Samsung Research China-Beijing at SemEval-2024 Task 3: A multi-stage framework for Emotion-Cause Pair Extraction in Conversations

Shen Zhang[⋆], **Haojie Zhang**[⋆✉], **Jing Zhang**
Xudong Zhang, Yimeng Zhuang, Jinting Wu
Samsung R&D Institute China-Beijing
{shen02.zhang, tayee.chang, jing97.zhang,
xudong.z1, ym.zhuang, jinting01.wu}@samsung.com

Abstract

In human-computer interaction, it is crucial for agents to respond to human by understanding their emotions. Unraveling the causes of emotions is more challenging. A new task named Multimodal Emotion-Cause Pair Extraction in Conversations is responsible for recognizing emotion and identifying causal expressions. In this study, we propose a multi-stage framework to generate emotion and extract the emotion causal pairs given the target emotion. In the first stage, Llama-2-based InstructERC is utilized to extract the emotion category of each utterance in a conversation. After emotion recognition, a two-stream attention model is employed to extract the emotion causal pairs given the target emotion for subtask 2 while MuTEC is employed to extract causal span for subtask 1. Our approach achieved first place for both of the two subtasks in the competition.

1 Introduction

Comprehending emotions plays a vital role in developing artificial intelligence with human-like capabilities, as emotions are inherent to humans and exert a substantial impact on our thinking, choices, and social engagements (Wang et al., 2023b). Dialogues, being a fundamental mode of human communication, abound with a variety of emotions (C. et al., 2008; Poria et al., 2019; Zahiri and Choi, 2017; Li et al., 2017; Xia and Ding, 2019; Ding et al., 2020; Wei et al., 2020; Fan et al., 2020). Going beyond simple emotion identification, unraveling the underlying catalysts of these emotions within conversations represents a more complex and less-explored challenge (Wang et al., 2023b). Hence, (Wang et al., 2023a, 2024) introduces a novel undertaking known as Recognizing Emotion Cause in Emotion-Cause-in-Friends (ECF). ECF contains 1,344 conversations and 13,509 utterances where 9,272 emotion-cause pairs are annotated, covering textual, visual, and acoustic modalities. All utterances are annotated by one of the seven emotion labels, which are neutral, surprise, fear, sadness, joy, disgust, and anger. Within ECF, a significant task is identified as Emotion-Cause Pair Extraction in Conversations (ECPEC). ECPEC is responsible for identifying causal expressions related to a specific utterance in conversations where the emotion is implicitly expressed. ECPEC provides two Multimodal Emotion Cause Analysis in Conversations (ECAC) subtasks:

- Subtask 1: Textual Emotion-Cause Pair Extraction in Conversations. Given a conversation containing the speaker and the text of each utterance $U = [U_1, U_2, ...U_n]$, the model is aim to predict emotion-cause pairs, which include emotion utterance's emotion category and the textual cause span in a specific cause utterance (e.g. U3_joy, U2_"You made up!").
- Subtask 2: Multimodal Emotion Cause Analysis in Conversations. Given a conversation including the speaker, text and audio-visual clip for each utterance, the model is aim to predict emotion-cause pairs, which include emotion category and a cause utterance (e.g. U5_Disgust, U5).

To address the above problem, Wang et al. (2023a) proposed a two-step approach. First, they extract the emotional utterances and causal utterances by a multi-task learning framework and then pair and filter them. Zhao et al. (2023) proposes an end-to-end method by leveraging multi-task learning in a pipeline manner. However, these methods still suffer from low evaluation performances.

Motivated by the phenomenon that the performance of the emotion recognition of utterances in

⋆: equal contributions. ✉: Corresponding Author.
Shen Zhang is in charge of the basic subtask-emotion recognition in conversation (ERC) and Haojie Zhang is responsible for the pipeline framework and causal pair extraction and causal span extraction subtasks.

Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024), pages 536–546

a conversation harnessed by the traditional manner is generally poor, we design a new pipeline framework. Firstly we utilize the Llama-2-based InstructERC (Lei et al., 2023a) to extract the emotion category of each utterance in a conversation. Then we consider the emotion causal pair extraction as the causal emotion entailment subtask and employ a two-stream attention model to extract the emotion causal pairs given the target emotion. For the causal span extraction, we employ MuTEC (Bhat and Modi, 2023) which is an end-to-end multi-task learning framework.

2 Related Works

2.1 Emotion Recognition in Conversation

Emotion recognition in conversation (ERC), which is a task to predict emotions of utterances during conversations, is crucial in both of the two ECAC subtasks. The existing methods can be divided into graph-based, RNN-based, Transformer-based, LLM-based, and knowledge-injecting methods.

