Multimodal Emotion-Cause Pair Extraction in Conversations

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

Abstract—Conversation is an important form of human communication and contains a large number of emotions. It is interesting to discover emotions and their causes in conversations. Conversation in its natural form is multimodal. Many studies have been carried out on multimodal emotion recognition in conversations, yet there is still a lack of work on multimodal emotion cause analysis. In this article, we introduce a new task named Multimodal Emotion-Cause Pair Extraction in Conversations, aiming to jointly extract emotions and the corresponding causes from conversations reflected in multiple modalities (i.e., text, audio and video). We accordingly construct a multimodal conversational emotion cause dataset, Emotion-Cause-in-Friends, which contains 9,794 multimodal emotion-cause pairs among 13,619 utterances in the *Friends* sitcom. We benchmark the task by establishing two baseline systems including a heuristic approach considering inherent patterns in the location of causes and emotions and a deep learning approach that incorporates multimodal features for emotion-cause pair extraction, and conduct the human performance test for comparison. Furthermore, we investigate the effect of multimodal information, explore the potential of incorporating commonsense knowledge, and perform the task under both Static and Real-time settings.

Index Terms—Affective computing, emotion analysis, emotion cause extraction, emotion-cause pair extraction, multimodal learning

1 Introduction

Emotions play an important role in human communication and decision-making [1]. In the field of textual emotion analysis, previous research mostly focused on emotion recognition [2], [3], [4], [5]. In recent years, emotion cause analysis, a new task which aims at extracting potential emotion causes, has received much attention due to its important application value. It contains two representative tasks: emotion cause extraction (ECE) and emotion-cause pair extraction (ECPE). ECE aims to extract the potential causes given the emotions [6], [7], [8]; ECPE was proposed to jointly extract the emotions and the corresponding causes in pairs [9], [10], [11], [12], thereby solving the problem of ECE's emotion annotation dependency and gaining more attention recently. These studies were normally carried out based on news articles, microblogs or fictions.

Conversation is an important form of human communication and contains a large number of emotions. Poria et al. [13] introduced the task to recognize emotion cause in textual conversations. 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 recognition in conversations [14], [15], [16], to the best of our knowledge, there is still a lack of research on multimodal emotion cause analysis in conversations at present.

In this work, we introduce a new task named Multimodal Emotion-Cause Pair Extraction in Conversations (MECPE), with the goal to extract all potential pairs of emotions and their corresponding causes from a conversation, in consideration of three modalities including 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 1,374 conversations and 13,619 utterances[1], where 9,794 emotion-cause pairs are annotated, covering three modalities. Fig. 1 displays a real example in our ECF dataset. In this conversation, it is expected to extract a set of six utterance-level emotion-cause pairs, e.g., Chandler's *Joy* emotion in Utterance 4 (U_4 for short) is triggered by the objective cause that he and Monica had made up and Monica's subjective opinion in U_3, forming the pairs (U_4, U_2) and (U_4, U_3); The cause for Phoebe's *Disgust* in U_5 is the objective event that Monica and Chandler were kissing in front of her (mainly reflected in the visual modality of U_5), forming the pair (U_5, U_5).

We benchmark the task with two baseline systems. One is a heuristic approach by using inherent patterns in the

- *The authors are with the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, Jiangsu 210094, China. E-mail: {ffwang, dingzixiang, rxia, zyli, jfyu}@njust.edu.cn.*

Manuscript received 31 January 2022; revised 27 November 2022; accepted 1 December 2022. Date of publication 5 December 2022; date of current version 13 September 2023.
This work was supported in part by the Natural Science Foundation of China under Grants 62076133 and 62006117, in part by the Natural Science Foundation of Jiangsu Province for Young Scholars under Grant BK20200463, and in part by Distinguished Young Scholars under Grant BK20200018.
(Corresponding author: Rui Xia.)
Recommended for acceptance by Special Issue On Neurosymbolic AI for Sentiment Analysis.
Digital Object Identifier no. 10.1109/TAFFC.2022.3226559

[1] An utterance is a unit of speech divided by the speaker's breath or pause.

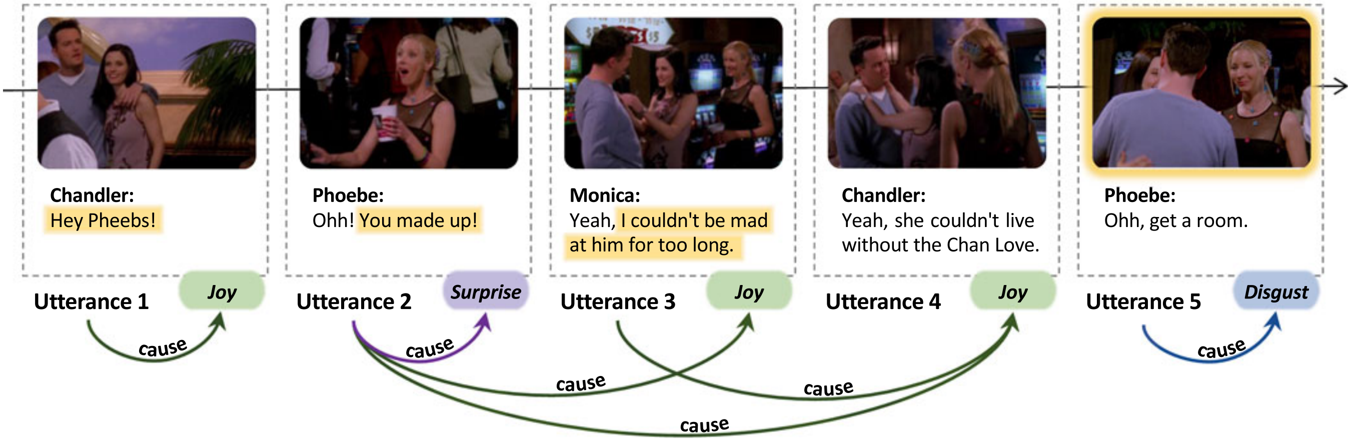


Fig. 1. An example of the annotated conversation in our ECF dataset. Each arc points from the cause utterance to the emotion it triggers. We have highlighted the cause evidence in the cause utterance. Background: Chandler and his girlfriend Monica walked into the casino, hugging each other (they had a quarrel earlier but made up soon), and then started a conversation with Phoebe.

location of causes and emotions; the other is a deep learning approach, called MECPE-2steps, adapted from a representative approach for ECPE in news articles [9]. We incorporate multimodal features for utterance representation, extract emotion utterances and cause utterances respectively, and finally construct the emotion-cause pairs.

We evaluate the performance of two simple baseline systems on the ECF dataset. To more comprehensively evaluate the MECPE task, we also conduct the human performance test and compare it with the evaluated baseline systems. The experimental results and further discussions demonstrate the potential of multimodal information for discovering emotions and causes in conversations.

The remainder of this paper is organized as follows: Section 2 gives the definition of the MECPE task. Section 3 describes the process of data annotation, as well as the statistics and analysis of our dataset. In Section 4, we introduce two baseline systems. The experimental results are presented and discussed in Section 5. Section 6 reviews current studies on the task of textual emotion cause analysis and multimodal emotion recognition in conversations. Finally, we conclude the paper and provide future directions of our research in Section 7.

2 Task

We first clarify the definitions of emotion and cause before introducing the task and dataset.

Emotion is a psychological state associated with thought, feeling and behavioral response [17]. 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* [18]. In conversations, emotions are usually annotated at the utterance level [2], [19], [20]. *Cause* refers to the explicitly expressed event or argument that is highly linked with the corresponding emotion [6], [7], [21]. 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.

Given a conversation $D = [U_1, \dots, U_i, \dots, U_{|D|}]$, in which each utterance is represented by the text, audio and video, i.e., $\boldsymbol{u}_i = [\boldsymbol{t}_i, \boldsymbol{a}_i, \boldsymbol{v}_i]$, we define the following two tasks:

Task 1: Multimodal Emotion-Cause Pair Extraction (MECPE). The goal of MECPE is to extract a set of emotion-cause pairs in consideration of three modalities

$$\mathcal{P} = \Big\{\dots, (U_j^e, U_k^c), \dots\Big\}, \tag{1}$$

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

Task 2: Multimodal Emotion-Cause Pair Extraction with Emotion Category (MECPE-Cat). In addition to MECPE, MECPE-Cat needs to further identify the corresponding emotion category for each emotion-cause pair

$$\mathcal{P} = \Big\{\dots, (U_j^e, U_k^c, y^e), \dots\Big\}, \tag{2}$$

where y^e denotes the emotion category in correspondence with (U_j^e, U_k^c).

