Streaming Active Learning for Adaptive Multimodal Sentiment Analysis in Spoken Dialogue Systems

2410004 Atsuto Ajichi

Recent dialogue systems are increasingly expected to respond empathetically by recognizing and adapting to users' emotional states. In human–agent interaction, affective information is conveyed not only through linguistic content but also through nonverbal behaviors such as vocal prosody, facial expressions, and body movements. Multimodal sentiment analysis (MSA), which integrates these heterogeneous signals, has therefore been widely studied. However, emotional expression patterns differ substantially across individuals in terms of dominant modalities and expression intensity, making it difficult for population-level models to capture user-specific affective characteristics.

One possible way to address this issue is to directly ask users about their emotional states during interaction. While self-reported labels are generally reliable, frequent queries may interrupt conversational flow and increase user burden. This creates a trade-off between improving emotion recognition accuracy and maintaining a natural user experience, highlighting the need for mechanisms that can decide when to request emotion labels in an adaptive manner.

To tackle this problem, this thesis formulates personalized multimodal sentiment analysis as a stream-based active learning problem, in which data arrive sequentially and label acquisition decisions must be made online. Based on this formulation, the thesis proposes RAL-MSA, a reinforcement-learning-based framework that extends Reinforced Active Learning (RAL) to multimodal settings. RAL-MSA estimates prediction uncertainty independently for linguistic, acoustic, and visual modalities, integrates these signals using dynamically learned modality weights, and updates both the uncertainty threshold and modality weights through reinforcement learning based on the outcomes of past queries.

The proposed framework is evaluated through simulation experiments using two human–agent dialogue corpora, Hazumi1902 and Hazumi1911, which contain self-reported sentiment annotations collected during Wizard-of-Oz interactions. Experimental results indicate that, under few-shot conditions, RAL-MSA can achieve competitive sentiment recognition performance while reducing unnecessary label requests compared with baseline strategies. Analyses of learned modality weights, uncertainty thresholds, and permutation-based feature importance suggest notable inter-subject variability, implying that the relevance of modalities and features differs across users.

Overall, this thesis presents a framework that explores the potential of reinforcement-based stream-based active learning for user-adaptive multimodal sentiment analysis in dialogue systems. While the evaluation is simulation-based, the results suggest that strategically learning when and how to query users may support more personalized emotion recognition with reduced user burden, providing a foundation for future investigation in real-time and long-term interaction scenarios.