Shun Katada

SANKEN, Osaka University, Japan

katada@sanken.osaka-u.ac.jp

1 Research interests

One of my research interests lies in **multimodal processing**. From a research perspective, multimodal is the science of heterogeneous and interconnected data (Liang et al., 2022). The multimodal processing methods mentioned here include linguistic, audio, visual, and biological information processing, which are important for understanding human behaviour. Computational representations and summarizing this information to reflect heterogeneity and interconnections are popular topics in this area. The author's current work focuses on creating multimodal spoken dialogue systems (SDSs) that can recognize human emotions/sentiments. The ultimate goal is to create an adaptive SDS that can change its behaviour by adapting a user's emotion based on multimodal processing.

Affective computing is another research interest that is not mutually exclusive. Affective computing relates to, arises from, or influences emotions (Picard, 2000). Although text modality has a dominant role in expressing emotion/sentiment during dialogue, nonverbal-based emotion recognition, such as facial expression and prosody, has been studied since the 1970s. Moreover, biosignals such as electroencephalograms (EEGs) and electrodermal activity (EDA) signals are often used in affective computing to detect emotional changes. Hence, the second topic focuses on these heterogeneous and interconnected data from an affective computing point of view, which is based on previous studies.

Biosignals have been used in previous works of the author; therefore, it is a candidate for the research topic but will be included in the above two topics.

1.1 Multimodal processing

First, the proposed method in previous work related to multimodal processing is shown (Katada et al., 2022). Language understanding has dramatically progressed through using large language models (LLMs), such as BERT and chatGPT, and has achieved excellent performance in emotion/sentiment estimation; however, using only linguistic information still has limitations. One of the issues is that sentiment is not necessarily expressed by users in human-agent interactions. To solve this issue, previous studies have proposed integrating token sequences derived from user utterances and time-series physiological (electrodermal) signals by multimodal processing. It was expected that integrating physiological signals into the language model can detect sentiment changes that are not expressed by user utterances. The Transformer architecture was applied to fuse text and physiological signals. As a result, our proposed methods significantly outperform the previous result, which is based on the simple early or late fusion method.

Second, a newly created multimodal dialogue corpus, called Hazumi2306, for developing an SDS with multimodal processing will be introduced, although it is not directly related to a new multimodal processing technique. The novelty of Hazumi2306 is that this dataset includes not only text, audiovisual, and physiological data but also frontal EEG data during human-agent interactions. The reason for collecting EEG data is that it has been the subject of focus in affective computing regions to capture unexpressed emotional changes in a controlled experimental environment. Approximately 500 minutes of chat dialogue were collected from thirty participants aged 20 to 70 years in total. The preliminary results of multimodal sentiment estimation based on conventional multimodal processing were also reported. It improved sentiment estimation performance when used with other modalities, although the simple EEG sensor used in this study has only three channels. This work has been published, and the corpus will be publicly available within the year (Katada et al., 2024). The analysis of this dataset by researchers will contribute to developing the SDS.

1.2 Affective computing

Multimodal analysis of human-agent interactions also sheds light on the emotional perception of humans. Basically, sentiment estimation based on multimodal processing considers only human observable signals such as linguistic, audio, and visual information. However, the contribution of the multimodal fusion of biosignals, which are unobservable by humans, has not been explored. In previous work (Katada et al., 2023), differences in the effect between observable (linguistic, audio, visual) and unobservable (physiological) signals were investigated in two different types of sentiment estimation, i.e., estimating sentiment labels annotated by the user and by a third party. Intuitively, a multimodal model based on the observable signal would be effective for estimating labels annotated by a third party since those labels are based on human observation (emotional perception). Addition-

The 20th Annual Meeting of the Young Researchers' Roundtable on Spoken Dialogue Systems, pages 30–31
September 16–17, 2024.

ally, a multimodal model based on the unobservable signal would be effective for estimating labels annotated by the users since those labels would include unexpressed sentiment. These assumptions are evaluated empirically, and the obtained results generally agree with these assumptions (Katada et al., 2023). The results suggest that physiological features are effective and that the fusion of linguistic representations with physiological features provides the best results for estimating self-sentiment labels. In contrast, the fusion of linguistic, audio, and visual features is effective for estimating sentiment labels based on third party, which can be derived from the corresponding signals that are observable by humans.

2 SDS research

Text-based dialogue systems have rapidly evolved in the past 10 years with the advent of deep learning, Transformer, BERT, and other LLMs. The number of model parameters and parallel computations continue to increase, and these efforts have enabled dialogue systems to produce accurate responses.

One simple perspective is that, unlike text-based dialogue systems, the SDS uses auditory data. Automatic speech recognition (ASR) may include some research topics related to LLMs. In a nonstationary noisy environment, the ASR performance degrades, and user utterance words that include word errors may be sent to an SDS equipped with an LLM. In this case, a dialogue breakdown may occur if the LLM cannot address the word error. Thus, handling word errors in ASR with LLM may be a research topic.

3 Suggested topics for discussion

Related to the abovementioned research, the need of working with signals that can be less invasive is one of the suggested topics. There are non-invasive techniques that may be useful for emotion recognition such as microgesture recognition, thermal imaging, sensors in mobiles, etc. It may be worth discussing what techniques with a multimodal spoken dialogue system would be valuable and practical.

References

Shun Katada, Shogo Okada, and Kazunori Komatani. 2022. Transformer-based physiological feature learning for multimodal analysis of self-reported sentiment. *International Conference on Multimodal Interaction (ICMI)* pages 349–358.

Shun Katada, Shogo Okada, and Kazunori Komatani. 2023. Effects of physiological signals in different types of multimodal sentiment estimation. *IEEE Transactions on Affective Computing* 14(3):2443–2457.

Shun Katada, Ryu Takeda, and Kazunori Komatani. 2024. Collecting human-agent dialogue dataset with frontal brain signal toward capturing unexpressed sentiment. *Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING)* pages 3518–3528.

Paul Pu Liang, Amir Zadeh, and Louis-Philippe Morency. 2022. Foundations and trends in multimodal machine learning: Principles, challenges, and open questions. *arXiv preprint arXiv:2209.03430* .

Rosalind W Picard. 2000. Affective computing. *MIT press* .

Biographical sketch

Shun Katada received a Ph.D. degree in life science from Tsukuba University, Japan, in 2014. He also received a Ph.D. degree in information science from JAIST, Japan, in 2022. He is currently a specially appointed assistant professor at SANKEN, Osaka University. He is a member of IPSJ and ACM.