wanner, Key-Sun Choi, Pum-Mo Ryu, Hsin-Hsi Chen, Lucia Donatelli, Heng Ji, Sadao Kurohashi, Patrizia Paggio, Nianwen Xue, Seokhwan Kim, Younggyun Hahm, Zhong He, Tony Kyungil Lee, Enrico Santus, Francis Bond, and Seung-Hoon Na, editors, *Proceedings of the 29th International Conference on Computational Linguistics*. International Committee on Computational Linguistics, Gyeongju, Republic of Korea, pages 266–284. https://aclanthology.org/2022.coling-1.21.

Ao Guo, Ryu Hirai, Atsumoto Ohashi, Yuya Chiba, Yuiko Tsunomori, and Ryuichiro Higashinaka. 2024. Personality prediction from task-oriented and open-domain human–machine dialogues. *Scientific Reports* 14(1):3868.

Hsien-Chin Lin, Shutong Feng, Christian Geishauser, Nurul Lubis, Carel van Niekerk, Michael Heck, Benjamin Ruppik, Renato Vukovic, and Milica Gašić. 2023. EmoUS: Simulating user emotions in task-oriented dialogues. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*. Association for Computing Machinery, New York, NY, USA, SIGIR '23, page 2526–2531. https://doi.org/10.1145/3539618.3592092.

Diane J. Litman, Mihai Rotaru, and Greg Nicholas. 2009. Classifying turn-level uncertainty using word-level prosody. In *Interspeech*. https://api.semanticscholar.org/CorpusID:279660.

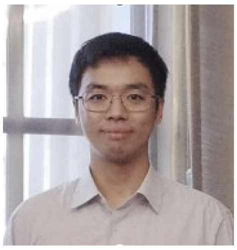
Stephanie Lukin, Pranav Anand, Marilyn Walker, and Steve Whittaker. 2017. Argument strength is in the eye of the beholder: Audience effects in persuasion. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*. pages 742–753.

Armand Stricker and Patrick Paroubek. 2024. A Unified Approach to Emotion Detection and Task-Oriented Dialogue Modeling. In *IWSDS*. Sapporo (Japon), Japan. https://hal.science/hal-04415809.

Carel van Niekerk, Andrey Malinin, Christian Geishauser, Michael Heck, Hsien-chin Lin, Nurul Lubis, Shutong Feng, and Milica Gasic. 2021. Uncertainty measures in neural belief tracking and the effects on dialogue policy performance. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, pages 7901–7914. https://doi.org/10.18653/v1/2021.emnlp-main.623.

Renato Vukovic, David Arps, Carel van Niekerk, Benjamin Matthias Ruppik, Hsien-Chin Lin, Michael Heck, and Milica Gašić. 2024. Dialogue ontology relation extraction via constrained chain-of-thought decoding. https://arxiv.org/abs/2408.02361.

Biographical sketch



Shutong Feng is a final-year PhD student at the Chair for Dialog System and Machine Learning, Heinrich Heine University Düsseldorf, Germany. He is supervised by Prof. Dr. Milica Gašić and co-supervised by Dr. Nurul Lubis. He is interested in modelling human affect in spoken dialogue systems. Shutong obtained his BA and MEng degrees from the University of Cambridge in 2019. He then worked as an engineer at Huawei Technologies Co. Ltd. before starting his PhD study in 2020.