JAIST Repository
https://dspace.jaist.ac.jp/

Title	文脈、感情の動態、および話者パーソナリティのモデリングを取り入れた感情認
Author(s)	XUE, JIEYING
Citation	
Issue Date	2025-09
Type	Thesis or Dissertation
Text version	ETD
URL	http://hdl.handle.net/10119/20078
Rights	
Description	Supervisor: NGUYEN, Minh Le, 先端科学技術研究科, 博士

JAIST
JAPAN
ADVANCED INSTITUTE OF
SCIENCE AND TECHNOLOGY

Japan Advanced Institute of Science and Technology

氏名	XUE Jieying		
学位の種類	博士（情報科学）		
学位記番号	博情第 559 号		
学位授与年月日	令和 7 年 9 月 24 日		
論文題目	Emotion Detection with Context, Emotional Dynamics, and Speaker Personality Modeling		
論文審査委員	Nguyen Le Minh	JAIST	Professor
	Kiyoaki Shirai	JAIST	Professor
	Shinobu Hasegawa	JAIST	Professor
	Naoya Inoue	JAIST	Associate Professor
	Ken Satoh	NII	Director

論文の内容の要旨

Our research encompasses primarily two interrelated areas: Emotion Recognition in Conversations (ERC) and multilingual multi-label emotion detection. The former aims to identify the emotional state of each utterance in a dialogue, while the latter addresses the detection of multiple emotions across languages within a given sample.

Within the ERC domain, we explore several critical yet underexplored dimensions: In emotion context modeling, traditional sequential models often capture only local emotional dependencies, overlooking long-range emotional influences that may arise between distant parts of a conversation. We argue that emotional states can be affected and transmitted by speakers and utterances throughout the conversation, regardless of their positional distance. In response to this limitation, we propose the *L*ong-range dependenc*Y* emotion*S* *M*odel (LYSM), which employs self-attention mechanisms to capture emotional dynamics throughout entire conversations, allowing the system to integrate emotional dependencies from both nearby and distant utterances. Experimental results on four benchmark datasets confirm its strong generalizability and effectiveness.

In utterance representation, unlike isolated sentence-level emotion tasks, the core challenge in dialogue emotion recognition lies in effectively representing a target utterance within its conversational context. Although pre-trained language models (PLMs) such as RoBERTa offer strong capabilities for context modeling, existing PLM-based approaches often fail to fully exploit their potential for fine-grained contextual encoding. To overcome these limitations, we introduce *Acc*umulating Word *R*epresentations in Multi-level Context Integration for ERC Task (AccWR), which aggregates multi-level contextual word information before inputting it into the PLM. This enriched contextual word aggregation enhances both semantic understanding and the model’s focus on the target utterance. Experimental results show that AccWR consistently outperforms baseline models on four benchmark datasets and demonstrates strong potential for broader applications such as response generation and semantic parsing.

For speaker modeling, ERC typically relies on spoken dialogues transcribed by automatic speech recognition systems. Individual traits such as linguistic style and personality significantly influence emotional expression, yet prior work often depends on implicitly learned speaker features, limiting interpretability and cross-domain generalization. To address this, we propose BiosERC: Integrating *Bio*graphy-Based *S*peaker Representations with Large Language Models for *E*motion *R*ecognition in *C*onversations, which employs LLMs with prompt-based techniques to extract explicit speaker profiles as external knowledge. These biography-based representations enhance the emotional understanding of each speaker, leading to more accurate and nuanced recognition, especially in complex or multi-party conversations. BiosERC achieves competitive or state-of-the-art (SOTA) performance on three benchmark datasets.

In multi-level context modeling, the spontaneous nature of conversations makes it difficult to capture transient and dynamically evolving emotional states using only contextual discourse and static speaker profiles. To address this, we propose TraceERC: *T*racking *R*elational *A*wareness of *C*ontextual, Character, and *E*motional States in *E*motion *R*ecognition in *C*onversations, which leverages LLMs to jointly encode dialogue context, speaker personality, and dynamic emotional cues. These enriched representations enable emotion predictions that are sensitive to both conversational flow and individual speaker characteristics, enhancing emotional understanding and adaptability. As one of the first LLM-based ERC models to incorporate contextual learning and in-context fine-tuning, TraceERC achieves SOTA results on MELD and strong performance across benchmarks.

In the multilingual domain, our team participated in SemEval-2025 Task 11, tackling both multi-label classification (Track A) and emotion intensity detection (Track B). We developed a generation-based framework leveraging multilingual PLMs and LLMs to support both high- and low-resource languages.

Keywords: Emotion Recognition in Conversations, Multilingual Multi-label Emotion Detection, Large Language Models, Speaker Personality Modeling, Emotion Context Modeling

論文審査の結果の要旨

This thesis presents a clear, well-organized set of studies that advance Emotion Recognition in Conversations (ERC) **and** multilingual, multi-label emotion detection**.**

The main contributions are:

- **LYSM** for modeling long-range links across a whole conversation,
- **AccWR** for a better representation of each utterance in its context, and
- **BiosERC** for using speaker profiles **(**from short bios) with large language models,
  leading to **TraceERC**, which tracks context, speaker traits, and changing emotion**s** together.

The experiments cover several public datasets, show consistent improvements**,** and reach top results on the standard data set **(MELD)**. The methods are well motivated and verified by careful experiments. This work also connects classic pretrained language model pipelines with LLMs**,** works across multiple languages**,** and makes ERC more interpretable through explicit speaker information. I do believe that the research could open next steps such as adding audio/vision cues, improving robustness to ASR errors, handling privacy in speaker profiles, speeding up models, and testing more domains and low-resource settings with fairness checks**.**

The candidate has published this research in international journals and conferences and received the **Best Description Paper Award at SemEval 2025**, showing both scholarly impact and community recognition.

Overall, this is a strong contribution that supports for conversational emotion understanding. It provides effective models and clear design ideas for others to build on, and it deserves strong recognition**.** This is an excellent dissertation, and we approve of awarding a doctoral degree to Ms. XUE Jieying.