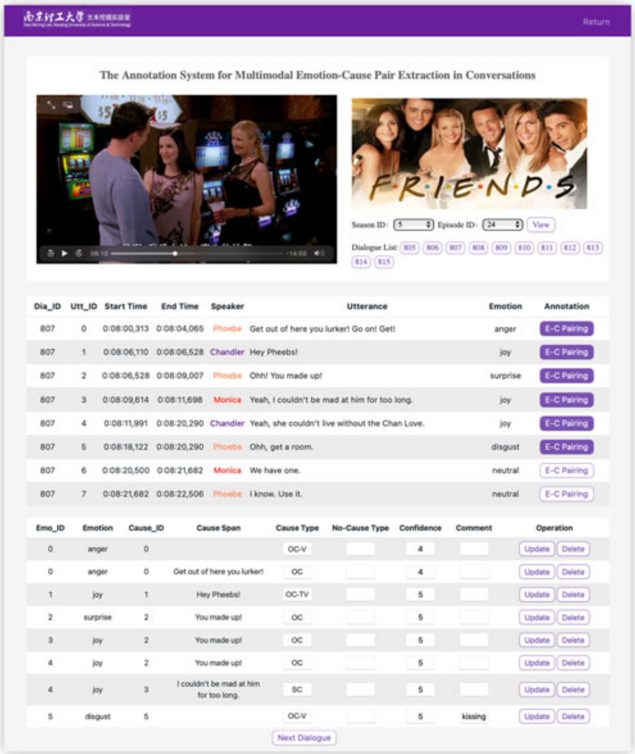
Taking the conversation in Fig. 1 as an example, the output of Task 1 is a set of utterance-level emotion-cause pairs: $\mathcal{P} = \{(U_1, U_1), (U_2, U_2), (U_3, U_2), (U_4, U_2), (U_4, U_3), (U_5, U_5)\}$. The emotion-cause pair in Task 2 includes an extra emotion category, e.g., (U_3, U_2, Joy).

It is worth noting that the above two tasks can be performed under two settings: the Static Setting and the Real-time Setting. The former can leverage the context of the entire conversation, while the latter relies only on historical utterances to detect the emotion as well as the cause [13].

3 Dataset

3.1 Dataset Source

The conversations in sitcoms usually contain more emotions than other TV series and movies. Hsu et al. [19] constructed the EmotionLines dataset for the Emotion Recognition in Conversations (ERC) task by selecting 1,000 conversations from the scripts of the popular American sitcom *Friends* and annotating each utterance with an emotion category. Poria et al. [20] extended EmotionLines to a multimodal dataset



The Annotation System for Multimodal Emotion-Cause Pair Extraction in Conversations

Season ID: 5 Episode ID: 24

Dialogue List:

Dia_ID	Utt_ID	Start Time	End Time	Speaker	Utterance	Emotion	Annotation
807	0	0:08:00,313	0:08:04,065	Phoebe	Get out of here you lurker! Go on! Get!	anger	E-C Pairing
807	1	0:08:06,110	0:08:06,528	Chandler	Hey Pheebs!	joy	E-C Pairing
807	2	0:08:06,528	0:08:09,007	Phoebe	Ohh! You made up!	surprise	E-C Pairing
807	3	0:08:09,614	0:08:11,698	Monica	Yeah, I couldn't be mad at him for too long.	joy	E-C Pairing
807	4	0:08:11,991	0:08:20,290	Chandler	Yeah, she couldn't live without the Chan Love.	joy	E-C Pairing
807	5	0:08:18,122	0:08:20,290	Phoebe	Ohh, get a room.	disgust	E-C Pairing
807	6	0:08:20,500	0:08:21,682	Monica	We have one.	neutral	E-C Pairing
807	7	0:08:21,682	0:08:22,506	Phoebe	I know. Use it.	neutral	E-C Pairing

Emo_ID	Emotion	Cause_ID	Cause Span	Cause Type	No-Cause Type	Confidence	Comment	Operation
0	anger	0		OC-V		4		Update Delete
0	anger	0	Get out of here you lurker!	OC		4		Update Delete
1	joy	1	Hey Pheebs!	OC-TV		5		Update Delete
2	surprise	2	You made up!	OC		5		Update Delete
3	joy	2	You made up!	OC		5		Update Delete
4	joy	2	You made up!	OC		5		Update Delete
4	joy	3	I couldn't be mad at him for too long.	SC		5		Update Delete
5	disgust	5		OC-V		5	kissing	Update Delete

Fig. 2. The interface of our developed annotation toolkit.

MELD, by extracting the audio-visual clip from the source episode, and re-annotating each utterance.

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.

3.2 Annotation Procedure

We first develop detailed annotation instructions and then employ three annotators who have reasonable knowledge of our task to annotate the entire dataset independently. Two of the three annotators are female and one is male.

Annotation Instructions. Given a multimodal conversation, the annotators first understand the context based on its raw transcripts and audio-visual clips. For the emotion category (one of Ekman's six basic emotions) labeled on each utterance, they should find the evidence of the corresponding causes from three modalities (text, audio and video) of the whole conversation. They are asked to not only annotate the utterance containing the cause evidence, but also label the type of the cause, and mark the textual cause span if the cause is explicitly expressed in the textual modality. If the annotator cannot identify the emotion cause from the limited conversation content, he may not make any annotations. For the example shown in Fig. 1, Phoebe was surprised that Chandler and Monica had made up and the evidence of the emotion cause exists in the text of U_2, so the cause utterance U_2, the textual cause span "*You made up*" and the type of cause "Objective Cause" will be annotated for Phoebe's *Surprise* emotion in U_2. In addition to the same objective cause, Chandler's *Joy* emotion in U_4 has another cause: the cause utterance U_3, the textual cause span "*I couldn't be mad at him for too long*" and the type of cause "Subjective Cause". Since Monica and Chandler were kissing in front of Phoebe in the video clip of U_5, which made Phoebe feel uncomfortable, we can annotate the cause utterance U_5, the type of cause "Objective Cause" for Phoebe's *Disgust* in U_5.

TABLE 1
The Inter-Annotator Agreement for Emotion Cause Annotations

Evaluation Granularity	Annotator Pair	Cohen's Kappa	Fleiss' Kappa	F1 Score
Utterance	A&B	0.6348	0.6044	0.6451
	A&C	0.6595		0.6716
	B&C	0.6483		0.6634
	Avg.	0.6475		0.6599
Span	A&B	0.5996	0.5621	0.6064
	A&C	0.6201		0.6277
	B&C	0.5930		0.6026
	Avg.	0.6046		0.6126

A, B, C represent the three annotators, respectively.

Annotation Toolkit. To improve the annotation efficiency, we furthermore develop a multimodal emotion cause annotation toolkit. It is a general toolkit for multimodal annotation in conversations, with the functions of multimodal signal alignment, quick emotion-cause pair selection, multiple user and task management, distributable deployment, etc. Fig. 2 displays the interface of this toolkit. The average duration of each conversation in our dataset is 31.6 seconds, and it takes an average of 8.6 minutes to annotate a conversation.

After annotation, we determine the cause utterances by majority voting (use the utterance agreed upon by at least two annotators as the gold annotation) and take the largest boundary (i.e., the union of the spans) as the gold annotation of textual cause span, similar as [22], [23]. If there are further disagreements, another expert is invited for the final decision.

3.3 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, Fleiss' Kappa, and F1 score. Cohen's Kappa is used to measure the consistency of any two annotators [24], while Fleiss' Kappa is used to measure the overall annotation consistency among three annotators [25]. We also report F1 score, as [26] has pointed out that F-measure is more suitable than Kappa coefficient to evaluate span annotations.

Specifically, we calculate the agreement score for each emotion respectively after converting the cause utterance/span annotations into binary labels at the utterance/token level, and then average them. The agreement scores are reported in Table 1. It can be seen that almost all the Kappa

[2] MELD is licensed under GNU General Public License v3.0 https://github.com/declare-lab/MELD.

TABLE 2
Basic Statistic of our ECF Dataset

Items	Number
Conversations	1,374
Utterances	13,619
Emotion (utterances)	7,690
Emotion (utterances) with causes	7,081
Emotion-cause (utterance) pairs	9,794
Emotion (utterances) with single cause	4,914
Emotion (utterances) with two causes	1,724
Emotion (utterances) with three causes	354

coefficients are higher than 0.6, which indicates a substantial agreement between the three annotators [27]. In addition, all metrics show relatively low agreement on cause span annotations, which is reasonable, given that annotating fine-grained spans is more difficult than just finding cause utterances.

3.4 Dataset Statistic and Analysis

3.4.1 Overall Statistics

As shown in Table 2, the ECF dataset contains 1,374 conversations and 13,619 utterances from three modalities, where 7,690 emotion utterances and 9,794 emotion-cause pairs have been annotated. In other words, about 55.73% of the utterances are labeled with one of the six basic emotions, and 91.34% of the emotions have the annotations of their corresponding causes in our dataset. Some of these emotions may be triggered by the causes involving multiple utterances, thus forming multiple pairs, resulting in a total number of pairs greater than 7,081. In addition, it should be noted that a small part of emotions (8.66%) are not annotated with the corresponding causes, because they may be triggered by latent causes which are not explicitly expressed in a limited multimodal conversation but rather need to be inferred through the understanding of the entire conversation or even the entire episode in *Friends*.

In Table 3, 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.

3.4.2 Emotion/Cause Distribution

For each emotion category, the proportion of emotion utterances annotated with causes is shown in Fig. 4. It can be seen that the distribution of emotion categories is unbalanced, and the proportion of emotion annotated with causes varies slightly with the emotion category. Among them, *Joy* covers the largest percentage among the six emotion categories (29.72%), and also has the higher proportion having cause annotations (93.03%). The percentage of *Surprise* annotated with causes is the highest (94.11%). In addition, *Disgust* and *Fear* account for a relatively low proportion among the six emotion categories, but *Disgust* has a much higher percentage having cause annotations (93.19%), whereas *Fear* has the lowest (78.24%), because the causes of *Fear* are often ambiguous and difficult to annotate in the less-informed conversations according to our observation.