Graph-based methods (Shen et al., 2021b; Li et al., 2024; Zhang et al., 2019; Taichi et al., 2020; Ghosal et al., 2019) aims to represent the correlations between emotions of utterances and speakers in the conversations. RNN-based methods (Hu et al., 2023; Lei et al., 2023c; Majumder et al., 2019; Hazarika et al., 2018; Poria et al., 2017) using GRU and LSTM (Wang et al., 2020) to capture the dependency of interlocutors and emotions of utterances. To model the emotional states during long-range context, Transformer-based methods (Song et al., 2022; Liu et al., 2023b; Chudasama et al., 2022; Shen et al., 2021a; Hu et al., 2022) utilize encoder-decoder framework or encoder-only models, such as BERT (Li et al., 2020) and RoBERTa (Kim and Vossen, 2021), to establish the correlation between long-range emotional states during conversations. Considering more than seven utterances in single conversation input, InstructERC (Lei et al., 2023b) defines the ERC task as a generative task based on LLMs, which unifies emotion labels between three common ERC datasets and utilizes auxiliary tasks (speaker identification and emotion prediction) by using instruction template to capture speaker relationships and emotional states in future utterances. Knowledge-injecting methods (Freudenthaler et al., 2022; Ghosal et al., 2020; Zhong et al., 2019; Zhu et al., 2021; Lei et al., 2023b) use external knowledge to analyze conversation scenarios.

2.2 Emotion Causes in Conversations

Poria et al. (2021) introduces the task of recognizing emotion causes in conversations and introduce two novel sub-tasks: Causal Span Extraction (CSE) and Causal Emotion Entailment (CEE), designed to identify the emotion cause at the span-level and utterance-level, respectively.

Causal Emotion Entailment Poria et al. (2021) define CEE as a classification task for utterance pairs and establish robust Transformer-based baselines for it. Wang et al. (2023a) introduces a multimodality conversation dataset Emotion-Cause-in-Friends (ECF) and propose a two-step approach to extract the causal pairs. They first extract the emotion utterances and the potential causal utterances individually and then pair and filter them. Li et al. (2022) introduce the social commonsense knowledge to propagate causal clues between utterances. Zhao et al. (2023) propose the Knowledge-Bridged Causal Interaction Network (KBCIN), which integrates commonsense knowledge (CSK) as three bridges called semantics-level bridge, emotion-level bridge and action-level bridge.

Causal Span Extraction involves identifying the causal span (emotion cause) for a given non-neutral utterance. Poria et al. (2021) first introduces the subtask and employs the pre-trained Transformer-based model to formulate the Causal Span Extraction as the Machine Reading Comprehension (MRC). Bhat and Modi (2023) propose a multi-task learning framework to extract the causal pairs and causal span in an utterance in a joint end-to-end manner. Besides, they also propose a two-step approach consisting of Emotion Prediction (EP), followed by Causal Span (CSE).

3 System Overview

3.1 System Architecture

The overview of the architecture of our proposed model is shown in Figure 1. The InstructERC aims to extract the emotion of utterances. TSAM model is a two-stream attention model utilized to extract the causal pairs given the predicted emotion utterance. The MuTEC is an end-to-end network designed to extract the causal span based on the causal pair extraction.

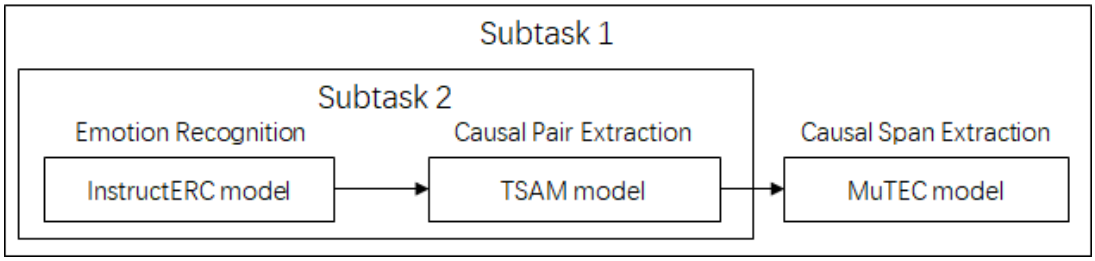

Figure 1: The overview of proposed model framework.

3.2 Emotion Recognition in Conversations

3.2.1 InstructERC for Emotion Recognition

InstructERC (Lei et al., 2023b) reformulate the ERC task from a discriminative framework to a generative framework and design a prompt template which comprises job description, historical utterance window, label set and emotional domain retrieval module. Besides emotion recognition task, InstructERC also utilizes speaker identification and emotion prediction tasks for ERC task. The performance of emotional domain retrieval module, which is based on Sentence BERT (Reimers and Gurevych, 2019), rely on the abundance of corpus. Taking into account that no additional data can be used, we only retain job description, historical utterance window and label statement in the instruct template.

3.2.2 Hierarchical Emotion Label

The hierarchical classification structure is shown in Figure 2. The emotion labels in dataset can be split into three categories: neutral, positive and negative, which positive set consists of surprise and joy while negative set includes fear, sadness, disgust and anger.

