ECPEC: Emotion-Cause Pair Extraction in Conversations

Wei Li, Yang Li, *Member, IEEE,* Vlad Pandelea, Mengshi Ge, Luyao Zhu, and Erik Cambria, *Fellow, IEEE*

Abstract—Conversational sentiment analysis (CSA) and emotion-cause pair extraction (ECPE) tasks have attracted increasing attention in recent years. The former aims to predict the sentiment states of speakers in a conversation, and the latter is about extracting emotion-cause clauses in a document. However, one drawback of CSA is that it cannot model the causal reasoning among emotion and neutral utterances from different speakers. In this work, we propose a new task: emotion-cause pair extraction in conversations (ECPEC), which aims to extract pairs of emotional utterances and corresponding cause utterances in conversations. The utterance-level ECPEC task is more challenging since the distance between emotion and cause utterances is larger than that of the clause-level ECPE task. To this end, we build a novel dataset ConvECPE and propose a specifically designed two-step framework for the new ECPEC task. Experimental results on ConvECPE dataset demonstrate the feasibility of the ECPEC task as well as the effectiveness of our framework.

Index Terms—Conversational sentiment analysis, emotional recurrent unit, contextual encoding, dialogue systems, multi-task learning

1 Introduction

Conversational sentiment analysis (CSA) is a task that consists in predicting the sentiment labels or sentiment intensities of a sequence of utterances in a dialogue [1]. It can be applied to many practical application scenarios, e.g., improving the performance of robot agents, tracking the mood of the speaker and so on. As a sub-task of CSA, Emotion-Cause Pair Extraction (ECPE) task, first proposed by [2], has aroused wide attention and become a hot research topic. The ECPE task aims to identify all potential pairs of emotions and the corresponding cause clauses in documents [2]. CSA predicts the sentiment states of utterances in conversations. Despite the increasing popularity and importance of CSA, emotional transmission (i.e., sentiment interactions) among speakers has not been taken into consideration in previous works on CSA to the best of our knowledge. To make full use of the sentiment states of all speakers and figure out how an utterance from one speaker affects the sentiment state of another speaker, we propose a new task: Emotion-cause Pair Extraction in Conversations (ECPEC), the aim of which is to extract emotion-cause utterance pairs (EC pairs) in conversations. This is important because the new task enables us to have a comprehensive understanding of the sentiment interactions among speakers. E.g., in the customer service system, this task can be applied to improving the response quality of the conversational agent by analyzing the interactions between user emotion and agent response.

We also present a new model since existing ECPE models like [3] and [4] are not specifically designed for the new ECPEC task. Compared with traditional documents, unique properties such as ungrammaticality, discontinuity, context-dependence and interactivity [5] make conversational data more difficult to analyze. For example, unlike the clause-level ECPE task, an utterance in ECPEC task is not necessarily to be a grammatically complete sentence. On the contrary, it can be built from single words, single phrases and non-lexical utterances (e.g., 'huh?') [6]. In addition, the distance between emotion and cause utterances in conversations is larger than the emotion-cause pair distance in documents. since the ECPEC task is formalized at utterance level instead of clause level. In the ECPE task, around 90% of the documents only have one emotion-cause pair [2] while the utterance-level ECPEC task has dozens of emotion-cause pairs in a conversation. Besides, an utterance may contain emotion and corresponding cause at the same time since there are several clauses in an utterance. These features make ECPEC a more challenging task, especially in long conversations.