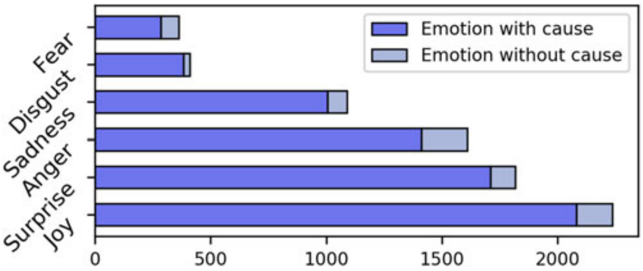


Fig. 4. The distribution of emotions (with/without cause) in different categories.

TABLE 3
A Summary of the Publicly Available Datasets for Traditional Emotion Cause Analysis, Emotion Recognition in Conversations and Conversational Emotion Cause Analysis

Dataset	Modality	Cause	Scene	# Ins
Emotion-Stimulus [28]	T	✓	–	2,414 s
ECE Corpus [22]	T	✓	News	2,105 d
NTCIR-13-ECA [29]	T	✓	Fiction	2,403 d
Weibo-Emotion [30]	T	✓	Blog	7,000 p
REMAN [31]	T	✓	Fiction	1,720 d
GoodNewsEveryone [23]	T	✓	News	5,000 s
IEMOCAP [32]	T,A,V	✗	Conv	7,433 u
DailyDialog [2]	T	✗	Conv	102,979 u
EmotionLines [19]	T	✗	Conv	14,503 u
SEMAINE [33]	T,A,V	✗	Conv	5,798 u
EmoContext [34]	T	✗	Conv	115,272 u
MELD [20]	T,A,V	✗	Conv	13,708 u
MELSD [35]	T,A,V	✗	Conv	20,000 u
RECCON-IE [13]	T	✓	Conv	665 u
RECCON-DD [13]	T	✓	Conv	11,104 u
ECF (ours)	T,A,V	✓	Conv	13,619 u

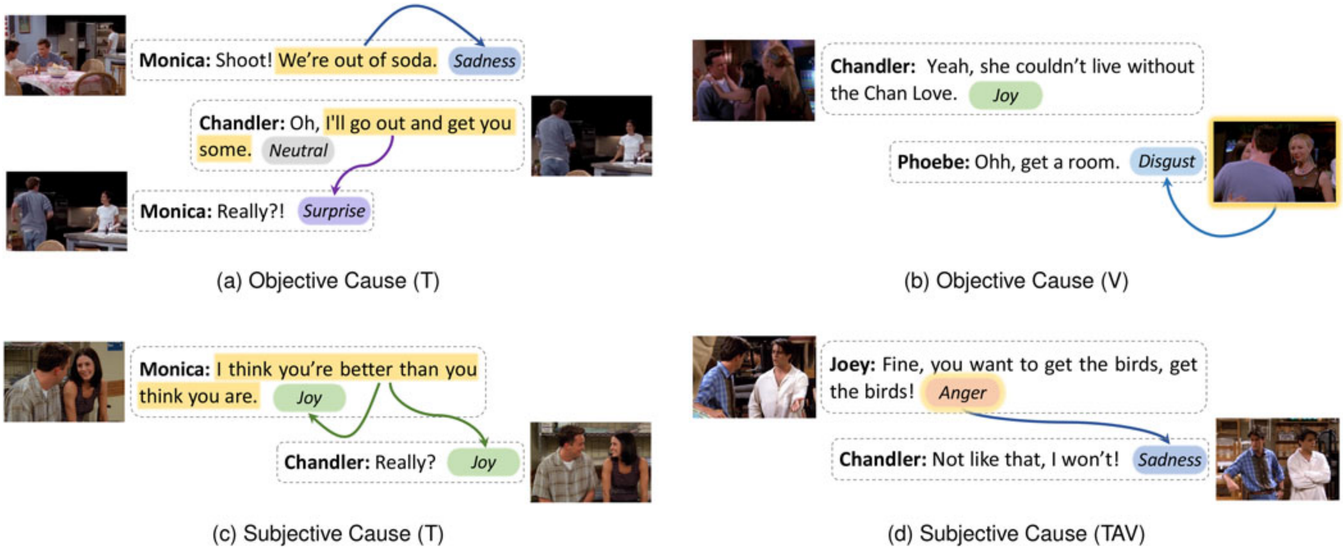
T, A and V refer to text, audio and video, respectively. Blog and Conv represent microblog and conversation. The units of the intances include sentence (s), document (d), post (p) and utterance (u).

3.4.3 Types of Emotion Causes

As stated in [36], [37], the reasons for human actions may be subjective or objective. Thus, we re-summarized the emotion causes into two types:

- *Objective Cause:* The publicly-observable things, events, facts, or conditions that have to do with the real outside world. As illustrated in Figs. 3a and 3b, the objective causes may be reflected in texts or visual scenes. For the example in Fig. 3b, what makes Phoebe disgust is that Monica and Chandler were kissing in front of her. It should be noted that there are also some objective reasons reflected in the audio, e.g., the speaker heard the baby crying.
- *Subjective Cause:* Personal feelings, opinions, experiences, or beliefs which cannot be proved right or wrong by any generally accepted criteria. In Fig. 3c, Monica's subjective opinion makes Chandler *Surprise*. Fig. 3d presents an example where the emotion of the target speaker is induced by the counterpart's emotion, which involves the three modalities.

In our dataset, the majority (71.23%) of emotions are triggered by objective causes, and a relatively small proportion (24.48%) of emotions are triggered by subjective causes. The


Monica: Shoot! We're out of soda. Sadness
Chandler: Oh, I'll go out and get you some. Neutral
Monica: Really?! Surprise
Chandler: Yeah, she couldn't live without the Chan Love. Joy
Phoebe: Ohh, get a room. Disgust
Monica: I think you're better than you think you are. Joy
Chandler: Really? Joy
Joey: Fine, you want to get the birds, get the birds! Anger
Chandler: Not like that, I won't! Sadness

(a) Objective Cause (T)

(b) Objective Cause (V)

(c) Subjective Cause (T)

(d) Subjective Cause (TAV)

Fig. 3. The examples of Objective Cause and Subjective Cause in our ECF dataset. Each arc points from the cause to the emotion it triggers. We have highlighted the cause evidence in the cause utterance. T, A, and V in parentheses refer to text, audio and video, respectively, indicating the modality that reflects the cause.

remaining emotions (4.29%) have both objective and subjective causes, such as those of U_4 in Fig. 1.

4 Baselines

We benchmark the MECPE task by establishing two preliminary baseline systems. One is a heuristic approach and the other is a deep learning approach.

4.1 Heuristic Approach

There are inherent patterns in the location of causes and emotions. For example, the cause often appears in a few utterances before the emotion (i.e, a pair of emotion and cause normally have a short relative distance). Based on this observation, we design four heuristic methods as follows.

We recognize the emotion utterances based on a trained emotion classifier (denoted by $\mathrm{E}_{\mathrm{Prediction}}$) at first, and then sample a cause utterance for the emotion according to two kinds of prior distributions of the relative positions (i.e., ..., -2, -1, 0, +1, ...) estimated on the training set:

- $\mathrm{C}_{\mathrm{Bernoulli}}$: We independently carry out a binary prediction for each relative position to determine whether its corresponding utterance is the cause utterance, based on a multi-variate Bernoulli distribution assumption.
- $\mathrm{C}_{\mathrm{Multinomial}}$: We sample one relative position from all possible positions, based on a Multinomial distribution assumption.

To test the upper bound of this approach, we also make use of the ground truth emotion annotations instead of emotion prediction in the first step, denoted by $\mathrm{E}_{\mathrm{Annotation}}$. A combination of two steps yields four heuristic methods: $\mathrm{E}_{\mathrm{Prediction}} + \mathrm{C}_{\mathrm{Bernoulli}}$, $\mathrm{E}_{\mathrm{Annotation}} + \mathrm{C}_{\mathrm{Bernoulli}}$, $\mathrm{E}_{\mathrm{Prediction}} + \mathrm{C}_{\mathrm{Multinomial}}$, and $\mathrm{E}_{\mathrm{Annotation}} + \mathrm{C}_{\mathrm{Multinomial}}$.

4.2 Deep Learning Approach

The deep learning approach MECPE-2steps shown in Fig. 5 is adapted from a representative approach for ECPE [9], where the scene shifts from news articles to conversation.

4.2.1 *Unimodal Feature Extraction*

We extract the features from three modalities and then concatenate them to obtain the independent multimodal representation of each utterance, i.e., $\boldsymbol{u}_i = [\boldsymbol{t}_i, \boldsymbol{a}_i, \boldsymbol{v}_i]$.

- *Text*: We initialize each token with pre-trained 300-dimensional GloVe vectors [38], feed them into a BiLSTM encoder with a standard attention mechanism, and then obtain the textual features of each utterance $\boldsymbol{t}_i$. In addition to BiLSTM, we also use pre-trained BERT [39] as the basic word encoder and feed each utterance into it independently.
- *Audio*: We extract the 6373-dimensional acoustic features $\boldsymbol{a}_i$ via the openSMILE toolkit [40] based on the INTERSPEECH 2009 Emotion Challenge feature set.
- *Video*: We apply one kind of 3D-CNN networks, named C3D [41], which learns spatio-temporal features using deep 3D ConvNet, to extract the 128-dimensional visual features $\boldsymbol{v}_i$ from the video of each utterance. Specifically, we sample 16 frames with a resolution of 171×128 from each video and feed them to the C3D network to extract a 4096-dimensional video descriptor, which is then followed by a linear layer for dimension reduction.

