QFNU_CS at SemEval-2024 Task 3: A Hybrid Pre-trained Model based Approach for Multimodal Emotion-Cause Pair Extraction Task

Zining Wang, Yanchao Zhao, Guanghui Han, Yang Song
School of Computer Science, Qufu Normal University, Ri Zhao, China

Abstract

This article presents the solution of Qufu Normal University for the Multimodal Sentiment Cause Analysis competition in SemEval2024 Task 3.The competition aims to extract emotion-cause pairs from dialogues containing text, audio, and video modalities. To cope with this task, we employ a hybrid pre-train model based approach. Specifically, we first extract and fusion features from dialogues based on BERT, BiLSTM, openSMILE and C3D. Then, we adopt BiLSTM and Transformer to extract the candidate emotion-cause pairs. Finally, we design a filter to identify the correct emotion-cause pairs. The evaluation results show that, we achieve a weighted average F1 score of 0.1786 and an F1 score of 0.1882 on CodaLab.

1 Introduction

The competition of multimodal emotion cause analysis(Gandhi et al., 2023) involves not only understanding linguistic content but also recognizing and comprehending various forms of information such as emotional expressions, sounds, and images. The significance of this competition lies in its ability to comprehensively understand and interpret emotions and motivations in human communication. By analyzing various forms of information in conversations, we can more accurately identify the sources and reasons for emotions, thereby enhancing our understanding of human behavior and communication methods. This holds importance across various fields including psychology, human-computer interaction, and affective computing, aiding in the development of more intelligent and human-centric technologies and systems, improving communication efficiency and quality, and promoting better understanding and communication among individuals.

This paper details our contribution to SemEval-2024 Task 3: Multimodal Emotion Cause Analysis in Conversations(Wang et al., 2024), encompassing two sub-tasks: extracting emotion-cause pairs(Xia and Ding, 2019) from text-only dialogues and from multimodal dialogues that include text, audio, and video modalities. In this task, we place a particular emphasis on implementing Sub-task 2.

For Sub-task 2: Multimodal Emotion-Cause Pair Extraction, we aim to extract emotion-cause pairs from dialogues that contain representations in text, audio, and video. Each pair includes an emotional utterance, its emotion category, and a cause utterance. The challenge lies in integrating information from multiple modalities to accurately identify emotional expressions and their related causes.

In our approach to task 2, we first preprocessed the dataset, mapping the text, audio, and video data of the Emotion Cause in Friends (ECF) dataset to a unified feature space. Then, we utilized a baseline model with a two-stage training scheme: emotion recognition and cause pair extraction. This approach focused on utilizing modalities, selecting models such as Bert(Devlin et al., 2018) and LSTM(Yu et al., 2019), and adjusting parameters for two phases of model training. After that, we predicted on test data in two stages using the trained models and evaluated the results through CodaLab to obtain corresponding F1 scores.

Our best-performing solution involved using Bert for emotion recognition followed by LSTM for cause pair extraction across all three modalities, achieving an F1 score of 0.1882. This methodological progression demonstrates our systematic approach to tackling the complexities of multimodal emotion cause analysis, highlighting our efforts in dataset preprocessing, model experimentation, and performance evaluation, while also proving the effectiveness of the baseline model(Wang et al.).

2 Methodology

In this section, we describe the E-MECPE method

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 349–353
June 20-21, 2024 ©2024 Association for Computational Linguistics