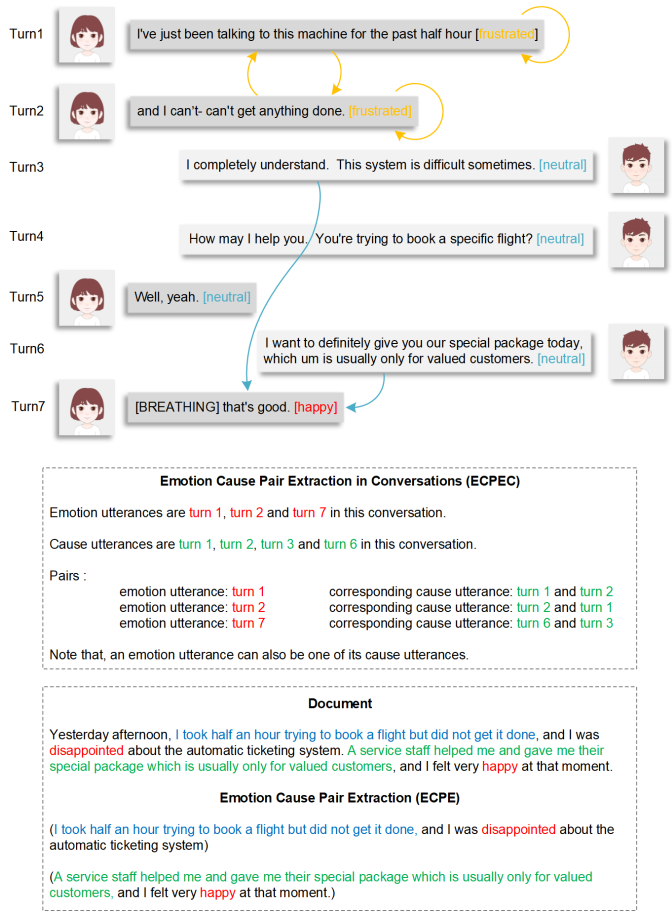
In Fig. 1, we illustrate the difference between the traditional ECPE task and our new ECPEC task. The snippet comes from a conversation in the IEMOCAP dataset [7] in which seven utterances are included. The document is adapted from the conversation snippet and composed of five clauses. In the document, the third and fifth clauses are emotion clauses containing emotions "happy" and "disappointed". Each of the emotional clauses has a corresponding cause clause. Here, the goal of the clause-level ECPE task is to extract the two EC pairs in this document. In the conversation, turns 1, 2 and 7 are emotional utterances. Each of them has two corresponding cause utterances. Taking turn 7 as an example, the emotion of the female speaker becomes "happy" because of the "understand" (turn 3) and "special package" (turn 6) from the male speaker. The goal of the utterance-level ECPEC task is to extract EC pairs in this conversation without knowing the emotion labels in advance.

To the best of our knowledge, there is no existing dataset for the ECPEC task. Therefore, we build an English conversational EC pair extraction dataset named ConvECPE, which is based on the famous interactive emotional dyadic motion capture database (IEMOCAP) [7]. It contains 7433 utterances that lie in 151 two-way conversations, where training and test sets contain 120 and 31 conversations, respectively. Each utterance carries one sentiment label from six sentiment labels, i.e., happy, sad, neutral, angry, excited and frustrated. Although Crawford [8] argued that this Ekman set of 6 basic emotions cannot capture the nuances of

- *W. Li, V. Pandelea, M. Ge, L. Zhu and E. Cambria are with School of Computer Science and Engineering, Nanyang Technological University, 639798, Singapore.*
  E-mail: wei008@e.ntu.edu.sg, cambria@ntu.edu.sg
- *Y. Li is with School of Automation, Northwestern Polytechnical University, China.*

(Corresponding Author: Erik Cambria)

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2022.3216551


Turn1
I've just been talking to this machine for the past half hour [frustrated]
Turn2
and I can't- can't get anything done. [frustrated]
Turn3
I completely understand. This system is difficult sometimes. [neutral]
Turn4
How may I help you. You're trying to book a specific flight? [neutral]
Turn5
Well, yeah. [neutral]
Turn6
I want to definitely give you our special package today, which um is usually only for valued customers. [neutral]
Turn7
[BREATHING] that's good. [happy]

Emotion Cause Pair Extraction in Conversations (ECPEC)

Emotion utterances are turn 1, turn 2 and turn 7 in this conversation.