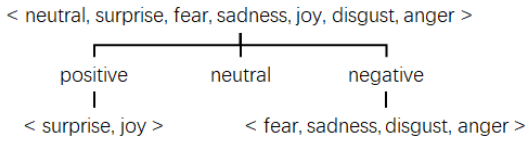

Figure 2: The Hierarchical Structure of Emotion labels.

3.2.3 Auxiliary Tasks and Instruct Design

Auxiliary tasks are proven as one of the efficient data augment methods (Lei et al., 2023b). Besides emotion recognition and speaker identification tasks, we add three auxiliary tasks in training data: sub-label recognition, positive recognition, and negative recognition tasks. The instruct template is depicted in Figure 3.

For emotion recognition and speaker identification task, we follow the format of instruct template in InstructERC, which consists of job description, historical content and label statement. For sub-label recognition (SR), positive recognition (PR) and negative recognition (NR) tasks, we utilize the corresponding label set which is mentioned in Section 3.2.2 to replace the label statement separately. The number of Speakers in the dataset is 304. The number of utterances from other speakers except the protagonist is far lower than the number of protagonists. Therefore, we unified all speakers other than the protagonist into 'Others'.

Visual data also plays an essential role in ERC. For video clips, we utilize LLaVA to generate descriptions of background, speaker movement and personal state. Therefore, we add background description, movement description and personal state description in instruct template. The background exhibits the information of scene in the conversation. The movement description depicts the action of speakers during corresponding utterances. The personal state description provides the observation of speakers' facial expressions. Considering the influence of the context, we have generated two sets of descriptions. The input of the first group only includes the clips corresponding to the utterances, while the second group adds the clips sequence corresponding to the historical utterances to the input of second group.

3.3 Emotion Cause Span Extraction

Emotion cause span extraction aims to extract the start position and end position of the causal utterance in a conversation. Typically, we can utilize a pipeline framework which firstly predicts the emotion and then predicts the cause span. For the cause span predictor, we can use SpanBERT (Joshi et al., 2020), RoBERTa (Liu et al., 2019) as the feature extractor and employ two heads on the top of them to extract the start and end positions given the causal

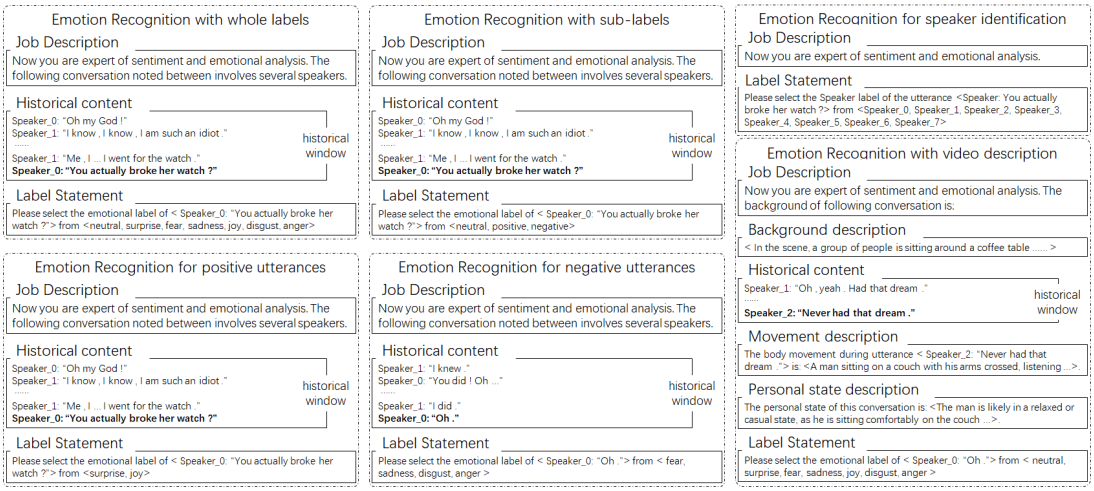

Figure 3: The Schematic of Instruct Template for ERC.

utterance. The two-step model offers an advantage in its modularity, allowing the application of distinct architectures for the emotion predictor and cause span predictor. However, it comes with two drawbacks: 1) Errors in the first step can propagate to the next, and 2) This approach assumes that emotion prediction and cause-span prediction are mutually exclusive tasks. In our system, we follow MuTEC Bhat and Modi (2023) and use an end-to-end framework in a joint multi-task learning manner to extract the causal span in a conversation.

During the training period, the input comprises the target utterance U_t, the candidate causes utterance U_i, and the historical context. MuTEC employs a pre-trained model (PLM) to extract the context representations. For emotion recognition, which is an auxiliary task, it employs a classification head on the top of the PLM. The end position is predicted by the prediction head of the concatenated representations of the given start index and the sequence output from the PLM. In this stage, the golden start index is used as the start index. The training loss is a linear combination of the loss for cause-span prediction and emotion prediction: $\mathcal{L}_{Loss} = \mathcal{L}_{CSE} + \beta\mathcal{L}_{Emotion}$.