4.2.2 *Emotion and Cause Extraction via Multi-Task Learning*

In order to extract a set of emotion utterances and a set of cause utterances for each conversation, respectively, we feed the independent utterance representations $\boldsymbol{u}_i$ into two utterance-level encoders: one for emotion extraction and another for cause extraction, and each encoder is either based on BiLSTM or Transformer. The hidden states of two utterance-level encoders $\boldsymbol{r}_i^e$ and $\boldsymbol{r}_i^c$, which can be viewed as the emotion-specific representation and cause-specific representation of utterance U_i, are respectively fed into two softmax layers to detect

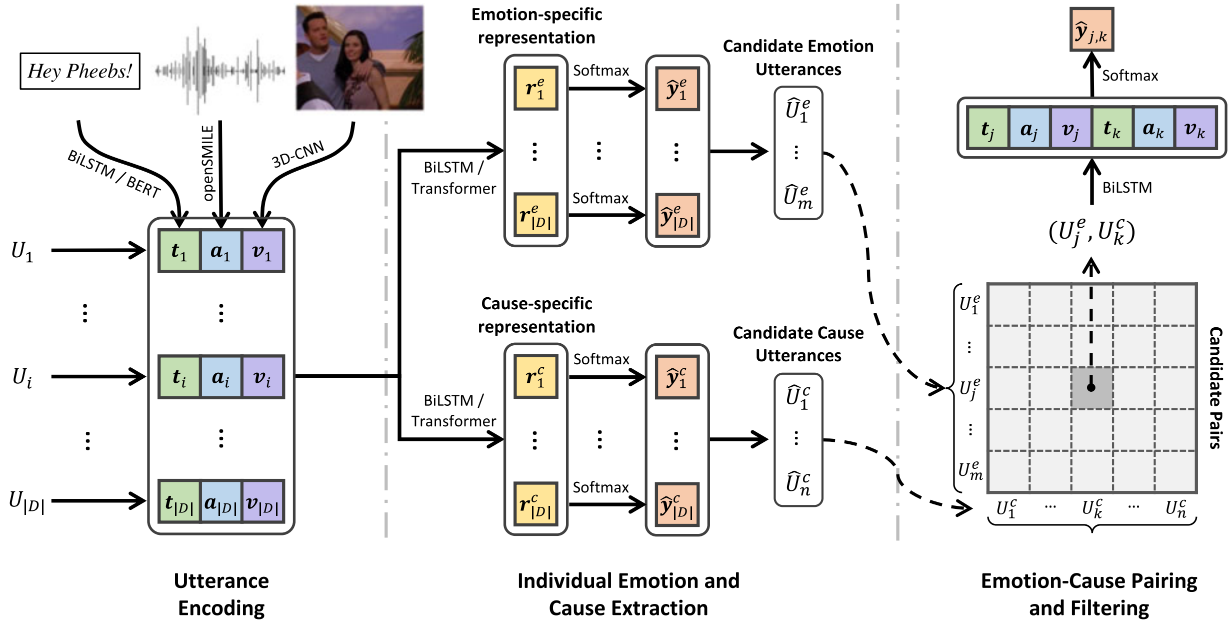


Fig. 5. The main structure of the deep learning baseline system MECPE-2steps.

whether U_i is an emotion/cause utterance

$$\hat{\mathbf{y}}_i^e = \mathrm{softmax}\left(\mathbf{W}^e\mathbf{r}_i^e + \mathbf{b}^e\right), \tag{3}$$
$$\hat{\mathbf{y}}_i^c = \mathrm{softmax}\left(\mathbf{W}^c\mathbf{r}_i^c + \mathbf{b}^c\right), \tag{4}$$

The loss of the first step is the sum of the cross-entropy loss of emotion extraction and cause extraction.

4.2.3 Emotion-Cause Pairing and Filtering

After obtaining a set of emotion utterances $E = [\hat{U}_1^e, \ldots, \hat{U}_m^e]$ and a set of cause utterances $C = [\hat{U}_1^c, \ldots, \hat{U}_n^c]$, we need to pair the two sets and extract a set of emotion-cause pairs with causal relationship.

First, we apply a Cartesian product to E and C, and obtain the set of candidate pairs. Second, we represent each pair by a feature vector $\mathbf{x}_{(\hat{U}_j^e, \hat{U}_k^c)}$, which concatenates the independent multimodal representations of the emotion utterance and the cause utterance as well as the distance vector between the two utterances. Finally, the pair representation is fed into a softmax layer to detect whether $(\hat{U}_j^e, \hat{U}_k^c)$ is a valid emotion-cause pair

$$\hat{\mathbf{y}}_{j,k} = \mathrm{softmax}\left(\mathbf{W}\mathbf{x}_{\left(\hat{U}_j^e, \hat{U}_k^c\right)} + \mathbf{b}\right). \tag{5}$$

5 Experiments

5.1 Experimental Setup

5.1.1 Experimental Settings

We evaluate the performance of two baseline systems on the ECF dataset. We divide the dataset into training, validation and testing sets in a ratio of 7:1:2 at the conversation level. 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. For BERT, we utilize the BERT-based model[3] with linear warmup and linear decay mechanism to the learning rate. The batch size and initial learning rate are set to 8 and 1e-5, respectively. We repeat all the experiments 20 times and report the average results.

5.1.2 Evaluation Metrics

Similar to [9], the Precision, Recall and F_1 score of the emotion-cause utterance pairs are used as the evaluation metrics for MECPE, which are calculated as follows:

$$P = \frac{\sum correct\_pairs}{\sum predicted\_pairs} \tag{6}$$
$$R = \frac{\sum correct\_pairs}{\sum annotated\_pairs} \tag{7}$$
$$F_1 = \frac{2 \times P \times R}{P + R} \tag{8}$$

where $predicted\_pairs$ denotes the number of emotion-cause pairs predicted by the model, $annotated\_pairs$ denotes the total number of emotion-cause pairs that are annotated in the dataset and the $correct\_pairs$ means the number of pairs that are both annotated and predicted as an emotion-cause pair.

In addition, we also evaluate the performance of emotion extraction and cause extraction. The Precision, Recall and F_1 score defined in [22] are used as the evaluation metrics.

[3] https://github.com/google-research/bert

TABLE 4
Experimental Results on MECPE (Task 1)

Methods		Emotion Extraction			Cause Extraction			Pair Extraction		
		P	R	F1	P	R	F1	P	R	F1
Heuristic	$\mathrm{E}_{\mathrm{Prediction}} + \mathrm{C}_{\mathrm{Bernoulli}}$	0.7362	0.7968	0.7648	0.5491	0.5028	0.5244	0.3699	0.2677	0.3101
	$\mathrm{E}_{\mathrm{Annotation}} + \mathrm{C}_{\mathrm{Bernoulli}}$	1.0000	1.0000	1.0000	0.6420	0.5485	0.5915	0.4958	0.3307	0.3967
	$\mathrm{E}_{\mathrm{Prediction}} + \mathrm{C}_{\mathrm{Multinomial}}$	0.7362	0.7968	0.7648	0.5488	0.5022	0.5239	0.3694	0.2671	0.3096
	$\mathrm{E}_{\mathrm{Annotation}} + \mathrm{C}_{\mathrm{Multinomial}}$	1.0000	1.0000	1.0000	0.6417	0.5483	0.5913	0.4951	0.3303	0.3963
Deep Learning	MECPE-2steps ($\mathrm{Text}_{\mathrm{LSTM}}$)	0.7377	0.8025	0.7679	0.6493	0.7528	0.6963	0.5538	0.4702	0.5070
	+ Audio	0.7411	0.8043	0.7706	0.6532	0.7685	0.7049	0.5474	0.4976	0.5194
	+ Video	0.7362	0.8113	**0.7713**	0.6486	0.7738	0.7049	0.5497	0.4885	0.5158
	+ Audio + Video	0.7474	0.7957	0.7700	0.6522	0.7776	**0.7084**	0.5483	0.5009	**0.5220**
	MECPE-2steps ($\mathrm{Text}_{\mathrm{BERT}}$)	0.7717	0.8136	0.7910	0.6747	0.7319	0.7013	0.5764	0.4872	0.5271
	+ Audio	0.7691	0.8168	**0.7917**	0.6725	0.7391	**0.7027**	0.5713	0.5034	**0.5348**
	+ Video	0.7710	0.8118	0.7903	0.6810	0.7251	0.7018	0.5777	0.4953	0.5321
	+ Audio + Video	0.7745	0.8110	0.7916	0.6842	0.7243	**0.7027**	0.5747	0.4981	0.5320
Human	Human #1 (Text)	0.7778	0.6985	0.7360	0.6247	0.6982	0.6594	0.5488	0.5024	0.5243
	+ Audio	0.7163	0.8324	0.7700	0.7070	0.7025	0.7048	0.5234	0.5630	0.5425
	+ Video	0.7248	0.8026	0.7617	0.6220	0.7980	0.6991	0.5088	0.5710	0.5381
	+ Audio + Video	0.7211	0.8525	**0.7813**	0.6262	0.8440	**0.7190**	0.5073	0.6344	**0.5635**
	Human #2 (Text)	0.7163	0.8051	0.7581	0.6117	0.7598	0.6761	0.5057	0.5721	0.5346
	+ Audio	0.7410	0.8178	0.7767	0.6449	0.7990	0.7137	0.5426	0.5823	0.5617
	+ Video	0.7266	0.8089	0.7655	0.6232	0.8165	0.7068	0.5388	0.5772	0.5557
	+ Audio + Video	0.7248	0.8760	**0.7933**	0.6252	0.8636	**0.7253**	0.5450	0.6047	**0.5724**

We also report the performance of emotion extraction and cause extraction in terms of recision, recall, and F_1 *score. All three subtasks below perform binary classification. The best results are in bold.*

For MECPE-Cat, we evaluate the emotion-cause pairs of each emotion category with F_1 score separately and further calculate a weighted average of F_1 scores across different emotion categories. Considering the imbalance of emotion categories described in Section 3.4, we also report the weighted average F_1 score of the four main emotion categories except *Disgust* and *Fear*.

