Multimodal Emotion-Cause Pair Extraction in Conversations

Fanfan Wang[*]   **Zixiang Ding**[*]   **Rui Xia**[†]   **Zhaoyu Li**   **Jianfei Yu**

School of Computer Science and Engineering,
Nanjing University of Science and Technology, China
{ffwang, dingzixiang, rxia, zyli, jfyu}@njust.edu.cn

Abstract

Emotion cause analysis has received considerable attention in recent years. Previous studies primarily focused on emotion cause extraction from texts in news articles or microblogs. It is also interesting to discover emotions and their causes in conversations. As conversation in its natural form is multimodal, a large number of studies have been carried out on multimodal emotion recognition in conversations, but there is still a lack of work on multimodal emotion cause analysis. In this work, we introduce a new task named Multimodal Emotion-Cause Pair Extraction in Conversations, aiming to jointly extract emotions and their associated causes from conversations reflected in multiple modalities (text, audio and video). We accordingly construct a multimodal conversational emotion cause dataset, Emotion-Cause-in-Friends, which contains 9,272 multimodal emotion-cause pairs annotated on 13,509 utterances in the sitcom *Friends*. We finally benchmark the task by establishing a baseline system that incorporates multimodal features for emotion-cause pair extraction. Preliminary experimental results demonstrate the potential of multimodal information fusion for discovering both emotions and causes in conversations.

1 Introduction

In the field of textual emotion analysis, previous research mostly focused on emotion recognition. In recent years, emotion cause analysis, a new task which aimed at extracting potential causes given the emotions (Lee et al., 2010; Chen et al., 2010; Gui et al., 2016b) or jointly extracting emotions and the corresponding causes in pairs (Xia and Ding, 2019; Ding et al., 2020a; Wei et al., 2020; Fan et al., 2020), has received much attention. These studies were normally carried out based on news articles or microblogs. Poria et al. (2021) further introduced

an interesting task to recognize emotion cause in textual dialogues.

However, conversation in its natural form is multimodal. Multimodality is especially important for discovering both emotions and their causes in conversations. For example, we do not only rely on the speaker's voice intonation and facial expressions to perceive his emotions, but also depend on some auditory and visual scenes to speculate the potential causes that trigger the speakers' emotions beyond text. Although a large number of studies have explored multimodal emotion analysis in conversations (Busso et al., 2008; McKeown et al., 2012; Poria et al., 2019), to our knowledge, at present there is still a lack of research on multimodal emotion cause analysis in conversations.

In this work, we introduce a new task named Multimodal Emotion-Cause Pair Extraction in Conversations (MC-ECPE), with the goal to extract all potential pairs of emotions and their corresponding causes from a conversation in consideration of three modalities (text, audio and video). We accordingly construct a multimodal emotion cause dataset, Emotion-Cause-in-Friends (ECF), by using the sitcom *Friends* as the source. The ECF dataset contains 1,344 conversations and 13,509 utterances[1], where 9,272 emotion-cause pairs are annotated, covering three modalities.

Figure 1 displays a real conversation in the ECF dataset, where Chandler and his girlfriend Monica walked into the casino, hugging each other (they had a quarrel earlier but made up soon), and then started a conversation with Phoebe. In Utterance 1 (U_1 for short), Chandler said hello to Phoebe with a *Joy* emotion (the cause is greeting). Phoebe's *Surprise* emotion in U_2 is caused by the event that Chandler and Monica had made up (reflected by the textual modality in U_2). This is also the cause of Monica's *Joy* emotion in U_3 and Chandler's

---

[*] Equal contribution.
[†] Corresponding author.

[1] An utterance is a unit of speech divided by the speaker's breath or pause.
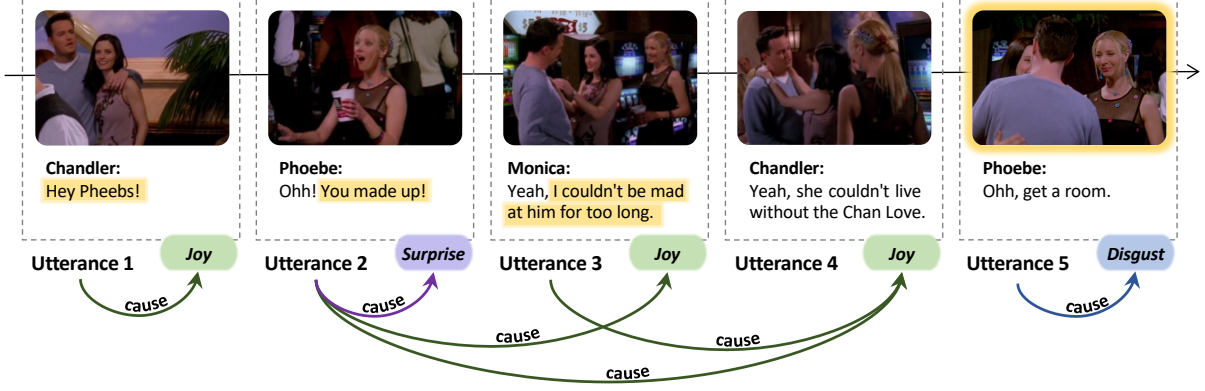
Figure 1: An example of the annotated conversation in our ECF dataset. Each arc points from the cause utterance to the emotion it triggers. We have highlighted the cause evidence which is mainly contained in a certain modality of the cause utterance.