Cause utterances are turn 1, turn 2, turn 3 and turn 6 in this conversation.

Pairs :

emotion utterance: turn 1	corresponding cause utterance: turn 1 and turn 2
emotion utterance: turn 2	corresponding cause utterance: turn 2 and turn 1
emotion utterance: turn 7	corresponding cause utterance: turn 6 and turn 3

Note that, an emotion utterance can also be one of its cause utterances.

Document

Yesterday afternoon, I took half an hour trying to book a flight but did not get it done, and I was disappointed about the automatic ticketing system. A service staff helped me and gave me their special package which is usually only for valued customers, and I felt very happy at that moment.

Emotion Cause Pair Extraction (ECPE)

(I took half an hour trying to book a flight but did not get it done, and I was disappointed about the automatic ticketing system)

(A service staff helped me and gave me their special package which is usually only for valued customers, and I felt very happy at that moment.)

Fig. 1: Comparison between ECPEC and traditional ECPE

emotional experience in the world, it is widely used in the CSA task and the corresponding dataset [9] since 6 basic emotions are concise and would benefit annotation consistency. Besides, it would be challenging for the application scenarios if incorporating the fine-grained sentiment in this task. Hence, we choose IEMOCAP with six basic emotional types as the basic dataset to build our ConvECPE dataset. Each non-neutral utterance has at least one cause utterance and at most three cause utterances. We do not label neutral utterances since sentiment interactions among the majority of neutral utterances are relatively weak. With this new dataset, we further develop an ECPEC-specific framework. The experimental results prove that, although challenging, the new ECPEC task is feasible. The main contributions of this work are summarized as follows:

- We propose a new task: emotion-cause pair extraction in conversations. It extends research on conversational emotion detection and helps the understanding of emotion interactions in conversations.
- Based on a CSA dataset, IEMOCAP, we build a high-quality dataset, ConvECPE for the new ECPEC task.
- We propose a framework consisting of two multi-task learning modules for this ECPEC task. The framework takes the properties of conversations into consideration.

We will release the ConvECPE dataset[1] and baseline systems online in the hope that this could contribute to research on EC pair extraction and a comprehensive understanding of sentiment interactions in conversations.

2 Related Work

2.1 EC Extraction

Lee et al. [10] first proposed the task of EC extraction, which is defined by extracting word-level cause of emotion expression in a given text. The authors manually built a corpus from Academia Sinica Balanced Chinese Corpus. After that, researchers developed plenty of works on this task setting, which can be categorized into rule-based models [11], [12], [13], [14] and machine learning models [15], [16], [17].

Considering that a clause may be a more appreciate unit for EC extraction, Chen et al. [15] transformed this word-level task into a clause-level task, and exploited six groups of linguistic cues to detect causes. Following this setting, numerous researchers make a contribution to algorithm achievement. Russo et al. [18] presented a linguistic patterns model augmented with common sense knowledge to classify the Italian sentence with a cause phrase. Gui et al. [19] extended the above idea into 25 groups of linguistic cues, then trained a machine learning model with SVM and CRF for EC extraction.

With the emerging trend of deep learning [20], an increasing number of researchers are applying deep neural networks for EC extraction. Gui et al. [21] expressed this task as a question answering task, and proposed a deep memory network with the convolution operation [22] to extract answers (cause) of questions (emotion). Li et al. [23] proposed a co-attention neural network to capture the correlation between emotion context and cause context. EC pair extraction was proposed by Xia and Ding [2] as a new task, whose objective is to extract the potential pairs of emotions and corresponding causes in documents. Compared with the existing document-level dataset, the conversation is more complex and has more application scenarios. Thus, it is essential to propose EC pair extraction in conversations.