5.1.3 Human Performance Test

We further carry out the human performance test on the MECPE task. Specifically, we employ two graduate students majoring in computer science as testers, to extract the emotion-cause utterance pairs in the testing set independently. Unlike the annotators of the ECF dataset, they have no formal training before testing.

To demonstrate whether the multimodal information is effective for humans to discover emotions and the corresponding causes in conversations, we designed a four-stage human testing strategy: providing the testers with the text-only information, text and audio information, text and silent video information, and all information from three modalities, respectively.

5.2 Experiments on MECPE (Task 1)

5.2.1 Main Results

In Table 4, we report the experimental results of our two baseline systems and the human performance test on Task 1.

First, we can see that the F_1 scores of human testers on emotion/cause/pair extraction are all improved significantly after being provided with audio or video, which shows that multimodal information is beneficial for humans to better understand the conversations and discover emotions and their corresponding causes.

Second, the heuristic methods perform very poorly, even when the ground truth annotations are used for emotion recognition. It suggests that MECPE is a complicated task and can not be well addressed by heuristics based only on the relative position distribution.

Finally, the deep learning baseline system, MECPE-2steps, performs much better in comparison with the heuristic methods. By adding acoustic and visual features, it can also gain significant improvements. When using BERT as the encoder, the performance of MECPE-2steps is very sound. It achieves a F_1 score of 0.5271 on pair extraction, which is comparable to human performance. We think it is reasonable because of BERT's strong representation ability and pre-training knowledge. Yet, the improvement by adding the multimodal features is slight in this case.

5.2.2 The Potential of Multimodal Information

In order to investigate the potential of multimodal information in our MECPE task, in Fig. 6 we compare the results of different methods before and after providing the audio and video information.

We can observe that introducing auditory and visual modalities leads to a significant increase in F_1 score of LSTM-based MECPE-2steps and two human testers. The improvements on pair extraction are 1.9 percentages under LSTM-based MECPE-2steps and nearly 4 percentages under the human performance test. Specifically, we find that introducing multimodal information will greatly increase the Recall score. As we have shown in the example in Figs. 1 and 3, a part of the emotion causes are expressed through

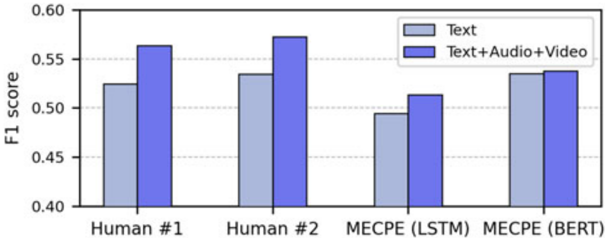


Fig. 6. Performance comparison of our baseline system and human testers on MECPE before and after providing the audio and video information. "MECPE (LSTM/BERT)" denotes our baseline system MECPE-2steps that uses Bi-LSTM/BERT as the textual encoder.

visual and auditory modalities. Hence, it makes sense to utilize multimodal information for our MECPE task.

As has been mentioned that when using BERT as the encoder, the improvement brought by the multimodal feature is limited. One possible reason is that BERT has already achieved high performance on the textual modality, which reduces the room for improvement brought by multimodal information. Another reason is as a baseline system, the representation and utilization of multimodal information in MECPE-2steps is very simple. It is not enough to perform deep scene understanding and reasoning required for multimodal emotion and cause extraction, leaving much room for further improvement in the future. For example, exploring visual commonsense reasoning [42], [43], [44], a new direction in computer vision, might be helpful for MECPE.

5.2.3 The Effect of Different Multimodal Fusion Methods

We report the performance of our baseline system using different multimodal fusion methods on the MECPE task in Table 5. *Concat*, *Add* and *LMF* fuse three unimodal feature vectors of each utterance before the utterance-level encoder by concatenation, addition, and the Low-rank Multimodal Fusion method, respectively. *Concat (Late)* first obtains the representations for each modality of the utterance separately via utterance-level encoders, and then concatenates them to make predictions (late fusion).

We can see that the performance of concatenation or addition is better than that of using LMF. The possible reason is that the introduction of the LMF model increases the complexity of the overall system and thus reduces its generalization performance [45], [46]. Moreover, the early fusion method outperforms the late fusion method because early fusion can well capture the inter-modal interaction. We will explore more advanced approaches for multimodal fusion in future work.

5.3 Experiments on MECPE-Cat (Task 2)

MECPE-Cat 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 MECPE-steps. The experimental results of pair extraction 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 on different emotion categories significantly varies with the proportion of emotion and cause annotation shown in Fig. 4. It should be noted that emotion category imbalance is actually an inherent problem in the ERC task [2], [19], [20], which is of great challenge and needs to be tackled in future work.

TABLE 5
Performance Comparison of Using Different Multimodal Fusion Methods on MECPE

Methods		Pair Extraction		
		P	R	F1
MECPE-2steps ($\text{Text}_{\text{LSTM}}$)		0.5538	0.4702	0.5070
Concat	+ Audio	0.5474	0.4976	0.5194
	+ Video	0.5497	0.4885	0.5158
	+ Audio + Video	0.5483	0.5009	**0.5220**
Add	+ Audio	0.5519	0.4984	0.5225
	+ Video	0.5471	0.4945	0.5180
	+ Audio + Video	0.5487	0.5046	**0.5243**
LMF	+ Audio	0.5400	0.4985	**0.5168**
	+ Video	0.5473	0.4871	0.5136
	+ Audio + Video	0.5325	0.5012	0.5144
Concat (Late)	+ Audio	0.5429	0.4843	0.5104
	+ Video	0.5437	0.4831	0.5102
	+ Audio + Video	0.5417	0.4865	**0.5112**

Similar to the conclusions drawn on MECPE, the performance of the baseline system is significantly improved after fusing the acoustic and visual features, which demonstrates that multimodal information is also helpful for MECPE-Cat. The relatively poor performance on this task indicates that it is more difficult to further predict the emotion categories based on MECPE.

5.4 The Potential of Commonsense Knowledge

We conduct additional experiments to explore the potential of incorporating the knowledge from a commonsense knowledge base (CSKB) ATOMIC_{20}^{20} [47] into our baseline system.

Specifically, we feed each utterance into a pre-trained neural knowledge model $\text{COMET-ATOMIC}_{20}^{20}$[4], which can automatically generate the emotion-oriented commonsense knowledge under two kinds of relation types, i.e., *xReact* and *oReact* (denoting how does the subject and the object in the utterance feel after the event). Given a utterance and a relation type, five tail phrases are generated as the commonsense knowledge for the utterance[5]. Next, we combine the commonsense knowledge under the two relation types, and feed them to a BiLSTM encoder to obtain the KB-based representation of each utterance, followed by concatenating the KB-based representation and the original utterance representation as the final utterance representation.

The experimental results on our two tasks are shown in Table 7. We can clearly observe that incorporating commonsense knowledge brings consistent improvements on emotion extraction, cause extraction, and pair extraction tasks.

[4] https://github.com/allenai/comet-atomic-2020

[5] For example, given the utterance (which is annotated with *Joy*) *"Yep, we sure showed those Hassidic jewellers a thing or two about softball."* and the relation type *xReact*, the model will generate the following phrases: {*good about themselves*, *proud of themselves*, *happy*, *proud*, *good*}.

TABLE 6
Experimental Results on MECPE-Cat (Task 2)

Methods		Pair Extraction							
		Anger	Disgust	Fear	Joy	Sadness	Surprise	w-avg. 6	w-avg. 4
Heuristic	$\mathrm{E}_{\mathrm{Prediction}}$ + $\mathrm{C}_{\mathrm{Bernoulli}}$	0.1669	0.0295	0.0227	0.2253	0.0994	0.2285	0.1732	0.1863
	$\mathrm{E}_{\mathrm{Annotation}}$ + $\mathrm{C}_{\mathrm{Bernoulli}}$	0.1755	0.0304	0.0241	0.2463	0.1080	0.2451	0.1864	0.2006
	$\mathrm{E}_{\mathrm{Prediction}}$ + $\mathrm{C}_{\mathrm{Multinomial}}$	0.1662	0.0290	0.0228	0.2250	0.0996	0.2289	0.1731	0.1861
	$\mathrm{E}_{\mathrm{Annotation}}$ + $\mathrm{C}_{\mathrm{Multinomial}}$	0.1759	0.0303	0.0241	0.2457	0.1074	0.2453	0.1863	0.2004
Deep Learning	MECPE-2steps ($\mathrm{Text}_{\mathrm{LSTM}}$)	0.2240	0.0408	0.0138	0.3375	0.1271	0.3964	0.2599	0.2803
	+ Audio	0.2493	0.0394	0.0223	0.3538	0.1622	0.4204	0.2821	0.3043
	+ Video	0.2245	0.0363	0.0150	0.3488	0.1351	0.4041	0.2659	0.2871
	+ Audio + Video	0.2534	0.0406	0.0233	0.3613	0.1746	0.4192	**0.2872**	**0.3098**
	MECPE-2steps ($\mathrm{Text}_{\mathrm{BERT}}$)	0.2439	0.0000	0.0071	0.3884	0.2160	0.4024	0.2932	0.3192
	+ Audio	0.2630	0.0046	0.0050	0.3962	0.2273	0.4135	**0.3047**	**0.3315**
	+ Video	0.2268	0.0019	0.0036	0.3855	0.2058	0.3999	0.2857	0.3111
	+ Audio + Video	0.2508	0.0009	0.0101	0.3849	0.2236	0.4064	0.2962	0.3224
Human	Human #1 (Text)	0.2455	0.1026	0.3333	0.3582	0.4674	0.4804	0.3567	0.3733
	+ Audio	0.3357	0.1000	0.2857	0.3231	0.4762	0.4538	0.3666	0.3874
	+ Video	0.5077	0.1053	0.2667	0.3194	0.4545	0.3273	0.3852	0.4084
	+ Audio + Video	0.4064	0.2051	0.2286	0.3700	0.5659	0.4595	**0.4166**	**0.4352**
	Human #2 (Text)	0.3899	0.2692	0.2778	0.4709	0.2851	0.3447	0.3766	0.3861
	+ Audio	0.3810	0.2857	0.1818	0.4286	0.5366	0.3077	0.3922	0.4011
	+ Video	0.3619	0.2692	0.2778	0.4261	0.3371	0.4381	0.3846	0.3964
	+ Audio + Video	0.5321	0.2222	0.2588	0.4179	0.3330	0.3921	**0.4135**	**0.4299**