During the inference period, as the start index is unknown, it uses top k start indices as the candidate start indices and gets k candidate end indices. Finally, it gets the final start-end indices by argmaxing the $k \times k$ start-end pairs.

3.4 Emotion-Cause Pair Extraction

3.4.1 TSAM Model

In our pipeline framework, for Subtask2, we first extract the emotion of the utterance and then extract the causal pairs given the emotional utterance in a conversation. The causal pairs extraction is typically modelled as the causal emotion entailment (CEE) task. In our system, we employ TSAM model from Zhang et al. (2022) as the causal pair extractor. TSAM mainly comprises three modules: Speaker Attention Network (SAN), Emotion Attention Network (EAN), and Interaction Network (IN). The EAN and SAN integrate emotion and speaker information simultaneously, and the subsequent interaction module efficiently exchanges pertinent information between the EAN and SAN through a mutual BiAffine transformation (Dozat and Manning, 2016).

Contextual Utterance Representation The pre-trained RoBERTa is employed as the utterance encoder, and we obtain contextual utterance representations by inputting the entire conversational history U_t, into the RoBERTa (Liu et al., 2019), separated by a special token [CLS], where $i = 0, 1, 2, ..., t$. We use the representation of [CLS] as the contextual representation of the utterance, which can be denoted as $h_u^i \in H_u$.

Emotion Attention Network To represent emotions, the EAN utilizes an emotion embedding network as the extractor of emotion representations, $X_e^k = Embedding(e_k)$, where e_k represents k-th emotion label. The embedding network can be considered as the lookup-table operation. The emotion embedding matrix is initialized using a random initializer and is fine-tuned throughout the training process. Employing a multi-head attention mechanism (Devlin et al., 2018), the EAN treats utterance representations as query vectors and emotion

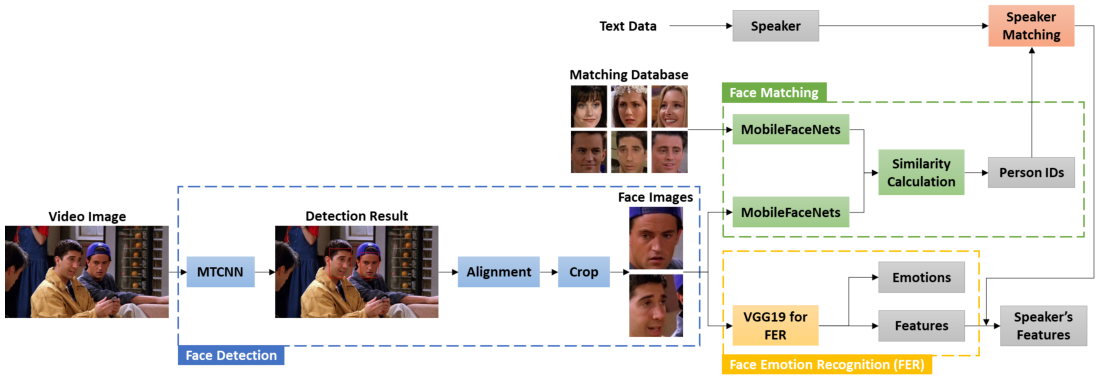

Figure 4: The framework of the face module.

representations as key and value vectors. The calculation process of the EAN mirrors that of a typical multi-head self-attention module (MHSA).

$$H_e = MHSA(Q, K, V) \tag{1}$$

where $Q = H_u, K = V = H_e$.

Speaker Attention Network The SAN facilitates interactions between utterances to incorporate speaker information by applying attention over the speaker relation graph. There are two types of relation edges: (1) Intra-relation type, which signifies how the utterance influences other utterances, including itself, expressed by the same speaker; (2) Inter-relation type, indicating how the utterance influences those expressed by other speakers. The speaker representation given a relationship can be formulated by the graphical attention mechanism (Zhang et al., 2022).

$$\begin{aligned} h_s^i &= \sum_{i \in \mathcal{R}} \sum_{j \in \mathcal{N}_i^r} \alpha_{ijr} W_r h_j^u \\ \alpha_{ijr} &= softmax(ReLu(\alpha_r^T W_r [h_i^u || h_j^u])) \end{aligned} \tag{2}$$

Interaction Network To efficiently exchange pertinent information between the EAN and SAN, a mutual Bi-Affine transformation is applied as a bridge (Dozat and Manning, 2016). In our Interaction Network, we integrate a masking mechanism to accommodate the existence of empty utterance speakers in some instances, which differs from the original approach. We denote this approach as the Masking Interaction Network (MIN).

