Toward Emotional Description via Meta-Dialogue: A Preliminary Framework for Data Collection

Kei Shimonishi[1], Koji Inoue[1], Yasuyuki Usuda[2], Rui Sakaida[3],
[1]Kyoto University, [2]Shizuoka Institute of Science and Technology, [3]Future University Hakodate
Correspondence: shimonishi@i.kyoto-u.ac.jp

Abstract

This study introduces a preliminary framework aimed at capturing richer and more nuanced emotional sentiments expressed in dyadic dialogues, moving beyond conventional emotion recognition research. We propose the concept of "meta-dialogue," a reflective conversational method where participants discuss their own prior dialogues, explicitly articulating their emotional states during those interactions. Preliminary results from recorded three sessions demonstrate the feasibility and potential of meta-dialogue for generating emotionally descriptive datasets, paving the way for enhanced dialogue systems and deeper insights into human emotional communication.

1 Introduction

With the advancement of large language models (LLMs), dialogue systems are now capable of generating highly empathetic responses (Fei et al., 2024; Sorin et al., 2024). However, to generate truly empathetic responses, it is essential to accurately recognize the user's emotions. Emotion recognition through speech and text has long been a central topic in media research (Poria et al., 2019b; Pereira et al., 2025), gaining attention not only for improving response generation but also for deepening our understanding of human interactions (Veltmeijer et al., 2021).

Research on emotion recognition relies on emotion models and datasets. Common models include categorical approaches, like Ekman's basic six emotions (Ekman, 1992), and dimensional frameworks using axes such as Valence and Arousal (Russell, 1980). However, dialogue datasets often come from acted or broadcast scenarios (Busso et al., 2008; Poria et al., 2019a), limiting their realism and diversity. This has caused a gap between model performance and understanding real human interactions.

To address this issue, we propose a novel concept named "psychosentience," defined as emotional descriptions that accurately reflect their dialogue context and situational nuances. For example, nuanced emotional expressions such as "I feel joy mixed with a touch of loneliness" can be naturally articulated. We posit that comprehension of such psychosentience can lead to deeper insights into dialogue and interaction research. Furthermore, advances in LLMs have now made this natural language-based approach technically feasible, offering promising applications across various domains such as counseling, elder care, and education, thus opening a new frontier in emotion recognition research.

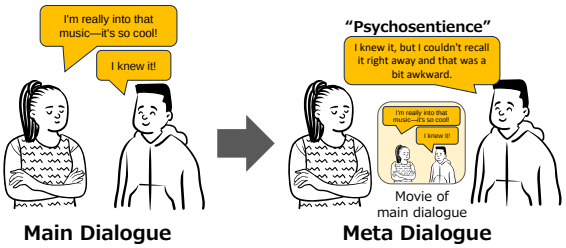

Figure 1: Concept of Meta-Dialogue

In this study, to gather data related to psychosentience (emotional description), we propose a new dialogue collection method, meta-dialogue. Meta-dialogue refers to a reflective dialogue in which participants talk about their own previous conversation, expressing the emotions they felt during it (see Figure 1). Such reflective dialogues resemble situations like a counselor reviewing a conversation with feedback, or post-match analyses in games like Chess. Our specific objective is to explore under what conditions meta-dialogue can be recorded, and to what extent and in what quality psychosentience expressions can be observed.

2 Meta-Dialogue Recording

As mentioned above, the goal of this study is to collect data on emotional sentiment in dialogues. We focus on dyadic (two-person) conversations and

Proceedings of the 29th Workshop on the Semantics and Pragmatics of Dialogue, September 3–5, 2025, Bielefeld, Germany.

Figure 2: Main-dialogue experiment setup

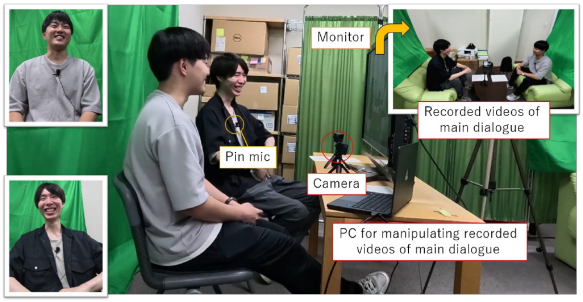

Figure 3: Meta-dialogue experiment setup

have participants engage in two rounds of dialogue. By limiting to dyadic dialogues, we ensure that both participants are always involved in the interaction, enabling us to collect emotional sentiment data specific to the dialogue itself.

The first round is referred to as the main dialogue, where participants engage in a typical conversation. The second round, the meta-dialogue, takes place immediately after the main dialogue, during which participants watch a recorded video of the main dialogue and discuss what they were thinking or feeling at each moment. By conducting the meta-dialogue immediately after the main one, we aim to capture participants' psychosentience while their memory is still fresh, allowing for sentiment data that closely reflects their actual emotional states during the main dialogue.