We report the F_1 *score under each emotion category as well as their weighted average results. "w-avg. 4" denotes the weighted-average* F_1 *score of the four main emotions categories except* Disgust *and* Fear. *The best results are in bold.*

TABLE 7
Experimental Results on MECPE and MECPE-Cat When Incorporating the Knowledge From $\mathrm{ATOMIC}^{20}_{20}$ Into Our Baseline System

Methods	Emotion Extraction			Cause Extraction			Pair Extraction				
	P	R	F1	P	R	F1	P	R	F1	w-avg. 6	w-avg. 4
MECPE-2steps (T)	0.7377	0.8025	0.7679	0.6493	0.7528	0.6963	0.5538	0.4702	0.5070	0.2599	0.2803
+ xReact	0.7548	0.7973	0.7749	0.6683	0.7445	0.7032	0.5713	0.4637	0.5105	0.2731	0.2928
+ oReact	0.7559	0.7977	0.7757	0.6669	0.7430	0.7020	0.5685	0.4596	0.5072	0.2718	0.2927
+ xReact + oReact	0.7527	0.8015	**0.7758**	0.6716	0.7413	**0.7037**	0.5713	0.4642	**0.5110**	**0.2747**	**0.2953**
MECPE-2steps (M)	0.7474	0.7957	0.7700	0.6522	0.7776	0.7084	0.5483	0.5009	0.5220	0.2872	0.3098
+ xReact	0.7621	0.7995	**0.7800**	0.6603	0.7782	0.7134	0.5538	0.4981	0.5227	0.2942	0.3162
+ oReact	0.7551	0.8059	0.7791	0.6621	0.7749	0.7131	0.5508	0.5011	0.5233	0.2949	0.3168
+ xReact + oReact	0.7572	0.8034	0.7791	0.6622	0.7776	**0.7145**	0.5519	0.5021	**0.5247**	**0.3006**	**0.3229**

"MECPE-2steps (T)" and "MECPE-2steps (M)" represent the baseline system under the text-only and multimodal settings, respectively. xReact *and* oReact *are the relation types of the knowledge. "w-avg. 6" and "w-avg. 4" evaluate the overall performance on MECPE-Cat.*

Moreover, since the generated commonsense knowledge is closely related to the emotion of subjects and objects, it is not surprising that the improvement on MECPE-Cat is more significant than that on MECPE. All these observations demonstrate the usefulness of incorporating external knowledge from existing CSKB to solve our two tasks.

5.5 Experiments Under the Real-Time Setting

The above experiments are all conducted on static conversations. In this subsection, we further perform our two tasks under the Real-time Setting mentioned in Section 2 and report the experimental results in Table 8.

Specifically, for the heuristic approach, we sample a cause utterance from the utterances before the emotion utterances according to prior distributions of the relative positions (i.e., ..., -2, -1, 0) estimated on the training set; for the deep learning approach, we encode only the historical context, i.e., replace the utterance-level BiLSTM to LSTM or mask the attention weights between the current utterance and future utterances in Transformer, and discard the candidate emotion-cause pairs where the cause locates after the emotion.

Compared with Table 4, the heuristic methods and BERT-based MECPE-2steps achieve better performance. The improvements are mainly in the Recall score, possibly because limiting the position of causes simplifies the task and makes it easier to find the correct emotion cause. However, the performance of the LSTM-based MECPE-2steps under the Real-time Setting deteriorates, especially on the MECPE-Cat task. We speculate that it is more sensitive to the lack of future information, which leads to a decrease in the F_1 scores of cause extraction and thus offsets the benefits of location restriction, while the stronger BERT-based system is basically not influenced.

TABLE 8
Experimental Results on MECPE and MECPE-Cat Under the Real-Time Setting

Methods		**Emotion Extraction**			**Cause Extraction**			**Pair Extraction**				
		P	**R**	**F1**	**P**	**R**	**F1**	**P**	**R**	**F1**	**w-avg. 6**	**w-avg. 4**
Heuristic	$E_{Prediction} + C_{Bernoulli}$	0.7415	0.7969	0.7676	0.5494	0.4922	0.5189	0.3824	0.2774	0.3213	0.1763	0.1904
	$E_{Annotation} + C_{Bernoulli}$	1.0000	1.0000	1.0000	0.6464	0.5408	0.5888	0.5122	0.3445	0.4119	0.1903	0.2056
	$E_{Prediction} + C_{Multinomial}$	0.7415	0.7969	0.7676	0.5491	0.5073	0.5270	0.3823	0.2880	0.3282	0.1805	0.1949
	$E_{Annotation} + C_{Multinomial}$	1.0000	1.0000	1.0000	0.6458	0.5573	0.5983	0.5121	0.3575	0.4210	0.1953	0.2110
Deep Learning	MECPE-2steps ($Text_{LSTM}$)	0.7394	0.8020	0.7688	0.6272	0.7374	0.6770	0.5622	0.4610	0.5054	0.2422	0.2630
	+ Audio	0.7406	0.8005	0.7688	0.6267	0.7469	0.6807	0.5517	0.4863	0.5161	0.2635	0.2858
	+ Video	0.7323	0.8088	0.7680	0.6252	0.7487	0.6804	0.5481	0.4868	0.5144	0.2492	0.2709
	+ Audio + Video	0.7340	0.8113	**0.7701**	0.6268	0.7488	**0.6818**	0.5460	0.4949	**0.5176**	**0.2721**	**0.2950**
	MECPE-2steps ($Text_{BERT}$)	0.7730	0.8115	0.7914	0.6746	0.7298	0.6996	0.5814	0.5049	0.5401	0.2931	0.3200
	+ Audio	0.7726	0.8134	**0.7919**	0.6659	0.7362	0.6966	0.5708	0.5219	0.5448	0.3002	0.3274
	+ Video	0.7706	0.8131	0.7906	0.6822	0.7222	**0.7012**	0.5774	0.5113	0.5416	0.2944	0.3210
	+ Audio + Video	0.7727	0.8085	0.7892	0.6793	0.7216	0.6987	0.5727	0.5216	**0.5451**	**0.3062**	**0.3343**

"w-avg. 6" and "w-avg. 4" evaluate the overall performance on MECPE-Cat.

6 Related Work

Previous research on emotion cause analysis focused on textual modality. Conversational emotion analysis, although involves multiple modalities, primarily studies emotion recognition. Therefore, we review current studies on the task of textual emotion cause analysis and multimodal emotion recognition in conversations as follows.

6.1 Textual Emotion Cause Analysis

The textual emotion cause extraction (ECE) task was originally proposed by [6], with the goal to extract cause spans of a given emotion in the text. Under the same task setting, one line of work used rule-based methods, including [48], [49] which automatically extracted conjunctive phrases related to emotion causes for cause identification from formal texts, and [50], [51] which detected the cause events in microblog posts via the proposed rule-based algorithms. Another line of work used machine learning methods, such as [28], [52].

By analyzing the corpus proposed by [6], Chen et al. [7] pointed out that clause may be a more suitable unit for cause annotation, and proposed to extract emotion cause at clause granularity. After that, some work based on this task setting sprung up [21], [53]. Especially, Gui et al. [22] released an open Chinese emotion cause dataset. This dataset has received extensive attention and become a benchmark dataset for the ECE task. Based on this corpus, in addition to many traditional machine learning methods [8], [22], [54] , a number of deep learning methods were put forward. For example, Gui et al. [55] and Li et al. [56] used deep memory networks and co-attention neural network to model context information, respectively; Yu et al. [57] proposed a hierarchical network-based clause selection framework; Ding et al. [58] introduced a relative position augmented embedding learning algorithm and transformed the independent prediction problem to a reordered prediction problem; etc.[59], [60].

However, there are two shortcomings in ECE: 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 [9] proposed a new task called emotion-cause pair extraction (ECPE) and constructed the ECPE dataset based on the benchmark corpus for ECE [22]. They further proposed a two-step pipeline framework ECPE-2steps, which first extracts an individual emotion set and cause set, and then pairs the corresponding emotions and causes. Following their work, many studies have been conducted on ECPE to solve the shortcomings of the existing methodology [10], [11], [12], [61].