2.2 Conversational Sentiment Analysis

Sentiment analysis is one of the most important tasks in natural language processing (NLP) because of its potential applications in a wide area of systems, including opinion mining [24], health-care [25], recommendation systems [26], education [27], etc.

In recent years, CSA attracted researchers' attention. Poria et al. [28] presented an LSTM-based model to preserve the sequential order of utterances and share information of consecutive utterances. Hazarika et al. [29] proposed Interactive COnversational memory Network to extract multimodal features from conversational videos and hierarchically capture the self- and inter-speaker emotional features into global memories. Besides, Majimder et al [30] keep track of states of speakers by modeling the party state, global state and emotional dynamics. Li et al. [1] utilized neural tensor networks to do context compositionality in conversations and proposed a two-channel feature extractor for sentiment analysis in conversations. However, knowing the sentiment of each utterance in conversations is not enough for us to understand the sentiment interactions among speakers and cannot provide informative knowledge for the application scenarios like the conversational agent in the customer service system. Hence, our work is proposed to deal with the drawback of the CSA task and promote the application of sentiment analysis in conversations.

[1] https://github.com/Maxwe11y/JointEC/tree/main/Dataset

3 Task Description and ConvECPE Dataset

3.1 Problem Definition

First of all, we give the definition of our ECPEC task. Given a multi-turn conversation $U = [u_1, u_2, ..., u_{|d|}]$, the ECPEC task is to extract a set of EC pairs in U:

$$P = \{..., (u^e, u^c), ...\}, \tag{1}$$

where u^e is an utterance with certain emotion and u^c is the corresponding cause utterance. Following the definition of cause events by Lee [10], the cause utterance of an emotion utterance is an utterance containing explicitly or implicitly expressed arguments or events evoking (or partially evoking) the presence of the corresponding emotion. Unlike the traditional ECPE task, the ECPEC task is defined at utterance level instead of clause level. Therefore, the "emotion" and "cause" used in this paper refer to "emotional utterance" and "cause utterance", respectively.

3.2 Dataset Annotation

As there is no existing dataset for the ECPEC task, we introduce a conversational emotion-cause pair extraction dataset (ConvECPE) in this paper. Similar to RECCON [31], our ConvECPE dataset is constructed based on the existing interactive emotional dyadic motion capture database (IEMOCAP) [7]. The IEMOCAP is a dataset of two-way conversations involving ten distinct participators [1], where six different sentiments (happy, sad, neutral, angry, excited and frustrated) are included. In this paper, we denote utterances (labeled as happy, sad, angry, excited or frustrated) as emotional utterances and neutral utterances as non-emotional utterances.

Because the sentiment interactions of most non-emotional utterances are relatively weak, we only annotate emotional utterances in this task. Due to the discontinuity and interactivity of conversation, emotional utterances with more than one cause utterance frequently occur while most of emotion clauses in documents have only one cause clause. Considering that the number of cause utterances of different emotional utterances varies from each other, we set the maximum number of causes to three to reduce the complexity of the ECPEC task. The predefined annotation rules are as follows:

1) The cause label of a given emotional utterance can be the emotional utterance itself or any other utterance within the conversation. If an emotional utterance contains both emotion related expression and its corresponding cause expression, then the current emotional utterance is regarded as its cause. Taking the utterance "I am not feeling good. I've been out of work." as an example, the emotion of this utterance is frustrated and we find that the event unemployment evokes this emotion. Therefore, this utterance contains emotion and corresponding cause in the mean time. In the original word-level emotion cause extraction (ECE) task [15], about 85% of the emotion causes are in the same clause where the emotion keywords are, which makes the ECE task less complex. In our corpus, the percentage that emotion and corresponding cause are in the same utterance is less than 30%.