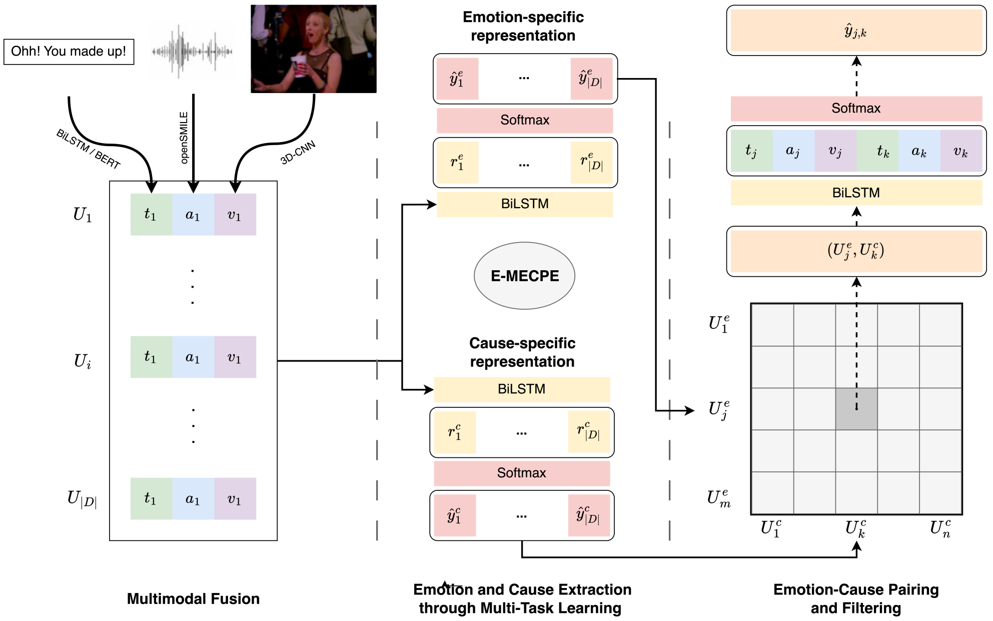

Figure 1: Framework diagram of the E-MECPE methodology

in depth. In general, this method is divided into three main parts: multimodal fusion, emotion and cause extraction through multi-task learning and emotion-cause pairing and filtering. The methodology of this paper is summarized in Fig. 1.

2.1 Multimodal Fusion

First, we obtain the representations of the three modalities from the text, audio and video modalities for their respective modalities. Then, the three modalities are stitched together in the order of text-audio-video to obtain the joint representation of the three modalities. The feature extraction method for each of these modalities is as follows:

For text, each token is initialized with pre-trained 300-dimensional GloVe vectors(Pennington et al., 2014). Subsequently, we used two different models to extract text features: the BiLSTM (Bidirectional Long Short-Term Memory Network) and the BERT (Bidirectional Encoder Representation Transformer).BiLSTM is a classical recurrent neural network that can effectively capture long-term dependencies in text sequences by a bidirectional structure that considers both forward and backward information. BERT, on the other hand, is a pre-trained language model based on the Transformer architecture, which is pre-trained on large-scale textual data and is able to capture rich semantic information. In this study, BiLSTM is used as a textual feature extractor to capitalize on its representational learning ability in sequential data; while BERT, as another textual feature extractor, acquires deeper semantic information by pre-training the model(Kim and Park, 2023). These two models are independently applied to discourse-level feature extraction tasks to evaluate their performance on sentiment and cause extraction tasks.

In the audio domain, we extract the 6373-dimensional acoustic features (a_i) using the openSMILE toolkit, leveraging the feature set designed for the INTERSPEECH 2013 Emotion Challenge. This comprehensive approach allows us to capture nuanced acoustic characteristics, providing a rich foundation for our subsequent analyses.

For video processing, we use a 3D-CNN network variant called C3D(Tran et al., 2015; Rao and Liu, 2020) to extract 16 frames from each video and process them through the C3D network to obtain 4096-dimensional video descriptors optimized for dimensionality reduction and to extract 128-dimensional visual features from each speech video.

2.2 Emotion Extraction

Our aim in sentiment extraction is to derive sentiment-related features from the discourse. We process each discourse to obtain sentiment-specific representations (re_i) by means of a specific bidirectional long short-term memory network (BiLSTM). The BiLSTM processes the forward and reverse sequences of the discourse separately by means of its two LSTM units, and ultimately combines the outputs of these two directions. This allows the network to take into account both the forward and backward information of the discourse, thus providing a more comprehensive understanding of the emotional content of the discourse.