The above studies mostly focused on emotion cause analysis in news articles [22], [23], [29], microblogs [30] and fictions [29], [31]. Recently, Poria et al. [13] introduced an interesting task of recognizing emotion cause in conversations and constructed a new dataset RECCON for this task. Considering that conversation itself is multimodal, we further propose to jointly extract emotions and their corresponding causes from conversations based on three modalities, and accordingly create a multimodal conversational emotion cause dataset.

6.2 Multimodal Emotion Recognition in Conversations

Although there is a lack of research on multimodal emotion cause analysis, many studies have been carried out on multimodal emotion recognition using textual, auditory, and visual modalities, especially in conversations, including convolutional neural network-based methods [62], [63], recurrent neural network-based methods [14], [64], graph structure-based methods [16], [65], and Transformer-based methods [15], [66].

In recent years, due to the increasing amount of open conversation data, the ERC task has received continuous attention in the field of affective computing. So far, there have been some publicly available datasets for ERC. IEMOCAP [32] contains multimodal dyadic conversations of ten actors performing the emotional scripts. SEMAINE [33] contains multimodal data of robot-human conversations, but only provides the attributes of four emotion dimensions. The above two datasets are relatively small in scale. DailyDialog [2] is a large dataset that contains the texts of two-person daily conversations covering 10 topics, but the neutral utterances in it account for a high proportion. EmoContext [34] has a large total number of utterances, but only contains two-person conversations in plain text, with only three utterances in each conversation. EmotionLines [19] 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. [20] 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. [35] 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.

7 Conclusion and Future Work

In this work, we introduce a new task named Multimodal Emotion-Cause Pair Extraction (MECPE), to conduct emotion cause analysis in conversations, and accordingly construct a multimodal conversational emotion cause dataset, Emotion-Cause-in-Friends (ECF). We benchmark the task by establishing two preliminary baseline systems, and conduct the human performance test for comparison. We also investigate the effect of multimodal information for this task, explore the potential of incorporating commonsense knowledge, and perform the task under both Static and Real-time settings.

As an important direction of affective computing, multimodal emotion cause analysis in conversation plays an important role in many real-world applications, such as intelligent customer service, intelligent companionship, and mental health monitoring. We think the new task and dataset may potentially bring new insights and perspectives to existing research in this direction.

It is also worth noting that MECPE is a challenging task. This work is only a preliminary study for the task. The proposed two baseline systems leave much room for further improvement. Although the multimodal features are shown to be effective for the MECPE task, the increase is quite limited. In summary, we think the following challenges are worthy of further study for the MECPE task:

- How to establish a multimodal conversation representation framework to efficiently align, interact and fuse the information from three modalities?
- How to effectively model the speaker relevance for both emotion recognition and cause extraction in conversations?
- How to perceive, understand and utilize the visual scenes to better assist emotion cause reasoning in conversations?
- How to utilize the external commonsense knowledge to bridge the gap between emotion and cause that are not explicitly reflected in the conversation?

Acknowledgments

Fanfan Wang and Zixiang Ding contributed equally to this work.

References

[1] T. Brosch, K. R. Scherer, D. M. Grandjean, and D. Sander, "The impact of emotion on perception, attention, memory, and decision-making," *Swiss Med. Weekly*, vol. 143, 2013, Art. no. w13786.

[2] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," in *Proc. 8th Int. Joint Conf. Natural Lang. Process.*, 2017, pp. 986–995.

[3] A. Chatterjee, K. N. Narahari, M. Joshi, and P. Agrawal, "Semeval-2019 task 3: Emocontext contextual emotion detection in text," in *Proc. 13th Int. Workshop Semantic Eval.*, 2019, pp. 39–48.

[4] J. Deng and F. Ren, "A survey of textual emotion recognition and its challenges," *IEEE Trans. Affective Comput.*, early access, Jan. 21 2021, doi: 10.1109/TAFFC.2021.3053275.

[5] A. Saxena, A. Khanna, and D. Gupta, "Emotion recognition and detection methods: A comprehensive survey," *J. Artif. Intell. Syst.*, vol. 2, no. 1, pp. 53–79, 2020.

[6] S. Y. M. Lee, Y. Chen, and C.-R. Huang, "A text-driven rule-based system for emotion cause detection," in *Proc. Workshop Comput. Approaches Anal. Gener. Emotion Text*, 2010, pp. 45–53.

[7] Y. Chen, S. Y. M. Lee, S. Li, and C.-R. Huang, "Emotion cause detection with linguistic constructions," in *Proc. CCF Int. Conf. Natural Lang. Process. Chin. Comput.*, 2010, pp. 179–187.

[8] L. Gui, R. Xu, Q. Lu, D. Wu, and Y. Zhou, "Emotion cause extraction, a challenging task with corpus construction," in *Proc. Chin. Nat. Conf. Soc. Media Process.*, 2016, pp. 98–109.

[9] R. Xia and Z. Ding, "Emotion-cause pair extraction: A new task to emotion analysis in texts," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 1003–1012.

[10] Z. Ding, R. Xia, and J. Yu, "ECPE-2D: Emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction," in *Proc. Assoc. Comput. Linguistics*, 2020, pp. 3161–3170.

[11] P. Wei, J. Zhao, and W. Mao, "Effective inter-clause modeling for end-to-end emotion-cause pair extraction," in *Proc. Assoc. Comput. Linguistics*, 2020, pp. 3171–3181.

[12] C. Fan, C. Yuan, J. Du, L. Gui, M. Yang, and R. Xu, "Transition-based directed graph construction for emotion-cause pair extraction," in *Proc. Assoc. Comput. Linguistics*, 2020, pp. 3707–3717.

[13] S. Poria et al., "Recognizing emotion cause in conversations," *Cogn. Comput.*, vol. 13, pp. 1–16, 2021.

[14] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "ICON: Interactive conversational memory network for multimodal emotion detection," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2018, pp. 2594–2604.

[15] X. Jin, J. Yu, Z. Ding, R. Xia, X. Zhou, and Y. Tu, "Hierarchical multimodal transformer with localness and speaker aware attention for emotion recognition in conversations," in *Proc. CCF Int. Conf. Natural Lang. Process. Chin. Comput.*, 2020, pp. 41–53.

[16] J. Hu, Y. Liu, J. Zhao, and Q. Jin, "MMGCN: Multimodal fusion via deep graph convolution network for emotion recognition in conversation," in *Proc. 59th Annu. Meeting Assoc. Comput. Linguistics 11th Int. Joint Conf. Natural Lang. Process.*, 2021, pp. 5666–5675.

[17] P. E. Ekman and R. J. Davidson, *The Nature of Emotion: Fundamental Questions.*, London, U.K.: Oxford Univ. Press, 1994.

[18] P. Ekman, "Universals and cultural differences in facial expressions of emotion," in *Proc. Nebraska Symp. Motivation*, 1971, pp. 207–283.

[19] C.-C. Hsu, S.-Y. Chen, C.-C. Kuo, T.-H. Huang, and L.-W. Ku, "EmotionLines: An emotion corpus of multi-party conversations," in *Proc. 11th Int. Conf. Lang. Resour. Eval.*, 2018, pp. 1597–1601.

[20] S. Poria, D. Hazarika, N. Majumder, G. Naik, E. Cambria, and R. Mihalcea, "MELD: A multimodal multi-party dataset for emotion recognition in conversations," in *Proc. 57th Annu. Meeting Assoc. Comput. Linguistics*, 2019, pp. 527–536.

[21] I. Russo, T. Caselli, F. Rubino, E. Boldrini, and P. Mart ínez-Barco, "Emocause: An easy-adaptable approach to emotion cause contexts," in *Proc. Workshop Comput. Approaches Subjectivity Sentiment Anal.*, 2011, pp. 153–160.

[22] L. Gui et al., "Event-driven emotion cause extraction with corpus construction," in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2016, pp. 1639–1649.

[23] L. A. M. Bostan, E. Kim, and R. Klinger, "Goodnewseveryone: A corpus of news headlines annotated with emotions, semantic roles, and reader perception," in *Proc. 12th Lang. Resour. Eval. Conf.*, 2020, pp. 1554–1566.

[24] J. Cohen, “A coefficient of agreement for nominal scales,” *Educ. Psychol. Meas.*, vol. 20, no. 1, pp. 37–46, 1960.

[25] M. L. McHugh, “Interrater reliability: The kappa statistic,” *Biochemia Medica*, vol. 22, no. 3, pp. 276–282, 2012.

[26] G. Hripcsak and A. S. Rothschild, “Agreement, the f-measure, and reliability in information retrieval,” *J. Amer. Med. Inform. Assoc.*, vol. 12, no. 3, pp. 296–298, 2005.

[27] J. R. Landis and G. G. Koch, “The measurement of observer agreement for categorical data,” *Biometrics*, vol. 33, pp. 159–174, 1977.

[28] D. Ghazi, D. Inkpen, and S. Szpakowicz, “Detecting emotion stimuli in emotion-bearing sentences,” in *Proc. Int. Conf. Intell. Text Process. Comput. Linguistics*, 2015, pp. 152–165.

[29] Q. Gao et al., “Overview of NTCIR-13 ECA task,” in *Proc. 13th NTCIR Conf.*, 2017, pp. 361–366.

[30] X. Cheng, Y. Chen, B. Cheng, S. Li, and G. Zhou, “An emotion cause corpus for chinese microblogs with multiple-user structures,” *ACM Trans. Asian Low-Resource Lang. Inf. Process.*, vol. 17, no. 1, pp. 1–19, 2017.