Joy emotion in U_4. Chandler's *Joy* emotion in U_4 has another cause - Monica's opinion in U_3 ("*I couldn't be mad at him for too long*"). The cause for Phoebe's *Disgust* in U_5 is the event that Monica and Chandler were kissing in front of her. This is not explicitly expressed by the textual modality, but is mainly reflected in the visual modality of U_5. For this example, it is expected to extract a set of six utterance-level emotion-cause pairs: $\{U_1\text{-}U_1, U_2\text{-}U_2, U_3\text{-}U_2, U_4\text{-}U_2, U_4\text{-}U_3, U_5\text{-}U_5\}$.

We finally benchmark the MC-ECPE task by establishing a baseline system adapted from a representative textual ECPE approach. We incorporate multimodal features for utterance representation, extract emotion utterances and cause utterances respectively, and finally construct emotion-cause pairs. The experimental results demonstrate the effect of multimodal information fusion for discovering both emotions and causes in conversations.

2 Related Work

Emotion Cause Analysis: Emotion cause extraction (ECE) is a subtask of emotion analysis. It was originally proposed by Lee et al. (2010), with the goal to extract cause spans of a given emotion in the text. Based on the same task setting, some researchers use rule-based methods (Neviarouskaya and Aono, 2013; Li and Xu, 2014; Gao et al., 2015a,b; Yada et al., 2017) or machine learning methods (Ghazi et al., 2015; Song and Meng, 2015) to extract emotion causes in their own corpus.

By analyzing the corpus proposed by Lee et al. (2010), Chen et al. (2010) pointed out that clause may be a more suitable unit for cause annotation, and proposed to extract emotion cause at clause

granularity. After that, a lot of work based on this task setting appeared (Russo et al., 2011; Gui et al., 2014). Especially, Gui et al. (2016a) released an open Chinese emotion cause dataset. This dataset has received extensive attention and become the benchmark dataset for the ECE task. Based on this corpus, many traditional machine learning methods (Gui et al., 2016a,b; Xu et al., 2017) and deep learning methods (Gui et al., 2017; Li et al., 2018; Yu et al., 2019; Xu et al., 2019; Ding et al., 2019; Xia and Ding, 2019) were put forward.

However, there are two shortcomings in the ECE task: 1) emotions must be manually annotated before cause extraction, which greatly limits its practical application; 2) the way of annotating emotions first and then extracting causes ignores the fact that emotions and causes are mutually indicative. To solve these problems, Xia and Ding (2019) proposed a new task called emotion-cause pair extraction (ECPE), aiming at extracting potential emotions and corresponding causes from documents simultaneously. They further constructed the ECPE dataset based on the benchmark corpus for ECE (Gui et al., 2016a). After that, a lot of work on the ECPE task has been put forward to solve the shortcomings of the existing methodology (Ding et al., 2020a,b; Wei et al., 2020; Fan et al., 2020).

The above studies mostly focused on emotion cause analysis in news articles (Gui et al., 2016a; Gao et al., 2017; Bostan et al., 2020), microblogs (Cheng et al., 2017) and fictions (Gao et al., 2017; Kim and Klinger, 2018). Recently, Poria et al. (2021) introduced an interesting task of recognizing emotion cause in conversations and constructed a new dataset RECCON for this task. Considering

Dataset	Modality	Cause	Scene	# Instances
Emotion-Stimulus (Ghazi et al., 2015)	T	✓	–	2,414 sentences
ECE Corpus (Gui et al., 2016a)	T	✓	News	2,105 documents
NTCIR-13-ECA (Gao et al., 2017)	T	✓	Fiction	2,403 documents
Weibo-Emotion (Cheng et al., 2017)	T	✓	Microblog	7,000 posts
REMAN (Kim and Klinger, 2018)	T	✓	Fiction	1,720 documents
GoodNewsEveryone (Bostan et al., 2020)	T	✓	News	5,000 sentences
IEMOCAP (Busso et al., 2008)	T,A,V	✗	Conversation	7,433 utterances
DailyDialog (Li et al., 2017)	T	✗	Conversation	102,979 utterances
EmotionLines (Hsu et al., 2018)	T	✗	Conversation	14,503 utterances
SEMAINE (McKeown et al., 2012)	T,A,V	✗	Conversation	5,798 utterances
EmoContext (Chatterjee et al., 2019)	T	✗	Conversation	115,272 utterances
MELD (Poria et al., 2019)	T,A,V	✗	Conversation	13,708 utterances
MELSD (Firdaus et al., 2020)	T,A,V	✗	Conversation	20,000 utterances
RECCON-IE (Poria et al., 2021)	T	✓	Conversation	665 utterances
RECCON-DD (Poria et al., 2021)	T	✓	Conversation	11,104 utterances
ECF (ours)	T,A,V	✓	Conversation	13,509 utterances

Table 1: A summary of datasets for emotion cause analysis and emotion recognition in conversations. T, A, V stand for text, audio and video respectively.

that conversation itself is multimodal, we further propose to jointly extract emotions and their corresponding causes from conversations based on multiple modalities, and accordingly create a multimodal conversational emotion cause dataset.