Next, the sentiment-specific representation (re_i) is fed into a softmax layer, the output of which can be considered as the probability that the discourse belongs to each sentiment category. the formula for the softmax layer is as follows:

$$\hat{\mathbf{y}}_i^e = \mathrm{softmax}\big(\mathbf{W}^e \mathbf{r}_i^e + \mathbf{b}^e\big) \tag{1}$$

where W_e and b_e are the weights and biases of the softmax layer, respectively, and $\hat{y}_i^e$ is the predicted sentiment distribution.

2.3 Cause Extraction

The purpose of the cause extraction part is to recognize causal relationships in discourse. We use another BiLSTM to extract cause-specific representations (rc_i). This BiLSTM works in a similar way to the BiLSTM used in sentiment extraction, but the parameters are not shared to ensure that the network learns the specific features for the cause extraction task.

The reason-specific representation (rc_i) is then fed into another softmax layer that focuses on determining the probability of different reason categories in the discourse. The formula for this softmax layer is as follows:

$$\hat{\mathbf{y}}_i^c = \mathrm{softmax}\big(\mathbf{W}^c \mathbf{r}_i^c + \mathbf{b}^c\big) \tag{2}$$

Here, W_c and b_c are the weights and biases of this softmax layer, and $\hat{y}_i^c$ denotes the predicted cause distribution.

2.4 Loss calculation

Our goal is to minimize the cumulative loss of the model on the emotion extraction and cause extraction tasks. The total loss L_{total} is the sum of the losses of the two tasks and is calculated by the following formula:

$$\mathcal{L}_{\text{total}} = -\sum_{i=1}^{N}\left(\sum_{j=1}^{C} y_i^{e,j}\log(\hat{y}_i^{e,j}) + \sum_{k=1}^{K} y_i^{c,k}\log(\hat{y}_i^{c,k})\right) \tag{3}$$

Where $y_{j,i}^e$ and $y_{k,i}^c$ denote the uniquely hot encoding of the true emotion and cause labels, respectively, N is the number of training samples, and C and K are the number of emotion and cause categories, respectively. This loss function ensures that the model learns to extract features related to emotions and reasons efficiently, thus improving the model's performance on both tasks.

2.5 Emotion-Cause Pairing and Filtering

Following the acquisition of Candidate Emotion Utterances and Candidate Cause Utterances, the pivotal task is to discern the existence of a causal relationship between sets E and C, ensuring the extraction of valid emotion-cause pairs. Initially, E and C are organized into a dot matrix, depicted in the third segment of Fig. 1, resulting in the generation of all conceivable candidate pairs denoted as $x(U_j^e; U_k^c)$. This vector amalgamates the self-contained multimodal representations of the emotion and cause expressions, along with a distance vector capturing the relational nuances between the two expressions.

The composite representation is then inputted into a softmax layer to determine the validity of the pairing x, filtering and extracting relevant emotion-cause pairs from numerous possibilities.

$$\hat{\mathbf{y}}_{j,k} = \mathrm{softmax}\left(\mathsf{W}\mathbf{x}_{(\hat{U}_j^e, \hat{U}_k^e)} + \mathbf{b}\right) \tag{4}$$

3 Experiments

3.1 Data Resources

The official dataset consists of three modalities: text, audio, and video clips, and includes 1,374 conversations and 13,619 utterances annotated for 9,794 emotion-cause pairs across the three modalities. The relevant connections are stored in a JSON file and correspond to independent video segments through specified IDs. In order to fully utilize all the multimodal data, we first preprocess and reduce the dimensionality of the data according to the methods described in the paper.

