Multimodal Emotion-Cause pair extraction

DS510: Aritificial Intelligence and Machine Learning Lab

AUTHORS

Ganta Naga Jaswanth
Chelluboina Siri

Contents

1 Introduction

The idea picked by our team is **'Multi modal Emotion Cause Pair Extraction'** in educational systems. The main goal of this project is to extract the emotion-cause pairs from multimodal data in educational systems. The multimodal data includes text, audio, and video data of the utterances of the participants in the educational system. This is important because understanding the emotions of the students can help in improving the educational systems. In an online educational system, the students interact with the system through video lectures from which the system can extract the emotions of the students. This can help the system to understand the emotions of the students and provide the necessary support to the students.

1.1 Problem Statement (Objective)

Given the text, audio, and video data for different speakers and utterances in an educational system, the task is to extract the emotion-cause pairs. The emotion-cause pairs are the pairs of emotions and the causes of those emotions.

2 Methodology

2.1 Research followed

The research paper followed for this project is **'Multimodal Emotion-Cause Pair Extraction in Conversations'** by **Wang and Ding**. The paper proposes both data and model for the task of extracting emotion-cause pairs from multimodal data.

2.2 Data

The data used in the paper is based on the **'MELD'** dataset. The MELD dataset is a multimodal dataset that contains text, audio, and video data. They annotated the original MELD dataset with emotion-cause pairs to create the ECF dataset. An emotion cause pair is a pair of emotion and the cause of that emotion in terms of utternace.

This dataset contains annotations for 1,374 conversations composed of 13,619 utterances for the 3 modalities. This dataset considers 7 types of emotions: **'anger', 'disgust', 'fear', 'joy', 'neutral', 'sadness', 'surprise'**.

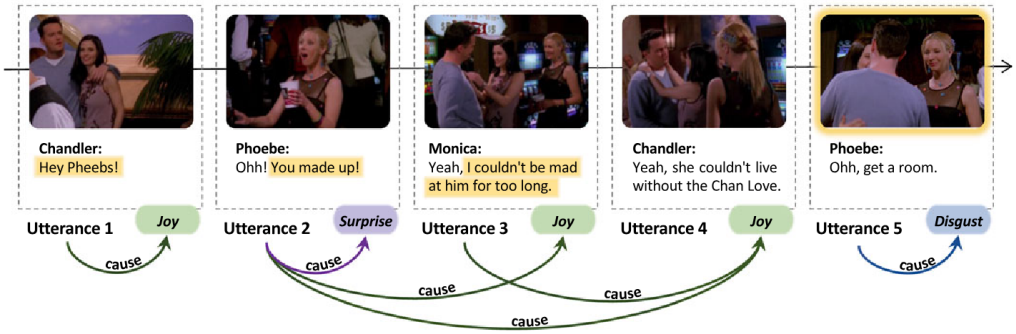

Figure 1: Example of emotion-cause pairs

2.3 Model

2.3.1 Unimodal Feature Extraction

- **Text Feature Extraction:** The text data is first initialized with the GloVe embeddings, which are then fed into a BiLSTM layer to extract the features.
- **Audio Feature Extraction:** Spatial audio features are extracted using openSMILE toolkit. The dimension of the audio features are 6373.
- **Video Feature Extraction:** The video data is first processed using C3D network (3D Convolutional Neural Network) to extract the features of 4096 dimensions.

2.3.2 Multimodal Feature Fusion

The features extracted from the unimodal data are then fused using simple concatenation.

2.3.3 Model Architecture

The model consists of 2 main components:

- **Extraction of Emotion utterances and Cause utterances:** The model first extracts the emotion and cause utterances from the concatenated multimodal data. This step is done using 2 independent BiLSTM layers one each for emotion and cause.
- **Emotion-Cause Pair Extraction:** On the obtained sets of emotion and cause utterances, first cartesian product is taken to get all possible pairs. Then, a BiLSTM layer is used to get the features of the pairs. Finally, a softmax layer is used to predict the probability of the emotion-cause pair.