Emotion Recognition in Conversations: Although there's a lack of research on multimodal emotion cause analysis, many studies have been carried out on multimodal emotion recognition using textual, acoustic, and visual modalities, especially in conversations (Busso et al., 2008; McKeown et al., 2012; Poria et al., 2019; Hazarika et al., 2018; Jin et al., 2020).

In recent years, due to the increasing amount of open conversation data, the Emotion Recognition in Conversations (ERC) task has received continuous attention in the field of NLP. So far, there have been some publicly available datasets for ERC. IEMOCAP (Busso et al., 2008) contains multimodal dyadic conversations of ten actors performing the emotional scripts. SEMAINE (McKeown et al., 2012) contains multimodal data of robot-human conversations (it does not provide emotion categories, but the attributes of four emotion dimensions). The above two datasets are relatively small in scale and do not contain multi-party conversations. DailyDialog (Li et al., 2017) is a large dataset that contains the texts of daily conversations covering 10 topics , but the neutral utterances in it account for a high proportion. EmoContext (Chat-

terjee et al., 2019) has a large total number of utterances, but only contains two-person conversations in plain text, with only three utterances in each conversation. EmotionLines (Hsu et al., 2018) contains two datasets: multi-party conversations from the sitcom *Friends* (Friends) and private chats on Facebook Messenger (EmotionPush) where all the utterances are labeled with emotion categories. Poria et al. (2019) extended EmotionLines (Friends) to the multimodal dataset MELD with raw videos, audio segments and transcripts, the size of which is moderate. Recently, Firdaus et al. (2020) constructed a large-scale multimodal conversational dataset MEISD from 10 famous TV series, where an utterance may be labeled with multiple emotions along with their corresponding intensities.

3 Task

We first clarify the definitions of emotion and cause in our work:

- **Emotion** is a psychological state associated with thought, feeling and behavioral response. In computer science, emotions are often described as discrete emotion categories, such as Ekman's six basic emotions including *Anger*, *Disgust*, *Fear*, *Joy*, *Sadness* and *Surprise* (Ekman, 1971). In conversations, emotions are usually annotated at the utterance level (Li et al., 2017; Hsu et al., 2018; Poria et al., 2019).

- **Cause** refers to the explicitly expressed event or argument that is highly linked with the corresponding emotion (Lee et al., 2010; Chen et al., 2010; Russo et al., 2011). In this work, we use an utterance to describe an emotion cause. Although we have annotated the textual span if the cause is reflected in the textual modality, we only consider utterance-level emotion and cause extraction in this work, in order to facilitate the representation and fusion of multimodal information.

There are two kinds of textual emotion cause analysis task: emotion cause extraction (ECE) (Gui et al., 2016b) and emotion-cause pair extraction (ECPE) (Xia and Ding, 2019). The goal of ECE is to extract the potential causes given the emotion annotation; while ECPE aims to jointly extract the emotions and the corresponding causes in pairs, which solves ECE's shortcoming of emotion annotation dependency and improves the performance of emotion and cause extraction.

Therefore, in this work we directly define the task of multimodal emotion cause analysis under ECPE rather than ECE. Given a conversation $D = [U_1, \ldots, U_i, \ldots, U_{|D|}]$, in which each utterance is represented by the text, audio and video , i.e., $U_i = [t_i, a_i, v_i]$, the goal of MC-ECPE is to extract a set of emotion-cause pairs:

$$\mathcal{P} = \{\ldots, U^e\text{-}U^c, \ldots\}, \tag{1}$$

where U^e denotes an emotion utterance and U^c is the corresponding cause utterance.

4 Dataset

4.1 Dataset Source

The conversations in sitcoms usually contain more emotions than other TV series and movies. Hsu et al. (2018) constructed the EmotionLines dataset from the scripts of the popular sitcom *Friends* for the ERC task. Poria et al. (2019) extended Emotion-Lines to a multimodal dataset MELD, by extracting the audio-visual clip from the source episode, and re-annotating each utterance with emotion labels.

We find that sitcoms also contain rich emotion causes, therefore we choose MELD[2] as the data source and further annotate the corresponding causes for the given emotion annotations. We drop a few conversations where three modalities are completely inconsistent in timestamps.
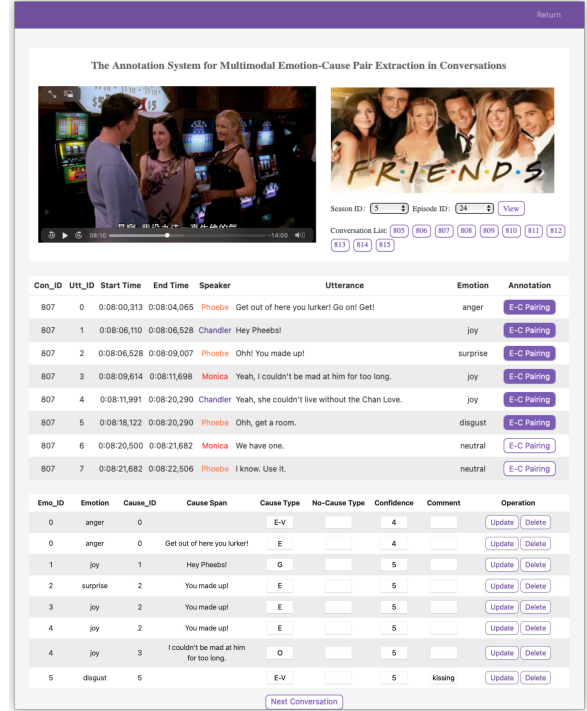
Figure 2: The interface of our developed annotation toolkit.

