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 natural and human-like conversations**, with a special focus on **emotions in task-oriented dialogue (ToD) systems**. ToD systems need to produce semantically and grammatically correct responses to fulfil the user's goal. Being able to perceive and express emotions pushes them one more step towards achieving human-likeness. To begin with, I constructed a dataset with meaningful emotion labels as well as a wide coverage of emotions and linguistic features in ToDs. Then, I improved emotion recognition in conversations (ERC) in the task-oriented domain by exploiting key characteristics of ToDs. Currently, I am working towards enhancing ToD systems with emotions.

1.1 Dataset Construction

Current research on emotions in conversations focuses on chit-chat dialogues because chit-chat dialogues are means for emotional expression and therefore are usually rich in emotions. Yet, emotions in ToDs, another important genre of spoken dialogues, are overlooked. In ToDs, users aim to achieve specific goals, such as hotel booking, by interacting with the system. While it is true that users do not express emotion the same way as they do in chit-chat dialogues, I observed that users do express various emotions concerning their goals. Users may talk about their feelings towards assorted situations that prompt them to interact with the system, such as a robbery or a vacation. It is also not uncommon to observe that users apologise to the system when they believe that they have caused trouble or confusion to the system, for example, when they try to correct or change their search criteria. In some worse scenarios, users may even insult the system. I am interested in such emotional nuances in users, which can have different implications for the system and would require different response strategies. This led me to construct **EmoWOZ**, **a corpus of task-oriented dialogues** where user emotions are annotated with our **tailored annotation scheme** (Feng et al., 2022).

1.1.1 Annotation Scheme for User Emotions

Existing ERC datasets make use of basic emotions from psychological theories. However, these emotion labels do not capture enough emotional nuances that are meaningful enough for ToDs. For example, to a ToD agent, it is unclear what "happiness" or "positive" means. What is missing that may influence system response here is whether the user emotion is elicited by the system.

In this spirit, I designed a tailored annotation scheme inspired by the Ortony, Collins, and Clore (OCC) model where emotions are defined as valenced reactions to various cognitive elicitors (Ortony et al., 1988). I devised a set of seven emotion labels considering three emotional aspects: **valence, elicitor, and conduct**. Valence concerns the positivity or negativity of emotions. Elicitor can be the system, including the entity proposed by the system, an event/fact, which is out of control of the system, or the user. The conduct aspect accounts for abusive behaviours.

1.1.2 Dialogue Collection and Annotation

I annotated user emotions in dialogues from two sources using Amazon Mechanical Turk. The first source is **MultiWOZ** (Budzianowski et al., 2018), one of the most well-established datasets for ToD modelling. Existing dialogue state labels in MultiWOZ allow us to investigate how task information can be leveraged to improve emotion recognition. The numerous benchmark results on MultiWOZ also allow us to directly assess the effectiveness of introducing emotion in ToD modelling tasks.

Since dialogues in MultiWOZ are human-to-human, and human operators rarely make mistakes, I additionally collected human-to-machine dialogues for balanced emotion coverage and diverse linguistic expressions. We refer to this sub-set as **DialMAGE** (**Dial**ogues with a **MA**chine **GE**nerated policy).

1.1.3 Annotation Quality Assurance

Given the difficulty and subjectivity in text emotion annotation, we adopted several quality assurance methods such as tutorials, qualification tests, hidden tests, and outlier detection. Each utterance was annotated by three English-speaking workers. The final inter-annotator agreement (Fleiss' Kappa) is 0.6, suggesting moderate to substantial agreement. This suggests a good usability of the dataset.

Proceedings of the 19th Annual Meeting of the Young Reseachers' Roundtable on Spoken Dialogue Systems, pages 54–56
September 11–12, 2023.

1.2 Improving ERC in ToDs

To build an emotion-aware ToD system, the first step is to give the system the ability to recognise user emotions. I first trained chit-chat ERC models with EmoWOZ and observed suboptimal results. This motivated me to exploit the characteristics of ToDs to improve ERC in ToDs. I proposed a framework called **ERToD** (**E**motion **R**ecogniser for **T**ask-**o**riented **D**ialogues), which effectively adapts chit-chat ERC models to the task-oriented domain by addressing three critical aspects: data, features, and objectives. First, I proposed two strategies of data augmentation to alleviate the class imbalance in EmoWOZ. Second, I used dialogue state as the task information encoding in combination with sentiment-aware text encoding. Third, I devised a multi-task learning objective and a novel emotion-distance weighted loss function. These approaches significantly improved the ERC performance of existing models.

1.3 Enhancing ToD Systems with Emotion

The ultimate goal of studying emotions in ToDs is to improve the system in either objective evaluation metrics or subjective user experience. Emotion is very important for a human operator, so can it influence all components in a modular ToD system. Correctly identifying the user emotion by the operator helps accurately identify the intent of the user and the status of the task completion, suggesting the potential of using emotion to improve downstream ToD modelling.

I showed that by considering emotion recognition as an auxiliary task in a multi-task learning framework, the joint goal accuracy of TripPy (Heck et al., 2020), a strong BERT-based dialogue state tracker, can be significantly improved. Our group has also developed an emotional user simulator (Lin et al., 2023), which exhibits diverse emotional expressions while achieving comparable task-related performance with other state-of-the-art generative user simulators. Currently, I am working towards incorporating emotion into other ToD modules, namely the dialogue policy and the natural language generator.

2 Spoken dialogue system (SDS) research

I expect in the future that SDSs can be more human-like by not only mimicking human responses but also mimicking the thinking process of humans. This should involve rationalising each decision of the system. Specific to my research interest, I envisage more use of emotion in task-oriented dialogue systems to push further the system performance as well as to improve the explainability of system behaviours via emotion. For example, the system should understand the user's situation and the cause of the user's emotion, which can hopefully lead to an optimal choice of dialogue acts as well as the system's emotional conduct.

3 Suggested topics for discussion

- **Ethics in Conversational AI:** When talking to computers, users are less refrained from showing impoliteness. What can we do to detect such behaviours? What is the proper response of a conversational AI? How can a conversational AI redirect the user towards good conduct?
- **Professionality:** What is the desired interpersonal skill and emotional behaviour of a ToD agent when it tries to show empathy?
- **Large Language Models (LLMs):** How can LLMs be applied to ToD when they are still prone to problems such as confabulation?

Biographical sketch

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

References

Paweł Budzianowski, Tsung-Hsien Wen, Bo-Hsiang Tseng, Iñigo Casanueva, Stefan Ultes, Osman Ramadan, and Milica Gašić. 2018. MultiWOZ - a large-scale multi-domain Wizard-of-Oz dataset for task-oriented dialogue modelling. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, pages 5016–5026. https://doi.org/10.18653/v1/D18-1547.

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 *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.

Michael Heck, Carel van Niekerk, Nurul Lubis, Christian Geishauser, Hsien-Chin Lin, Marco Moresi, and Milica Gasic. 2020. TripPy: A triple copy strategy

for value independent neural dialog state tracking. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*. Association for Computational Linguistics, 1st virtual meeting, pages 35–44. https://aclanthology.org/2020.sigdial-1.4.

Hsien-Chin Lin, Shutong Feng, Christian Geishauser, Nurul Lubis, Carel van Niekerk, Michael Heck, Benjamin Ruppik, Renato Vukovic, and Milica Gasic. 2023. Emous: Simulating user emotions in task-oriented dialogues.

Andrew Ortony, Gerald L. Clore, and Allan Collins. 1988. *The Cognitive Structure of Emotions*. Cambridge University Press. https://doi.org/10.1017/CBO9780511571299.