$$\begin{aligned} \dot{H}_e &= softmax(Mask(H_e W_1 H_s^T)) H_s \\ \dot{H}_s &= softmax(Mask(H_s W_2 H_e^T)) H_e \end{aligned} \tag{3}$$

Cause Predictor The ultimate utterance representation for U_i is acquired by concatenating the output $\dot{H}_e$ and $\dot{H}_s$ from the L-layer TSAM. Subsequently, the concatenated vector undergoes classification using a fully-connected network. Given the target utterance U_i, the causal probability of the U_j can be formulated as follows:

$$p_{i,j} = sigmoid(fc(H_s^j || H_e^j)) \tag{4}$$

Multi-task Learning Auxiliary Task (MTLA) One drawback of the pipeline framework is that the extraction of utterance emotion and causal information are treated as separate tasks, potentially limiting the exploration of implicit relationships between them. Therefore, we incorporate emotion prediction as an auxiliary task within a multi-task learning framework. For emotion prediction, we utilize a classification head atop the Transformer-based model and apply the Dice loss (Li et al., 2019) as the multi-category classification loss.

3.5 Infusion of Video and Audio Information

The video data potentially carries rich knowledge for emotion analysis and existing research (Caridakis et al., 2007) has underscored the significance of multi-modal information in augmenting the semantic prediction capabilities of models. Our study leverages the visual and auditory cues present in conversational contexts with the aim of bolstering the efficacy of our language models in emotion analysis tasks.

3.5.1 Embedding and Concating Strategy

We set up specific embedding and fusion strategies for different language models. For BERT, we use the concatenation of textual and multi-modal features in the hidden layer. For Large Language Models (LLMs), our approach is characterized by the utilization of visual captions as supportive prompts, thereby furnishing the LLMs with an enriched informational context.

Models	LLM	w-avg F1	Accuracy
Origin InstructERC	Llama-2-7B-chat	53.83	50.87
Origin InstructERC	Llama-2-13B-chat	55.50	48.93
Ours-ERC-7B	Llama-2-7B-chat		
+ 3 auxiliary tasks		56.88	61.38
+ 3 auxiliary tasks & historical clips desc		57.74	57.02
+ 3 auxiliary tasks & utterance clips desc		58.42	57.92
Ours-ERC-13B	Llama-2-13B-chat		
+ 3 auxiliary tasks		57.85	**61.45**
+ 3 auxiliary tasks & historical clips desc		**58.64**	60.83
+ 3 auxiliary tasks & utterance clips desc		58.50	61.04

Table 1: Results of ERC task on test set without neutral utterances.

3.5.2 Extract Audio Feature Set

Audio data contains valuable information for emotion analysis, including tone, pitch, speed, and intensity of speech, as well as non-linguistic sounds and pauses, which together convey rich emotional cues. We use openSMILE (Eyben et al., 2010) to extract two comprehensive feature sets: GeMAPS (Eyben et al., 2016) and ComParE (Schuller and Batliner, 2013). GeMAPS is proposed for its effectiveness in capturing emotion-relevant vocal characteristics and ComParE encompasses a wide range of descriptors.

3.5.3 Video Image to Text

Integrating multi-modal features directly into the hidden layers of Large Language Models (LLMs) presents a significant challenge, primarily due to the prohibitive requirements for data and computational resources, such as GPUs. Although some finetuning strategies like prompt tuning could achieve it by addiing features to the input layer, we convert video to text with captioning where we can leverage our well-trained ERC model.

The performance of image captioning has been further enhanced with the outstanding NLU ability of LLMs. Large VLMs like LLaVA (Liu et al., 2023a) provide GPT-4 level multi-modal capability by visual instruction tuning. Furthermore, the Audio-Visual Language Model, Video-Llama (Zhang et al., 2023a), integrates both visual and audio encoders, enabling the comprehensive fusion of entire video content into LLMs. Without further training the VLMs as lack data, a well-designed prompt instructs the model to generate an emotion-related description. Our prompt asks the model to generate information from the front-ground event and place to character movements, the main character, facial expression, and finally emotion. The use of Chain-of-Thought (Wei et al., 2022) prompting further guides the model through a step-by-step process to derive the final emotion label. The output generated at each step is then incorporated into the ERC model, enriching it with a more detailed informational context.

3.5.4 Video image to Face Embedding

The faces in the video images contain rich emotion-related information, so pre-trained models are used to extract the face embeddings and correspond the identity of the face to the speaker in the text. The framework of the face module is shown in Figure 4.