4.2 Annotation Procedure

Given a multimodal conversation, for the emotion (one of Ekman's six basic emotions) labeled on each utterance, the annotator should annotate the utterances containing corresponding causes, label the types of causes, and mark the textual cause spans if the causes are explicitly expressed in the textual modality.

We first develop detailed annotation instructions and then employ three annotators who have reasonable knowledge of our task to annotate the entire dataset independently. After annotation, we determine the cause utterances by majority voting and take the largest boundary (i.e., the union of the spans) as the gold annotation of textual cause span, similar as (Gui et al., 2016a; Bostan et al., 2020). If there are further disagreements, another expert is invited for the final decision.

To improve the annotation efficiency, we furthermore develop a multimodal emotion cause annotation toolkit[3]. It is a general toolkit for multimodal annotation in conversations, with the functions of multimodal signal alignment, quick emotion-cause pair selection, multiple users and tasks manage-

Type	Explanation	Modality	%	Example
Event	Something that happens in a particular situation, which is normally a fact.	T	60.14%	[U1] Phoebe: Ohh! You made up! (*Surprise*) **Emotion-Cause Pair:** (U1, U1)
		A	0.60%	[U1] Chandler: What is wrong with Emma? (*Sadness*) [U2] Monica: Oh she misunderstood, she thought she was moving to Tulsa. (*Neutral*) **Emotion-Cause Pair:** (U1, U1) *Note: Chandler heard Emma crying.
		V	7.56%	[U1] Phoebe: Ohh, get a room. (*Disgust*) **Emotion-Cause Pair:** (U1, U1) *Note: In the video, Monica and Chandler were kissing in front of Phoebe, as shown in Figure 1.
Opinion	Someone's feelings or thoughts about people or things rather than a fact.	T	25.11%	[U1] Monica: Yeah, I couldn't be mad at him for too long. (*Joy*) [U2] Chandler: Yeah, she couldn't live without the Chan Love. (*Joy*) **Emotion-Cause Pair:** (U2, U1)
Emotional Influence	The speaker's emotion is induced by the counterpart's emotion.	T,A,V	3.74%	[U1] Joey: Fine, you want to get the birds, get the birds! (*Anger*) [U2] Chandler: Not like that, I won't! (*Sadness*) **Emotion-Cause Pair:** (U2, U1)
Greeting	People tend to be happy when they meet and greet each other.	T,V	2.85%	[U1] Chandler: Hey Pheebs! (*Joy*) **Emotion-Cause Pair:** (U1, U1)

Table 2: A summary of emotion cause types. In the example, the emotion and textual cause span are colored in red and blue accordingly.

Annotator Pair	A&B	A&C	B&C	Avg.
Cohen's Kappa	0.6348	0.6595	0.6483	0.6475
Fleiss' Kappa			0.6044	

Table 3: The inter-annotator agreement for utterance-level emotion cause annotations. A, B, C represent the three annotators respectively.

Items	Number
Conversations	1,344
Utterances	13,509
Emotion (utterances)	7,528
Emotion (utterances) with cause	6,876
Emotion-cause (utterance) pairs	9,272

Table 4: Basic statistic of our ECF dataset.

ment, distributable deployment, etc. Figure 2 displays the interface of the toolkit.

4.3 Annotation Quality Assessment

To evaluate the quality of annotation, we measure the inter-annotator agreement on the full set of annotations, based on Cohen's Kappa and Fleiss' Kappa. Cohen's Kappa is used to measure the consistency of any two annotators (Cohen, 1960), while Fleiss' Kappa is used to measure the overall annotation consistency among three annotators (McHugh, 2012). The agreement scores are reported in Table 3.

It can be seen that the Kappa coefficients are all higher than 0.6, which indicates a substantial agreement between three annotators (Landis and Koch, 1977).

4.4 Dataset Statistic and Analysis

Overall Statistics: As shown in Table 4, the ECF dataset contains 1,344 conversations and 13,509 utterances from three modalities, where 7,528 emo-

tion utterances and 9,272 emotion-cause pairs have been annotated. In other words, about 55.73% of the utterances are annotated with one of the six basic emotions, and 91.34% of the emotions are annotated with the corresponding causes in our dataset. The number of pairs is larger than 6,876, which indicates that one emotion may be triggered by multiple causes in different utterances.

In Table 1, we compare our ECF dataset with the related datasets in the field of emotion cause analysis and emotion recognition in conversations, in terms of modality, scene, and size.