In dialogue-based emotion analysis, it is reported that emotion recognition becomes more accurate when visual information such as facial expressions are used in addition to transcripts and speech phonetics, which have traditionally been the main sources (Poria et al., 2017). Regarding facial expression analysis, methods have been proposed to finely evaluate expression intensity tailored to individual scales (Shimonishi et al., 2024), enabling more precise detection of subtle facial changes. Based on these findings, this study records dialogue in the following experimental environment.

Main Dialogue Figure 2 shows the experimental setup for the main dialogue. Participants sit face-to-face and engage in a natural conversation. Three cameras are used to record: one captures a side view of the conversation, and the other two are placed in front of each participant to record gestures and facial expressions.

In Figure 2, the side-view footage is overlaid with front-facing images of the participants. Additionally, each participant wears a pin mic to clearly capture individual speech, facilitating transcription and phonetic analysis. This setup enables simultaneous recording of each participant's (1) speech (textual and phonetic information), (2) gestures and facial expressions (visual information).

Meta Dialogue Figure 3 shows the experimental setup for the meta-dialogue. Unlike the main dialogue, participants sit side-by-side and view the main dialogue video on the same monitor. As with the main dialogue, three cameras are used to record the conversation and each participant's face, while pin mic capture speech.

During meta-dialogue, participants watch the video recorded from the side view of the main dialogue, allowing them to reflect from an external perspective. The video is played continuously without pausing. Although a design where the video is paused for each utterance could allow pinpoint reflections, we opted not to pause due to several reasons: (1) it would impose a heavy time burden on participants; (2) it might result in unnatural or forced sentiment expressions that diverge from genuine feelings. Thus, while it may not be possible to obtain one-to-one mappings of sentiment to each utterance, we expect to observe natural reflections on how participants felt or how their emotional states evolved during the dialogue sequence.

3 Conclusion

We presented a new scheme for recording dialogue data, named meta dialogue, aimed at collecting emotional descriptions. So far, we have empirically recorded three sessions involving six participants, and we are currently analyzing the data to see whether the expected emotional descriptions can be observed within these meta dialogues.

Acknowledgments

This work was supported by JSPS KAKENHI JP25H01382.

Limitations

This study is preliminary and has several limitations. First, the dataset is small (three sessions, six participants), all of whom are native Japanese speakers. This limits the generalizability of our findings to other populations and cultures. Second, the meta-dialogue method relies on participants' memory and self-report, which may be affected by recall bias or social desirability, especially since the same partner is present during reflection. Third, because the main dialogue video is not paused during meta-dialogue, it is difficult to precisely align emotional descriptions with specific utterances. Future work should expand the dataset and include participants from diverse backgrounds to validate and generalize the proposed framework.

Ethical Considerations

This research involves collecting sensitive emotional data from participants, requiring careful ethical consideration. First, the meta-dialogue method may cause psychological discomfort or affect relationships when emotions are verbalized. To mitigate this, we obtain informed consent, explain all procedures and risks, and allow participants to withdraw at any time. Second, the dialogue data collected in this study, especially the content of meta-dialogues which includes personal feelings, constitutes highly sensitive data. Therefore, we need to handle this data with the utmost care. Third, emotion descrition generation technology could be misused for manipulation or surveillance. We commit to transparency about its capabilities and limitations, and will contribute to guidelines for responsible use.

References

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language Resources and Evaluation (LREC)*, 42(4):335–359.

Paul Ekman. 1992. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200.

Hao Fei, Han Zhang, Bin Wang, Lizi Liao, Qian Liu, and Erik Cambria. 2024. EmpathyEar: An open-source avatar multimodal empathetic chatbot. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 61–71.

Patrícia Pereira, Helena Moniz, and Joao Paulo Carvalho. 2025. Deep emotion recognition in textual conversations: A survey. *Artificial Intelligence Review*, 58(1):1–37.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 873–883.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Rishi Naik, Erik Cambria, and Alexander Hoffmann. 2019a. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 527–536.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard Hovy. 2019b. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

James A. Russell. 1980. A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6):1161–1178.

Kei Shimonishi, Kazuaki Kondo, Hirotada Ueda, and Yuichi Nakamura. 2024. Ordinal scale evaluation of smiling intensity using comparison-based network. In *AAAI Workshop, Machine Learning for Cognitive and Mental Health Workshop (ML4CMH)*, pages 64–73.

Vera Sorin, Dana Brin, Yiftach Barash, Eli Konen, Alexander Charney, Girish Nadkarni, and Eyal Klang. 2024. Large language models and empathy: Systematic review. *Journal of Medical Internet Research*, 26:e52597.

Emmeke A Veltmeijer, Charlotte Gerritsen, and Koen V Hindriks. 2021. Automatic emotion recognition for groups: A review. *IEEE Transactions on Affective Computing*, 14(1):89–107.