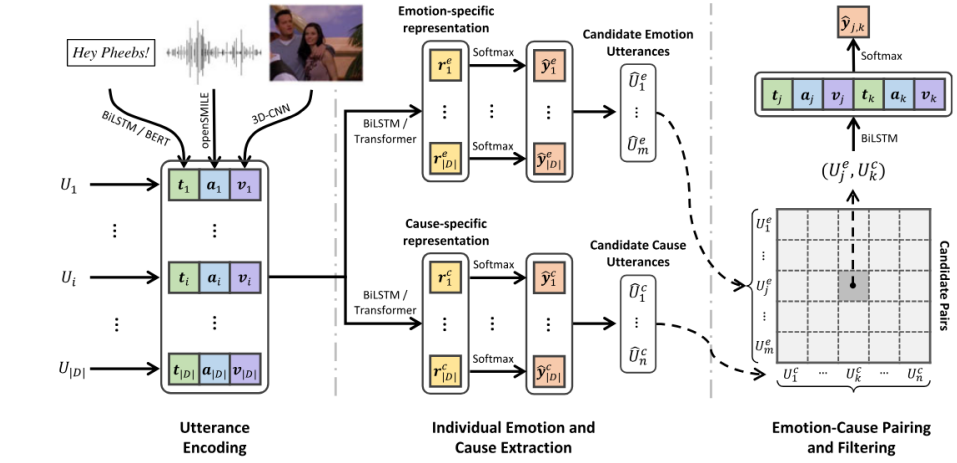

Fig. 5. The main structure of the deep learning baseline system MECPE-2steps.

Figure 2: Model Architecture

3 Experimentation & Results

In the experimental setting, the proposed method is compared with the MECPE method. The results are shown in Table 1. The entire dataset is split into train, test and validation sets in the ratio of $70:10:20$. During training, the number of utterances per conversation is set to 35 and the number of tokens in an utterance is also set to 35. The number of hidden units in the LSTM layer is set to 200. The number of epochs is set to 20. The learning rate is set to 0.001 and batch size is set to 32.

	Accuracy	Precision	F1
MECPE	0.55	0.47	0.50
Ours	0.53	0.44	0.46

Table 1: Comparison of performance metrics

The results show that the proposed method performs slightly worse than the MECPE method. This could be due to the discrepancy in feature extraction methods.

4 Deployment

As a part of our work, we have developed a visualizer for emotion-cause pairs. The visualizer is a web application that takes the input of a conversation and displays the emotion-cause pairs extracted from the conversation.

The 3 example uses a conversation from test split of the ECF dataset.

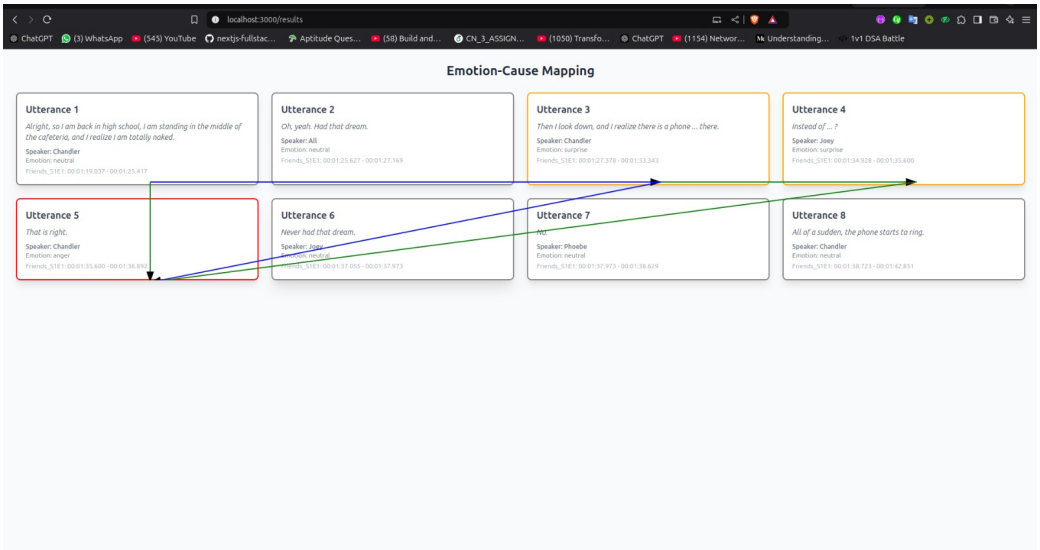

Figure 3: Visualizer for emotion-cause pairs

But in real-life conversations, the conversation is not always in the form of a list of sentences. So, we need to extract the utternaces in text using the audio of the conversation which could be a challenging task. We need to separate the signals in the conversation and align them sequentially to get the text modality.

5 Conclusion & Future Work

The novel task of emotion-cause pair extraction from conversations can be used in various applications like customer service, mental health, and social media analysis.

From an architectural perspective, the model can be improved by using a transformer-based model like BERT or RoBERTa for better performance. Also, the audio and video feature extraction can be further imporved using fine-grained emotion recognition models to get better emotion representations or attention based models like Vision Transformer (ViT) for better video feature extraction.

In an educational setting, this is particularly useful for understanding the emotional state of students and the causes behind them. This will enhance the learning experience and help in providing better support to students. But the current model is not directly deployable in an offline setting since it requires infrastructure for audio and video processing in the classroom.

References

[1] Wang, Fanfan and Ding, Zixiang and Xia, Rui and Li, Zhaoyu and Yu, Jianfei, "Multimodal Emotion-Cause Pair Extraction in Conversations", *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1832-1844, 2023. doi: 10.1109/TAFFC.2022.3226559.