Emotion/Cause Distribution: For each emotion category, the proportion of emotion utterances annotated with causes is shown in Figure 4. It can be seen that the distribution of emotion categories is unbalanced, and the proportion of emotion having causes varies slightly with the emotion category.

Types of Emotion Causes: In Table 2, we furthermore summarized the emotion causes in our dataset into four types.
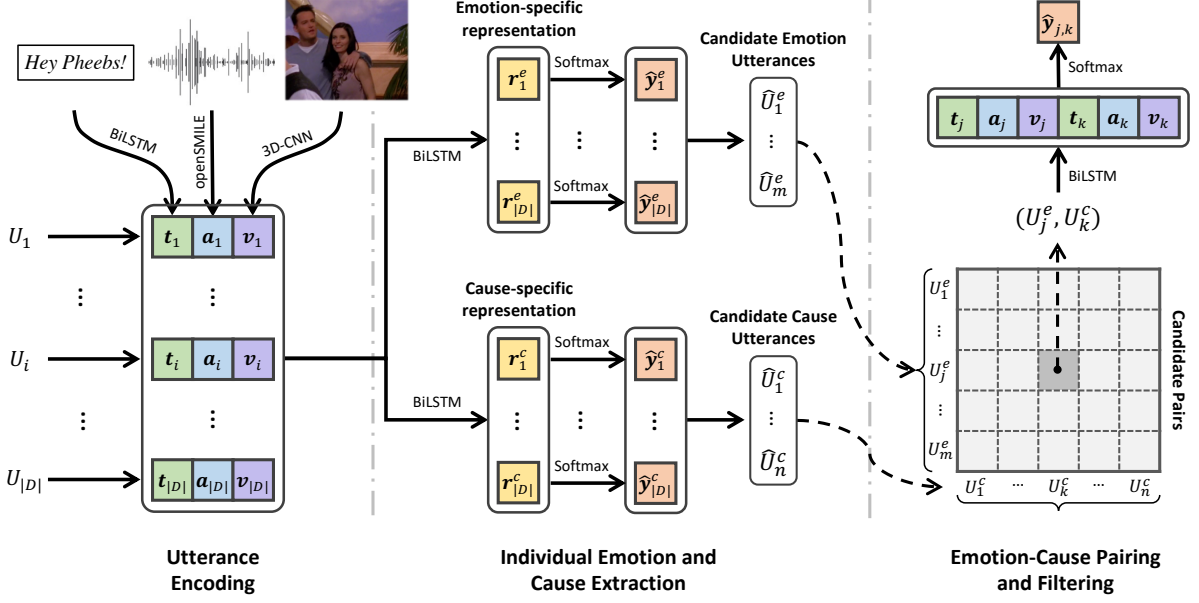
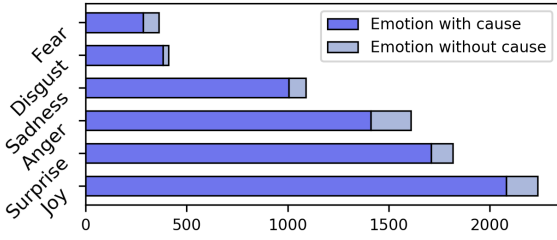Figure 3: Overview of the baseline system MC-ECPE-2steps.

Figure 4: The distribution of emotions (with/without cause) in different categories.

- **Event:** The speaker's emotion is caused by something that happens in a particular situation, which is normally a fact. This type of cause may be reflected in the three modalities.

- **Opinion:** The speaker's emotion is triggered by someone's feelings or thoughts about people or things rather than a fact. Such causes are only expressed in texts.

- **Emotional Influence:** The speaker's emotion is sometimes induced by the counterpart's emotion. This type of cause is normally embodied in three modalities jointly.

- **Greeting:** As an act of giving a sign of welcome or recognition in communication, the greeting is a cause for the *Joy* emotion in daily conversations. It can be reflected in both textual and visual modalities.

We can see that "Event" covers the largest per-centage of emotion causes (68.30%), followed by "Opinion" (25.11%), which suggests that most of the emotions in conversations are triggered by specific events or subjective opinions.

It is also worth noting that 8% of the emotion causes in our dataset are the events mainly reflected in the acoustic or visual modalities. For example, Phoebe was disgust because Monica and Chandler were kissing in front of her in the scene shown in Figure 1, we cannot speculate on such causes only based on the textual content obviously.

5 Baseline

In this section, we benchmark our MC-ECPE task by proposing a baseline system named MC-ECPE-2steps, which is adapted from a representative ECPE-2steps (Xia and Ding, 2019) approach for ECPE in news articles.

5.1 Main Framework

The main framework of MC-ECPE-2steps is shown in Figure 3.

Step 1 aims to extract a set of emotion utterances and a set of cause utterances individually via multi-task learning. We first obtain the independent utterance representations u_i through word-level encoder and then feed them into two utterance-level encoders. The hidden states r_i^e and r_i^c, which can be viewed as the emotion-specific representation and cause-specific representation of utterance U_i, are used to perform emotion prediction and cause

prediction respectively.