Specifically, during the preprocessing stage, for the audio data in the video, we use the ffmpeg tool to extract the corresponding audio files for each video segment. We then utilize the open-source tool called openSMILE (Eyben et al., 2010) and apply The INTERSPEECH 2013 ComParE feature set (Schuller et al., 2013), which is the default feature set of openSMILE, to extract features from the audio data. As a result, we obtain a 6373-dimensional acoustic feature vector.For video data, we refer to the C3D model structure to extract video features and obtain a 4096-dimensional representation.As for text data, following the same approach as described in the paper, we utilize pre-trained Glove word vectors to obtain text embeddings.

3.2 Training

The training process is divided into two parts: the first part is emotion extraction and cause extraction, and the second part is the extraction of emotion-cause pairs. We explored three different training conditions: utilizing only textual modalities, combining textual and audio modalities, combining textual and video modalities, and leveraging all data modalities and fine-tune the model parameters based on the baseline to select the appropriate parameters to obtain the best score.

Emotion extraction and cause extraction: The initial phase of our experiment compared emotion extraction using Bert and BiLSTM model architectures, conducted on an RTX 4070Ti Super GPU setup. Key training parameters were carefully selected to enhance model performance. The batch size for BiLSTM was fixed at 16, while for Bert, it was set to 4, with the training spanning 15 epochs. The loss weights for both emotion extraction and cause extraction tasks were set to 1.0, indicating their equal importance in our training objectives.

Emotion-cause pairs extraction: In the subsequent phase focusing on cause pair identification, the same model architecture was employed, trained under identical conditions to assess the effect of data modality on performance. The batch size was increased to 200 to potentially improve generalization, with a learning rate of 0.005 aimed at optimal convergence. A 0.5 dropout keep probability for word embeddings was introduced for added regularization, while maintaining a 1.0 keep probability for the softmax layer. The l2 regularization coefficient remained at 1e-5, consistent with our approach to model complexity control.

3.3 Evaluation

Similar to baseline, we utilize the macro-averaged F1 score (Gui et al., 2018) as the primary evaluation metric for our task. This metric accounts for both precision and recall, providing a balanced assessment of model performance. The F1 score is calculated using the following formula:

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{5}$$

where:

- P denotes precision, calculated as the ratio of correctly predicted emotion-cause pairs to the total predicted pairs.
- R denotes recall, calculated as the ratio of correctly predicted emotion-cause pairs to the total annotated pairs.

In our evaluation, F_1 is the harmonic mean of precision and recall, indicating the model's balance in detecting emotion-cause pairs: a higher F_1 score signifies better performance.

Table 1: Experimental Results

Model	Modality	$F1_{emotion}$	$F1_{caution}$	$F1_{pair}$
BiLSTM	T	0.7441	0.7008	0.5041
	TA	0.7398	0.6986	0.5104
	TV	0.7431	0.7016	0.5162
	TAV	0.7422	0.6993	0.5226
Bert	T	0.7362	0.6687	0.5104
	TA	0.7356	0.6637	0.5160
	TV	0.7365	0.6700	0.5104
	TAV	0.7363	0.6648	0.5246

3.4 Results and analysis

We assessed Bert and BiLSTM models on various modalities: text (T), text-audio (TA), text-video (TV), and their combination (TAV), as shown in Table 1. Results underline the models' proficiency in extracting sentiment-cause pairs from multimodal dialogues, with distinct performance variations across modalities.

The BiLSTM model demonstrates incremental improvements in $F1_{pair}$ scores from T to TAV, indicating the advantage of utilizing multimodal data. The highest performance is observed in the TAV setup with a score of 0.5226, underscoring the benefits of combining text, audio, and video.

Conversely, the Bert model showcases superior performance in the TAV modality, achieving an $F1_{pair}$ score of 0.5246. This performance highlights Bert's ability to effectively leverage deep contextual embeddings across modalities for more accurate extraction of sentiment-cause pairs. The robustness of Bert, particularly in the multimodal TAV setup, confirms its efficacy in handling complex multimodal data.

Overall, Bert emerges as the preferred model for extracting sentiment-cause pairs across all modalities, with a peak performance in the TAV configuration, reflected by a weighted average F1 score of 0.1786 and an F1 score of 0.1882 on CodaLab. These findings advocate for the continued exploration of multimodal approaches, particularly leveraging models like Bert that excel in contextual understanding and integration of multimodal data.