Firstly, the Multi-Task Convolutional Neural Network (MTCNN) (Zhang et al., 2016) is used to detect the bounding boxes and key points of the faces. Next, the face images are affine transformed to a forward and intermediate state, and the faces are cropped and resized. The cropped images are then used for two subtasks: face matching and Face Emotion Recognition (FER). During face matching, two images of each protagonist are selected to build a matching database. With the help of MobileFaceNets (Chen et al., 2018), the embeddings of the face images are extracted, and the identity of each face image is obtained by calculating its similarity with the embeddings of faces in the matching database. During FER, the emotion-related embedding of the face image corresponding to the speaker is extracted by VGG19 (Simonyan and Zisserman, 2015) for subsequent multimodal analysis. When the speaker is a supporting character that is not included in the matching database, the features of the face image with the largest area are selected. When no face is detected or the speaker cannot be matched, the output features are filled with 0.

3.6 Model Ensemble

Ensembling models has been proven to be effective in boosting system performance across various tasks (Zhang et al., 2023b). For the extraction of

Model	Pre-trained Model	Test Pos.F1*	Eval Pos.F1**
Origin TSAM	RoBERTa-base	74.3	-
Ours-CEE	base		
+MIN	RoBERTa-base	75.5	-
+MIN & MTLA	RoBERTa-base	75.9	-
+MIN & MTLA	RoBERTa-large	76.9	-
+MIN & MTLA & Ensemble	RoBERTa-large	**78.0**	**38.7**
Ours-CSE	BERT-base	-	31.62 (w-avg.)
Ours-CSE	RoBERTa-large	-	**32.23 (w-avg.)**

* The results are based on ground truth emotion labels.
** The results are based on emotion labels given by ERC.

Table 2: Results of our models for the causal emotion entailment subtask.

causal pairs, we utilize various models for ensemble learning. We utilize a majority voting mechanism to determine the final prediction, aiming for optimal performance on the test dataset.

4 Experimental Setup

4.1 Training Data

The split of dataset is same as SHARK (Wang et al., 2023b). The ECF dataset is divided into training, validation and test sets, which incclude 9966, 1087, 2566 utterances.

4.2 Training Details

For ERC task, we use InstructERC with Llama-2-7B-chat and LLamMA2-13B-chat, which retain default parameters. We finetune ERC model by peft on single A100 with batch size 8. The length of historical window is 12.

For both the causal emotion entailment subtask and the causal span extraction subtask, we adopt the default hyperparameter settings from the respective original papers. We found that conducting a hyperparameter grid search did not yield any additional performance improvements.

5 Results and Discuss

5.1 Emotion Recognition

We use weight average F1 score and accuracy to evaluate the performance of the model. It should be noted that according to the rules of the competition, we removed the neutral utterances when computing F1 score and accuracy. The result of ERC on test set is shown in Table 1. All models is trained on four auxiliary tasks mentioned by in Section 3.2.3. The best weight average F1 score is 58.64, which is achieved by Llama-2-13B with historical clips descriptions. The descriptions which contains information with the emotions of speakers improve 0.79 (from 57.85 to 58.64). As for accuracy, the Llama-2-13B without video clips descriptions achieves the highest score of 61.45. Compared with InstructERC's training data strategy, we have added additional auxiliary tasks and improve 12.52 on accuracy.

5.2 Emotion Cause Span Extraction

We utilize an end-to-end framework for cause span extraction and achieve a final performance of 32.23 in weighted average proportional F1 score on the official evaluation dataset as is shown in the Table 2. Our result significantly surpasses the result of 26.40 above $\sim +6.0$ achieved by the second-place participant. Furthermore, our results achieved the highest scores across all other official evaluation metrics, validating the effectiveness of our approach for subtask 1.

5.3 Causal Emotion Entailment

In our initial experiments focusing solely on text modality, we utilize the TSAM model as our baseline for the causal pair extraction subtask. As is shown in Table 2, After incorporating the MIN, our positive F1 score improves by $+1.2$. Furthermore, with the introduction of emotional multi-task learning as an auxiliary task, our result sees an additional improvement of $+0.4$. Furthermore, we achieve an additional improvement of approximately $\sim +1.1$ in the official final evaluation dataset through model ensembling.

We also conduct experiments involving other modalities, including audio and vision, as is show in Table 3. For both audio and vision features, we concatenate them with the pure textual features. Regarding audio, we experiment with two public feature sets: GeMAPS and ComParE. The GeMAPS feature has a dimension of 62, while the ComParE

Modality	Feature Set	Feature Selection	Feature Dimension	Test Pos.F1
Audio	GeMAPS	×	62	39.0
	ComParE	×	top 1000	62.4
	ComParE	√	352	67.6
	ComParE	√	296	70.5
	ComParE	√	128	**73.9**
Vision	Max Img	×	128	70.7
	Speaker Img	×	128	74.3
	Emotional Speaker Img	×	512	**74.8**

Table 3: Results of multi-modality experiments for the causal emotion entailment subtask.

feature has a dimension of 6373. For the ComParE features, we employ an L1-based logistic regression classifier for feature selection, and we find that the best performance is achieved with a feature selection dimension of 128, resulting in a performance of 73.9. For the vision modality, we achieve a performance of 74.8, which is comparable to the result of the audio modality. However, upon introducing either audio or visual modalities, we observe a decreasing trend compared to the pure textual modality. This observation inspires us to develop a more reasonable approach to incorporate multi-modality in conversation analysis.