Step 2 performs emotion-cause pairing and filtering. We combine all predicted emotions and causes into pairs, obtain the pair representation through BiLSTM and attention mechanism, and finally filter out the pairs that do not contain a causal relationship via a feed-forward neural network.

5.2 Multimodal Features

Since the emotions and causes in this paper are defined on the multimodal utterances, we further extract the features from three modalities and then concatenate them to obtain the independent multimodal representation of each utterance, i.e., $u_i = [t_i, a_i, v_i]$.

Text: We initialize each token with pre-trained 300-dimensional GloVe vectors (Pennington et al., 2014), feed them into a BiLSTM and then obtain the textual features of each utterance t_i after an attention mechanism.

Audio: We adopt the 1611-dimensional acoustic features a_i extracted by Poria et al. (2019) using openSMILE in their MELD dataset.

Video: We apply 3D-CNN (Ji et al., 2012) to extract the 128-dimensional global scene features v_i from the video of each utterance.

6 Experiments

6.1 Settings and Metrics

The maximum number of utterances in each conversation and the maximum number of words in each utterance are both set to 35. The dimensions of word embedding and relative position are set to 300 and 50, respectively. The hidden dimension of BiLSTM is set to 100. All models are trained based on the Adam optimizer with a batch size of 32 and a learning rate of 0.005. The dropout ratio is set to 0.5, and the weight of L_2-norm regularization is set to 1e-5.

We divide the dataset into training, validation and testing sets in a ratio of 7:1:2 at the conversation level. In order to obtain statistically credible results, we repeat all the experiments 20 times and report the average results. The precision, recall and F_1 score defined in Xia and Ding (2019) are used as the evaluation metrics for the MC-ECPE task.

6.2 Overall Performance

In addition to MC-ECPE-steps, we further design four simple statistical methods for comparison,

Approach	P	R	F_1
$E_{Pred} + C_{Multi\text{-}Bernoulli}$	0.3662	0.2024	0.2605
$E_{True} + C_{Multi\text{-}Bernoulli}$	0.4940	0.2522	0.3339
$E_{Pred} + C_{Multinomial}$	0.3658	0.2024	0.2604
$E_{True} + C_{Multinomial}$	0.4933	0.2518	0.3334
MC-ECPE-2steps	0.4943	0.5376	**0.5132**
$-$Audio	0.5391	0.4778	0.5045
$-$Video	0.4989	0.5289	0.5116
$-$Audio $-$ Video	0.5565	0.4465	0.4942

Table 5: Experimental results on the MC-ECPE task.

based on the observation that there is a certain trend in the relative positions between emotion utterances and cause utterances (i.e., most cause utterances are either the emotion utterances themselves or are immediately before their corresponding emotion utterances). In the training phase, we separately train an emotion classifier, and calculate the prior probability distribution of relative positions between the cause utterances and their corresponding emotion utterances. In the testing phase, we first obtain emotion utterances in two alternative ways: 1) emotion prediction (E_{Pred}), which is based on the trained emotion classifier, 2) emotion annotations (E_{True}), which are the ground truth labels of emotion prediction. Next, the relative position is assigned to each utterance in the document, which is the position of the current utterance relative to the given emotion utterance (for example, -2, -1, 0, +1, etc.). Then we use the following two strategies to randomly select a cause utterance for each emotion utterance according to the prior probability distribution.

- $C_{Multi\text{-}Bernoulli}$: We independently carry out a binary decision for each relative position to determine whether its corresponding utterance is the cause utterance. The selection probability of each relative position is calculated from the training set.

- $C_{Multinomial}$: We randomly select a relative position from all relative positions, and its corresponding utterance is the cause utterance. The selection probability of each relative position is calculated from the training set.

The experimental results on the MC-ECPE task are reported in table 5. We can see that the statistical methods perform much poorer than our baseline system MC-ECPE-2steps, which shows that it is not enough to select the emotion cause only based on the relative position.

Approach	Anger	Disgust	Fear	Joy	Sadness	Surprise	w-avg. 6	w-avg. 4
$E_{Pred} + C_{Multi\text{-}Bernoulli}$	0.1352	0.0302	0.0146	0.1937	0.0812	0.1905	0.1457	0.1567
$E_{True} + C_{Multi\text{-}Bernoulli}$	0.1415	0.0318	0.0145	0.2091	0.0885	0.2013	0.1552	0.1670
$E_{Pred} + C_{Multinomial}$	0.1346	0.0302	0.0147	0.1928	0.0812	0.1907	0.1453	0.1563
$E_{True} + C_{Multinomial}$	0.1416	0.0322	0.0153	0.2082	0.0878	0.2017	0.1550	0.1667
MC-ECPE-2steps	**0.2837**	0.0466	0.0260	**0.3673**	**0.1820**	0.4211	**0.3000**	**0.3237**
$-$Audio	0.2400	**0.0630**	**0.0268**	0.3511	0.1511	0.4043	0.2764	0.2970
$-$Video	0.2837	0.0509	0.0262	0.3661	0.1781	**0.4213**	0.2993	0.3227
$-$Audio $-$ Video	0.2233	0.0522	0.0175	0.3417	0.1216	0.3982	0.2625	0.2827

Table 6: Experimental results on the task of MC-ECPE with emotion category. *Note: "w-avg. 4" denotes the weighted-average F_1 score of the main four emotions except *Disgust* and *Fear*.