[31] E. Kim and R. Klinger, “Who feels what and why? annotation of a literature corpus with semantic roles of emotions,” in *Proc. 27th Int. Conf. Comput. Linguistics*, 2018, pp. 1345–1359.

[32] C. Busso et al., “IEMOCAP: Interactive emotional dyadic motion capture database,” *Lang. Resour. Eval.*, vol. 42, no. 4, pp. 335–359, 2008.

[33] G. McKeown, M. Valstar, R. Cowie, M. Pantic, and M. Schroder, “The SEMAINE database: Annotated multimodal records of emotionally colored conversations between a person and a limited agent,” *IEEE Trans. Affect. Comput.*, vol. 1, no. 3, pp. 5–17, First Quarter 2012.

[34] A. Chatterjee, U. Gupta, M. K. Chinnakotla, R. Srikanth, M. Galley, and P. Agrawal, “Understanding emotions in text using deep learning and big data,” *Comput. Hum. Behav.*, vol. 93, pp. 309–317, 2019.

[35] M. Firdaus, H. Chauhan, A. Ekbal, and P. Bhattacharyya, “MEISD: A multimodal multi-label emotion, intensity and sentiment dialogue dataset for emotion recognition and sentiment analysis in conversations,” in *Proc. 28th Int. Conf. Comput. Linguistics*, 2020, pp. 4441–4453.

[36] A. Sepielli, “Subjective and objective reasons,” in *The Oxford Handbook of Reasons and Normativity*. London, U.K.: Oxford Univ. Press, 2018.

[37] M. Schroeder, “Having reasons,” *Philos. Stud.*, vol. 139, no. 1, pp. 57–71, 2008.

[38] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2014, pp. 1532–1543.

[39] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. Conf. North Amer. Chapter Assoc. Comput. Linguistics: Hum. Lang. Technol.*, 2019, pp. 4171–4186.

[40] F. Eyben, M. Wöllmer, and B. Schuller, “Opensmile: The munich versatile and fast open-source audio feature extractor,” in *Proc. 18th ACM Int. Conf. Multimedia*, 2010, pp. 1459–1462.

[41] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, “Learning spatiotemporal features with 3D convolutional networks,” in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 4489–4497.

[42] R. Zellers, Y. Bisk, A. Farhadi, and Y. Choi, “From recognition to cognition: Visual commonsense reasoning,” in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, 2019, pp. 6720–6731.

[43] J. Lei, L. Yu, M. Bansal, and T. Berg, “TVQA: Localized, compositional video question answering,” in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2018, pp. 1369–1379.

[44] J. Lei, L. Yu, T. Berg, and M. Bansal, “Tvqa+: Spatio-temporal grounding for video question answering,” in *Proc. 58th Annu. Meeting Assoc. Comput. Linguistics*, 2020, pp. 8211–8225.

[45] P. Liang et al., “Multibench: Multiscale benchmarks for multimodal representation learning,” in *In Proc. Neural Inf. Process. Syst. Conf.*, 2021, pp. 1–20.

[46] D. Gkoumas, *Quantum Cognitively Motivated Context-Aware Multimodal Representation Learning for Human Language Analysis*. Milton Keynes, U.K.: Open Univ., 2021.

[47] J. D. Hwang et al., “(Comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs,” in *Proc. AAAI Conf. Artif. Intell.*, 2021, pp. 6384–6392.

[48] A. Neviarouskaya and M. Aono, “Extracting causes of emotions from text,” in *Proc. Int. Joint Conf. Natural Lang. Process.*, 2013, pp. 932–936.

[49] S. Yada, K. Ikeda, K. Hoashi, and K. Kageura, “A bootstrap method for automatic rule acquisition on emotion cause extraction,” in *Proc. IEEE Int. Conf. Data Mining Workshops*, 2017, pp. 414–421.

[50] W. Li and H. Xu, “Text-based emotion classification using emotion cause extraction,” *Expert Syst. Appl.*, vol. 41, no. 4, pp. 1742–1749, 2014.

[51] K. Gao, H. Xu, and J. Wang, “Emotion cause detection for chinese micro-blogs based on ECOCC model,” in *Proc. Pacific-Asia Conf. Knowl. Discov. Data Mining*, 2015, pp. 3–14.

[52] S. Song and Y. Meng, “Detecting concept-level emotion cause in microblogging,” in *Proc. World Wide Web*, 2015, pp. 119–120.

[53] L. Gui, L. Yuan, R. Xu, B. Liu, Q. Lu, and Y. Zhou, “Emotion cause detection with linguistic construction in chinese weibo text,” in *Proc. CCF Int. Conf. Natural Lang. Process. Chin. Comput.*, 2014, pp. 457–464.

[54] R. Xu, J. Hu, Q. Lu, D. Wu, and L. Gui, “An ensemble approach for emotion cause detection with event extraction and multi-kernel svms,” *Tsinghua Sci. Technol.*, vol. 22, no. 6, pp. 646–659, 2017.

[55] L. Gui, J. Hu, Y. He, R. Xu, Q. Lu, and J. Du, “A question answering approach to emotion cause extraction,” in *Proc. Empir. Methods Natural Lang. Process.*, 2017, pp. 1593–1602.

[56] X. Li, K. Song, S. Feng, D. Wang, and Y. Zhang, “A co-attention neural network model for emotion cause analysis with emotional context awareness,” in *Proc. Empir. Methods Natural Lang. Process.*, 2018, pp. 4752–4757.

[57] X. Yu, W. Rong, Z. Zhang, Y. Ouyang, and Z. Xiong, “Multiple level hierarchical network-based clause selection for emotion cause extraction,” *IEEE Access*, vol. 7, pp. 9071–9079, 2019.

[58] Z. Ding, H. He, M. Zhang, and R. Xia, “From independent prediction to re-ordered prediction: Integrating relative position and global label information to emotion cause identification,” in *Proc. AAAI Conf. Artif. Intell.*, 2019, pp. 6343–6350.

[59] B. Xu, H. Lin, Y. Lin, Y. Diao, L. Yang, and K. Xu, “Extracting emotion causes using learning to rank methods from an information retrieval perspective,” *IEEE Access*, vol. 7, pp. 15573–15583, 2019.

[60] R. Xia, M. Zhang, and Z. Ding, “RTHN: A RNN-transformer hierarchical network for emotion cause extraction,” in *Proc. Int. Joint Conf. Artif. Intell.*, 2019, pp. 5285–5291.

[61] Z. Ding, R. Xia, and J. Yu, “End-to-end emotion-cause pair extraction based on sliding window multi-label learning,” in *Proc. Conf. Empir. Methods Natural Lang. Process.*, 2020, pp. 3574–3583.

[62] S. Poria, I. Chaturvedi, E. Cambria, and A. Hussain, “Convolutional MKL based multimodal emotion recognition and sentiment analysis,” in *Proc. IEEE 16th Int. Conf. Data Mining*, 2016, pp. 439–448.

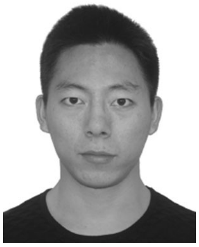
[63] I. Chaturvedi, R. Satapathy, S. Cavallari, and E. Cambria, “Fuzzy commonsense reasoning for multimodal sentiment analysis,” *Pattern Recognit. Lett.*, vol. 125, pp. 264–270, 2019.

[64] F. Chen, Z. Sun, D. Ouyang, X. Liu, and J. Shao, “Learning what and when to drop: Adaptive multimodal and contextual dynamics for emotion recognition in conversation,” in *Proc. 29th ACM Int. Conf. Multimedia*, 2021, pp. 1064–1073.

[65] Y. Fu et al., “Context-and knowledge-aware graph convolutional network for multimodal emotion recognition,” *IEEE MultiMedia*, vol. 29, no. 3, pp. 91–100, Third Quarter 2022.

[66] Y. Mao, G. Liu, X. Wang, W. Gao, and X. Li, “DialogueTRM: Exploring multi-modal emotional dynamics in a conversation,” in *Proc. Findings Assoc. Comput. Linguistics*, 2021, pp. 2694–2704.

Fanfan Wang received the BS degree in automation from the Nanjing University of Science and Technology, Nanjing, China, in 2019. She is currently working toward the PhD degree in computer science and technology with the School of Computer Science and Engineering, Nanjing University of Science and Technology. Her research interests include natural language processing and affective computing.



Zixiang Ding received the BS degree from the School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing, China, in 2016. He is currently working toward the PhD degree with the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests include natural language processing and affective computing. He has received the ACL2019 Outstanding Paper Award.

Rui Xia received the PhD degree from the Institute of Automation, Chinese Academy of Sciences, in 2011. He is currently a professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. His research interests include natural language processing, data mining, and affective computing. He has published more than 60 papers in top journals and conferences. His work on emotion-cause pair extraction has received the ACL2019 Outstanding Paper Award.

Zhaoyu Li received the BS degree in network engineering from the Dalian University of Technology, Dalian, China, in 2016, and the MS degree in computer science from the Nanjing University of Science and Technology, Nanjing, China, in 2022. His research interests include natural language processing, and multimodal learning.

Jianfei Yu received the BS and MS degrees from the Nanjing University of Science and Technology, Nanjing, China, in 2012 and 2015, respectively, and the PhD degree from Singapore Management University, Singapore, in 2018. He is currently an Associate Professor with the School of Computer Science and Engineering, Nanjing University of Science and Technology. His research interests include natural language processing, machine learning, and data mining.

▷ **For more information on this or any other computing topic, please visit our Digital Library at** www.computer.org/csdl.