FACULTY OF APPLIED SCIENCES
UNIVERSITY
OF WEST BOHEMIA

STUDENT
SCIENTIFIC CONFERENCE 2024

Dialogue Representation and Multimodal Fusion for Emotion Cause Analysis

Josef Baloun[1], Jiří Martínek[2], Ladislav Lenc[3], Pavel Král[4], Matěj Zeman[5], Lukáš Vlček [6]

1 Introduction

In this paper, we present an approach for solving *SemEval-2024 Task 3: The Competition of Multimodal Emotion Cause Analysis in Conversations* (Wang et al. , 2024). The task includes two subtasks that focus on emotion-cause pair extraction using text, video, and audio modalities. Our approach is composed of encoding all modalities (MFCC and Wav2Vec for audio, 3D-CNN for video, and transformer-based models for text) and combining them in an utterance-level fusion module. The model is then optimized for link and emotion prediction simultaneously. Our approach achieved 6th place in both subtasks. The full leaderboard can be found at `https://codalab.lisn.upsaclay.fr/competitions/16141#results`.

2 Task

The goal is to extract potential pairs of emotions and corresponding causes in a conversation/document and/or other source of dialogue as illustrated in Fig. 1.

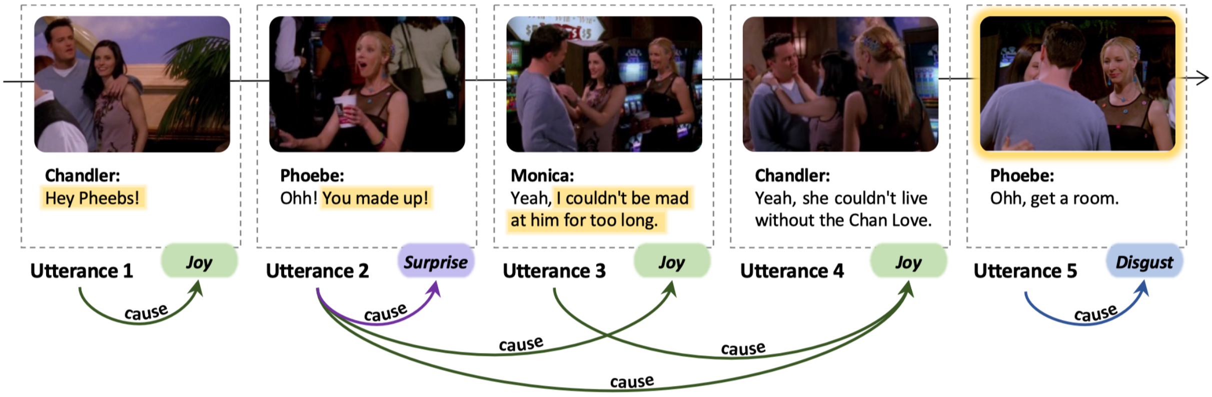

Figure 1: Emotion-cause pair extraction task illustration

[1] student of the doctoral degree program Applied Sciences, field of study Informatics and Cybernetics, e-mail: balounj@students.zcu.cz

[2] e-mail: jimar@kiv.zcu.cz

[3] e-mail: llenc@kiv.zcu.cz

[4] e-mail: pkral@kiv.zcu.cz

[5] student of the master degree program Applied Sciences, field of study Informatics and Cybernetics, e-mail: zemanm98@students.zcu.cz

[6] student of the master degree program Applied Sciences, field of study Informatics and Cybernetics, e-mail: vlcek0@students.zcu.cz

3 System Overview

We decomposed the main objective into emotion and link prediction (the estimation of pairs) tasks. The final result then consists of source and target utterances provided by the link and emotion of the target utterance.

The architecture is depicted in Figure 2. First, we encode the different modalities at the utterance or dialogue level to incorporate more context. Next, we fuse the representations at the utterance level. Once we have representations of all individual utterances in a dialogue, we predict links and emotions.

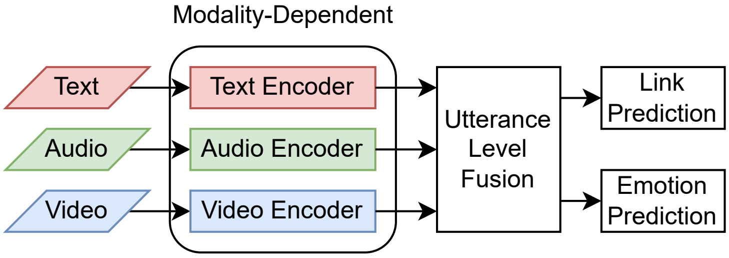

Figure 2: System architecture

4 Conclusions

Our model incorporates all modalities (text, audio and video features). The multi-task learning may help for a single task, but it is very likely suboptimal. We have obtained better overall results with two separate models and encountered conflicts in fusion strategies (whether to use the aggregation or the fusion token). Our best model for emotion prediction is text-only with no fusion mechanism, while the best model for linking benefits from the aggregation fusion strategy. Our key findings during the result analysis are as follows:

1. The primary source of information resides within the text. Processing of audio/video modalities has brought only a small positive impact in the case of the link prediction task.
2. The information about emotion showed to be important for the link prediction task and significantly improves the results.
3. The context of the whole dialogue (processing multiple utterances at once) is crucial for link prediction.

Acknowledgement

This work has been partly supported by the OP JAC project DigiTech no. CZ.02.01.01/00/23_021/0008402 and by the Grant No. SGS-2022-016 Advanced methods of data processing and analysis. Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic.

References

Wang, F., Ma, H., Xia, R., Yu, J., and Cambria, E. (2024). Semeval-2024 task 3: Multimodal emotion cause analysis in conversations. *In Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pp. 2022-2033.