6.3 The Effectiveness of Multimodal Features

To explore the effectiveness of different multimodal features, we conduct extended experiments and also report the results in Table 5.

It can be seen that removing the acoustic features or visual features from the baseline system MC-ECPE-2steps ($-$Audio/$-$Video) leads to a decrease in F_1 score. When both the acoustic features and visual features are removed, the F_1 score of the system even drops by about 1.9% ($-$Audio $-$ Video). Specifically, by integrating multimodal features, the baseline system can predict more causes that are reflected in the auditory and visual scenes, resulting in the great improvement in the recall rate. These results illustrate that it's beneficial to introduce multimodal information into the MC-ECPE task.

In addition, we found that the improvement brought by visual features is slightly lower than that brought by acoustic features. This indicates that it is challenging to perceive and understand the complex visual scenes in conversations, hence leaving much room for extra improvement in multimodal feature representation and multimodal fusion.

6.4 Experiments on MC-ECPE with Emotion Category

We further conduct experiments on the extended task named "MC-ECPE with emotion category" which needs to predict an additional emotion category for each emotion-cause pair. Specifically, we convert the binary emotion classification to multi-class emotion classification in the first step of MC-ECPE-steps. We first evaluate the emotion-cause pairs of each emotion category with F_1 score separately. To evaluate the overall performance, we further calculate a weighted average of F_1 scores across different emotion categories. Considering the imbalance of emotion categories in the dataset described in Section 4.4, we also report the weighted average F_1 score of the main four emotion categories except *Disgust* and *Fear*.

The experimental results on this task are reported in Table 6. An obvious observation is that the performance on *Surprise* is the best, while that on *Fear* is the worst. The performance for different emotion categories significantly varies with the proportion of emotion and cause annotation shown in Figure 4. It should be noted that emotion category imbalance is actually an inherent problem in the ERC task (Li et al., 2017; Hsu et al., 2018; Poria et al., 2019), which is of great challenge and needs to be tackled in future work.

Similar to the conclusion drawn on MC-ECPE, the performance of the baseline system is significantly reduced if not utilizing the acoustic and visual features, which demonstrates that multimodal fusion is also helpful for the task of MC-ECPE with emotion categories. Although there is much room for further improvement, our model is still effective and feasible. The relatively poor performance on this task indicates that it is more difficult to further predict the emotion categories based on MC-ECPE.

7 Conclusions and Future Work

In this work, we introduce a new emotion cause analysis task named Multimodal Emotion-Cause Pair Extraction (MC-ECPE) in Conversations. Secondly, we accordingly construct a multimodal conversational emotion cause dataset, Emotion-Cause-in-Friends (ECF), based on the American sitcom *Friends*. Finally, we establish a baseline system and demonstrate the importance of multimodal information for the MC-ECPE task.

MC-ECPE is a challenging task, leaving much room for future improvements. The focus of this work is the introduction of the task and datasets,

and we only propose a simple baseline system to benchmark the task. In the future, the following issues are worth exploring in order to further improve the performance of the task:

- How to effectively model the impact of speaker relevance on emotion recognition and emotion cause extraction in conversations?

- How to better perceive and understand the visual scenes to better assist emotion cause reasoning in conversations?

- How to establish a multimodal conversation representation framework to efficiently align, interact and fuse the information from three modalities?

Acknowledgments

References

Laura Ana Maria Bostan, Evgeny Kim, and Roman Klinger. 2020. Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 1554–1566.

Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N Chang, Sungbok Lee, and Shrikanth S Narayanan. 2008. Iemocap: Interactive emotional dyadic motion capture database. *Language resources and evaluation*, 42(4):335–359.

Ankush Chatterjee, Umang Gupta, Manoj Kumar Chinnakotla, Radhakrishnan Srikanth, Michel Galley, and Puneet Agrawal. 2019. Understanding emotions in text using deep learning and big data. *Computers in Human Behavior*, 93:309–317.

Ying Chen, Sophia Yat Mei Lee, Shoushan Li, and Chu-Ren Huang. 2010. Emotion cause detection with linguistic constructions. In *Computational Linguistics (COLING)*, pages 179–187.

Xiyao Cheng, Ying Chen, Bixiao Cheng, Shoushan Li, and Guodong Zhou. 2017. An emotion cause corpus for chinese microblogs with multiple-user structures. *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, 17(1):1–19.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Zixiang Ding, Huihui He, Mengran Zhang, and Rui Xia. 2019. From independent prediction to re-ordered prediction: Integrating relative position and global label information to emotion cause identification. In *AAAI Conference on Artificial Intelligence (AAAI)*, pages 6343–6350.

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

Zixiang Ding, Rui Xia, and Jianfei Yu. 2020b. End-to-end emotion-cause pair extraction based on sliding window multi-label learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3574–3583.

P. Ekman. 1971. Universals and cultural differences in facial expressions of emotion. *Nebraska Symposium on Motivation. Nebraska Symposium on Motivation*, Vol. 19.

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 *Association for Computational Linguistics (ACL)*, pages 3707–3717.