4 Conclusion

In this paper, we present Effective Multimodal Emotion-Cause Pair Extraction (E-MECPE) method. We used this method to perform emotional cause analysis on the Emotion-Cause-in-Friends (ECF) dataset. Ablation experiments were conducted for text unimodal and multimodal under different text encoders, respectively, and the relevant parameters associated with the experiments were tuned. The experimental results show that BERT encoding-based text representation and multimodal joint representation help in the extraction of emotional cause pairs, and that the parameter settings are crucial for the performance enhancement of this task. This finding not only validates the effectiveness of our method, but also points out an important direction for future research in the field of sentiment analysis kresearch by pointing out an important direction.

5 Prospects for Advancement

Due to the late entry time, limited hardware resources, and short submission period, we only had time to fine-tune and conduct ablation experiments based on the baseline. However, we believe that there is still a lot of room for improvement in adjusting this model. For example, further attempts can be made in aligning and filtering methods for multimodal data, selecting more encoders, and enhancing the model's understanding of causal relationships. We will also continue exploring on top of this model to continuously advance the development of this research direction.

6 Acknowledgements

This research is supported by Natural Science Foundation of Shandong Province [grant number ZR2022QF088]; Natural Science Foundation of Rizhao [grant number RZ2021ZR30]; and ShanDong Students' Platform for innovation and entrepreneurship training program [grant number S2023104460124]

References

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Florian Eyben, Martin Wöllmer, and Björn Schuller. 2010. Opensmile: the munich versatile and fast open-source audio feature extractor. In *Proceedings of the 18th ACM International Conference on Multimedia*, MM '10, page 1459–1462, New York, NY, USA. Association for Computing Machinery.

Ankita Gandhi, Kinjal Adhvaryu, Soujanya Poria, Erik Cambria, and Amir Hussain. 2023. Multimodal sentiment analysis: A systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions. *Information Fusion*, 91:424–444.

Lin Gui, Ruifeng Xu, Dongyin Wu, Qin Lu, and Yu Zhou. 2018. Event-driven emotion cause extraction with corpus construction. In *Social Media Content Analysis: Natural Language Processing and Beyond*, pages 145–160. World Scientific.

Kyeonghun Kim and Sanghyun Park. 2023. Aobert: All-modalities-in-one bert for multimodal sentiment analysis. *Information Fusion*, 92:37–45.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Chengping Rao and Yang Liu. 2020. Three-dimensional convolutional neural network (3d-cnn) for heterogeneous material homogenization. *Computational Materials Science*, 184:109850.

Björn Schuller, Stefan Steidl, Anton Batliner, Alessandro Vinciarelli, Klaus Scherer, Fabien Ringeval, Mohamed Chetouani, Felix Weninger, Florian Eyben, Erik Marchi, et al. 2013. The interspeech 2013 computational paralinguistics challenge: Social signals, conflict, emotion, autism. In *Proceedings INTERSPEECH 2013, 14th Annual Conference of the International Speech Communication Association, Lyon, France*.

Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. 2015. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497.

Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. Multimodal emotion-cause pair extraction in conversations. *IEEE Transactions on Affective Computing*, 14(3):1832–1844.

Fanfan Wang, Heqing Ma, Rui Xia, Jianfei Yu, and Erik Cambria. 2024. Semeval-2024 task 3: Multimodal emotion cause analysis in conversations. In *Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024)*, pages 2022–2033, Mexico City, Mexico. Association for Computational Linguistics.

Rui Xia and Zixiang Ding. 2019. Emotion-cause pair extraction: A new task to emotion analysis in texts. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1003–1012.

Yong Yu, Xiaosheng Si, Changhua Hu, and Jianxun Zhang. 2019. A review of recurrent neural networks: Lstm cells and network architectures. *Neural computation*, 31(7):1235–1270.