6 Conclusion

In this paper, we propose a joint pipeline framework for Subtask1 and Subtask2. Firstly, we utilize the Llama-2-based Instruct ERC model to extract the emotional content of utterances in a conversation. Next, we employ a two-stream attention model to identify causal pairs based on the predicted emotional states of the utterances. Lastly, we adopt an end-to-end framework using a multi-task learning approach to extract causal spans within a conversation. Our approach achieved first place in the competition, and the effectiveness of our approach is further confirmed by the ablation study. In future work, we plan to explore the integration of audio and visual modalities to enhance the performance of the task.

References

Ashwani Bhat and Ashutosh Modi. 2023. Multi-task learning framework for extracting emotion cause span and entailment in conversations. In *Transfer Learning for Natural Language Processing Workshop*, pages 33–51.

Busso C., Bulut M., and Lee et al. CC. 2008. IEMOCAP: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42:335–359.

George Caridakis, Ginevra Castellano, Loic Kessous, Amaryllis Raouzaiou, Lori Malatesta, Stelios Asteriadis, and Kostas Karpouzis. 2007. Multimodal emotion recognition from expressive faces, body gestures and speech. In *Artificial Intelligence and Innovations 2007: from Theory to Applications*, pages 375–388, Boston, MA. Springer US.

Sheng Chen, Yang Liu, Xiang Gao, and Zhen Han. 2018. Mobilefacenets: Efficient cnns for accurate real-time face verification on mobile devices. In *Biometric Recognition: 13th Chinese Conference, CCBR 2018, Urumqi, China, August 11-12, 2018, Proceedings 13*, pages 428–438. Springer.

Vishal Chudasama, Purbayan Kar, Ashish Gudmalwar, Nirmesh Shah, Pankaj Wasnik, and Naoyuki Onoe. 2022. M2fnet: Multi-modal fusion network for emotion recognition in conversation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 4652–4661.

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*.

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020. ECPE-2D: Emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction. In *In Association for Computational Linguistics (ACL)*, page 3161–3170.

Timothy Dozat and Christopher D Manning. 2016. Deep biaffine attention for neural dependency parsing. *arXiv preprint arXiv:1611.01734*.

Florian Eyben, Klaus R. Scherer, Björn W. Schuller, Johan Sundberg, Elisabeth André, Carlos Busso, Laurence Y. Devillers, Julien Epps, Petri Laukka, Shrikanth S. Narayanan, and Khiet P. Truong. 2016. The geneva minimalistic acoustic parameter set (gemaps) for voice research and affective computing. *IEEE Transactions on Affective Computing*, 7(2):190–202.

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.

Chuang Fan, Chaofa Yuan, Jiachen Du, Lin Gui, Min Yang, and Ruifeng Xu. 2020. Transition-based directed graph construction for emotion-cause pair extraction. In *In Association for Computational Linguistics (ACL)*, page 3707–3717.

Bernhard Freudenthaler, Jorge Martinez-Gil, Anna Fensel, Kai Höfig, Stefan Huber, and Dirk Jacob. 2022. KI-Net: Ai-based optimization in industrial manufacturing—a project overview. In *International Conference on Computer Aided Systems Theory*, pages 554–561. Springer.

Deepanway Ghosal, Navonil Majumder, Alexander Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. Cosmic: Commonsense knowledge for emotion identification in conversations. *arXiv preprint arXiv:2010.02795*.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. 2019. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. *arXiv preprint arXiv:1908.11540*.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. ICON: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2594–2604.

Dou Hu, Yinan Bao, Lingwei Wei, Wei Zhou, and Songlin Hu. 2023. Supervised adversarial contrastive learning for emotion recognition in conversations. *arXiv preprint arXiv:2306.01505*.

Guimin Hu, Ting-En Lin, Yi Zhao, Guangming Lu, Yuchuan Wu, and Yongbin Li. 2022. UniMSE: Towards unified multimodal sentiment analysis and emotion recognition. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7837–7851.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the association for computational linguistics*, 8:64–77.

Taewoon Kim and Piek Vossen. 2021. EmoBERTa: Speaker-aware emotion recognition in conversation with roberta.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2023a. Instructerc: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *arXiv preprint arXiv:2309.11911*.

Shanglin Lei, Guanting Dong, Xiaoping Wang, Keheng Wang, and Sirui Wang. 2023b. InstructERC: Reforming emotion recognition in conversation with a retrieval multi-task llms framework. *arXiv preprint*, arXiv:2309.11911.