Mauajama Firdaus, Hardik Chauhan, Asif Ekbal, and Pushpak Bhattacharyya. 2020. Meisd: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4441–4453.

Kai Gao, Hua Xu, and Jiushuo Wang. 2015a. Emotion cause detection for chinese micro-blogs based on ecocc model. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD)*, pages 3–14.

Kai Gao, Hua Xu, and Jiushuo Wang. 2015b. A rule-based approach to emotion cause detection for chinese micro-blogs. *Expert Systems with Applications*, 42(9):4517–4528.

Qinghong Gao, Jiannan Hu, Ruifeng Xu, Gui Lin, Yulan He, Qin Lu, and Kam-Fai Wong. 2017. Overview of ntcir-13 eca task. In *Proceedings of the NTCIR-13 Conference*.

Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. 2015. Detecting emotion stimuli in emotion-bearing sentences. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 152–165. Springer.

Lin Gui, Jiannan Hu, Yulan He, Ruifeng Xu, Qin Lu, and Jiachen Du. 2017. A question answering approach to emotion cause extraction. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1593–1602.

Lin Gui, Dongyin Wu, Ruifeng Xu, Qin Lu, Yu Zhou, et al. 2016a. Event-driven emotion cause extraction with corpus construction. In *EMNLP*, pages 1639–1649. World Scientific.

Lin Gui, Ruifeng Xu, Qin Lu, Dongyin Wu, and Yu Zhou. 2016b. Emotion cause extraction, a challenging task with corpus construction. In *Chinese National Conference on Social Media Processing*, pages 98–109.

Lin Gui, Li Yuan, Ruifeng Xu, Bin Liu, Qin Lu, and Yu Zhou. 2014. Emotion cause detection with linguistic construction in chinese weibo text. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 457–464. Springer.

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.

Chao-Chun Hsu, Sheng-Yeh Chen, Chuan-Chun Kuo, Ting-Hao Huang, and Lun-Wei Ku. 2018. Emotionlines: An emotion corpus of multi-party conversations. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Shuiwang Ji, Wei Xu, Ming Yang, and Kai Yu. 2012. 3d convolutional neural networks for human action recognition. *IEEE transactions on pattern analysis and machine intelligence*, 35(1):221–231.

Xiao Jin, Jianfei Yu, Zixiang Ding, Rui Xia, Xiangsheng Zhou, and Yaofeng Tu. 2020. Hierarchical multimodal transformer with localness and speaker aware attention for emotion recognition in conversations. In *CCF International Conference on Natural Language Processing and Chinese Computing*, pages 41–53. Springer.

Evgeny Kim and Roman Klinger. 2018. Who feels what and why? annotation of a literature corpus with semantic roles of emotions. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1345–1359.

J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174.

Sophia Yat Mei Lee, Ying Chen, and Chu-Ren Huang. 2010. A text-driven rule-based system for emotion cause detection. In *NAACL HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 45–53.

Weiyuan Li and Hua Xu. 2014. Text-based emotion classification using emotion cause extraction. *Expert Systems with Applications*, 41(4):1742–1749.

Xiangju Li, Kaisong Song, Shi Feng, Daling Wang, and Yifei Zhang. 2018. A co-attention neural network model for emotion cause analysis with emotional context awareness. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 4752–4757.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 986–995.

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282.

Gary McKeown, Michel Valstar, Roddy Cowie, Maja Pantic, and Marc Schroder. 2012. The semaine database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent. *IEEE Transactions on Affective Computing*, 1(3):5–17.

Alena Neviarouskaya and Masaki Aono. 2013. Extracting causes of emotions from text. In *International Joint Conference on Natural Language Processing (IJCNLP)*, pages 932–936.

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.

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. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 527–536.

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*, pages 1–16.

Irene Russo, Tommaso Caselli, Francesco Rubino, Ester Boldrini, and Patricio Martínez-Barco. 2011. Emocause: an easy-adaptable approach to emotion cause contexts. In *Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*, pages 153–160.

Shuangyong Song and Yao Meng. 2015. Detecting concept-level emotion cause in microblogging. In *World Wide Web (WWW)*, pages 119–120.

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

Bo Xu, Hongfei Lin, Yuan Lin, Yufeng Diao, Liang Yang, and Kan Xu. 2019. Extracting emotion causes using learning to rank methods from an information retrieval perspective. *IEEE Access*, 7:15573–15583.

Ruifeng Xu, Jiannan Hu, Qin Lu, Dongyin Wu, and Lin Gui. 2017. An ensemble approach for emotion cause detection with event extraction and multi-kernel svms. *Tsinghua Science and Technology*, 22(6):646–659.

Shuntaro Yada, Kazushi Ikeda, Keiichiro Hoashi, and Kyo Kageura. 2017. A bootstrap method for automatic rule acquisition on emotion cause extraction. In *IEEE International Conference on Data Mining Workshops*, pages 414–421.

Xinyi Yu, Wenge Rong, Zhuo Zhang, Yuanxin Ouyang, and Zhang Xiong. 2019. Multiple level hierarchical network-based clause selection for emotion cause extraction. *IEEE Access*, 7(1):9071–9079.