Shanglin Lei, Xiaoping Wang, Guanting Dong, Jiang Li, and Yingjian Liu. 2023c. Watch the speakers: A hybrid continuous attribution network for emotion recognition in conversation with emotion disentanglement. In *2023 IEEE 35th International Conference on Tools with Artificial Intelligence (ICTAI)*, pages 881–888.

Jiang Li, Xiaoping Wang, Guoqing Lv, and Zhigang Zeng. 2024. GraphCFC: A directed graph based cross-modal feature complementation approach for multimodal conversational emotion recognition. *IEEE Transactions on Multimedia*, 26:77–89.

Jiangnan Li, Fandong Meng, Zheng Lin, Rui Liu, Peng Fu, Yanan Cao, Weiping Wang, and Jie Zhou. 2022. Neutral utterances are also causes: Enhancing conversational causal emotion entailment with social commonsense knowledge. *arXiv preprint arXiv:2205.00759*.

Jingye Li, Meishan Zhang, Donghong Ji, and Yijiang Liu. 2020. Multi-task learning with auxiliary speaker identification for conversational emotion recognition. *arXiv preprint arXiv:2003.01478*.

Xiaoya Li, Xiaofei Sun, Yuxian Meng, Junjun Liang, Fei Wu, and Jiwei Li. 2019. Dice loss for data-imbalanced nlp tasks. *arXiv preprint arXiv:1911.02855*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. DailyDialog a manually labelled multi-turn dialogue dataset. *arXiv preprint*, arXiv:1710.03957.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning.

Xiao Liu, Jian Zhang, Heng Zhang, Fuzhao Xue, and Yang You. 2023b. Hierarchical dialogue understanding with special tokens and turn-level attention. *arXiv preprint arXiv:2305.00262*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria. 2019. DialogueRNN: An attentive rnn for emotion detection in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6818–6825.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the*

55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 873–883.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. MELD: A multimodal multi-party dataset for emotion recognition in conversations. *arXiv preprint*, arXiv:1810.02508.

Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, et al. 2021. Recognizing emotion cause in conversations. *Cognitive Computation*, 13:1317–1332.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Bjorn Schuller and Anton Batliner. 2013. *Computational Paralinguistics: Emotion, Affect and Personality in Speech and Language Processing*, 1st edition. Wiley Publishing.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021a. DialogXL: All-in-one xlnet for multi-party conversation emotion recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13789–13797.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021b. Directed acyclic graph network for conversational emotion recognition. *arXiv preprint arXiv:2105.12907*.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Xiaohui Song, Longtao Huang, Hui Xue, and Songlin Hu. 2022. Supervised prototypical contrastive learning for emotion recognition in conversation. *arXiv preprint arXiv:2210.08713*.

Ishiwatari Taichi, Yasuda Yuki, Miyazaki Taro, and Goto Jun. 2020. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7360–7370. Association for Computational Linguistics.

Fanfan Wang, Zixiang Ding, Rui Xia, Zhaoyu Li, and Jianfei Yu. 2023a. 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.

Fanfan Wang, Jianfei Yu, and Rui Xia. 2023b. Generative emotion cause triplet extraction in conversations with commonsense knowledge. In *In Findings of the Association for Computational Linguistics: EMNLP 2023*, page 3952–3963.

Yan Wang, Jiayu Zhang, Jun Ma, Shaojun Wang, and Jing Xiao. 2020. Contextualized emotion recognition in conversation as sequence tagging. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 186–195.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Ed Huai hsin Chi, F. Xia, Quoc Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. *ArXiv*, abs/2201.11903.

Penghui Wei, Jiahao Zhao, and Wenji Mao. 2020. Effective inter-clause modeling for end-to-end emotion-cause pair extraction. In *In Association for Computational Linguistics (ACL)*, page 3171–3181.

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.

Sayyed M. Zahiri and Jinho D. Choi. 2017. Emotion detection on tv show transcripts with sequence-based convolutional neural networks. *arXiv preprint*, arXiv:1708.04299.

Dong Zhang, Liangqing Wu, Changlong Sun, and et.al. 2019. Modeling both context-and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *IJCAI*, pages 5415–5421.

Duzhen Zhang, Zhen Yang, Fandong Meng, Xiuyi Chen, and Jie Zhou. 2022. Tsam: A two-stream attention model for causal emotion entailment. *arXiv preprint arXiv:2203.00819*.

Hang Zhang, Xin Li, and Lidong Bing. 2023a. Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. *arXiv preprint arXiv:2306.02858*.

Haojie Zhang, Xiao Li, Renhua Gu, Xiaoyan Qu, Xiangfeng Meng, Shuo Hu, and Song Liu. 2023b. Samsung research china-beijing at semeval-2023 task 2: An al-r model for multilingual complex named entity recognition. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 114–120.

Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503.

Weixiang Zhao, Yanyan Zhao, Zhuojun Li, and Bing Qin. 2023. Knowledge-bridged causal interaction network for causal emotion entailment. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(11):14020–14028.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 165–176.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1571–1582.