2) Assuming that there are three emotional utterances u_a, u_b and u_c in a conversation, u_b and u_c are both the cause of u_a and u_c is also the cause of u_b. In this case, the priority is given to u_b when labeling the causes of u_a since u_b is the direct cause of u_a. For example, there is a conversation snippet:

A : `I was admitted to Stanford University`
(u_c, ***happy***)
B : `Congrats! That is a great university.`
(u_b, ***happy***)

ConvECPE	Zero	One	Two	Three
Match	992	2558	1794	381
Ratio	17.33%	44.68%	31.34%	6.65%

TABLE 1: The labeling result in round one.

A : `Thanks! this is my dream school.`
(u_a, ***excited***)

In this example, both u_c and u_b are the causes of u_a. Obviously, the emotion of u_b is affected by the good news in u_c. The feedback from speaker B further strengthens the positive emotion of the speaker A in u_a. Therefore, u_b is regarded as the most direct cause of u_a in this case.

Based on the annotation rules mentioned above, we design an annotation workflow. In the first round labeling, we invite two expert annotators who have enough knowledge for sentiment analysis and dialogue systems to independently label each conversation. Due to the number of cause utterances for emotional utterances varies from one to three, it is impossible to use kappa to measure the interrater agreement. Hence, we use the statistical information in Table 1 to show the quality of the first round annotation. Specifically, only 17.33% of the annotations from two annotators for the same utterance have no intersection, and around 82.67% have at least one intersection, indicating that the judgment between our two annotators is basically consistent. Then we collect the labeling results and compare the differences between them. We only keep the results that are both labeled as cause utterances by the two annotators. In the second round, we invite a third expert annotator to process those emotional utterances that have no intersection in the first round (only 17.33% of them need to be processed). We get the union set of the labels given by the previous two annotators for each utterance. In this case, the third annotator is supposed to select the most relevant labels from the union set; Otherwise, we invite experts to discuss those utterances and obtain the final results by means of a majority vote (around one percent of the utterances need to be processed). Discussion is used as the final solution to process the conflicts between annotators [9].

3.3 ConvECPEC Dataset Analysis

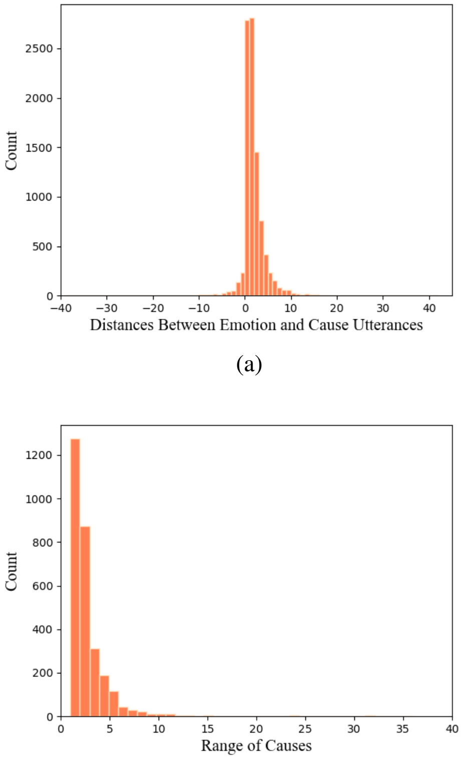
The ConvECPE dataset consists of 151 dialogues with 7,433 utterances, among which 1,708 are neutral and 5,725 are emotional. The proportion of emotional utterances with the different number of cause utterances is shown in Table 2: 47.12% have one cause, 36.82% have two causes and only 14.32% have three; 29.39% are the causes for themselves.

Fig. 2a shows the distribution of distances between emotional utterances and their causes. The distance is calculated over $9,473$ EC pairs according to the formula $id_e - id_c$. Here, id_e and id_c are the index of the emotional utterance and corresponding cause utterance, respectively. Here, the sign of the distance corresponds to the relative directions instead of the real values. As in Fig. 2a, the distances between emotion and cause utterances are within the range of $[id_e - 42, id_e + 39]$. And the majority of the cause utterances of current emotional utterance are within a window in which the index ranges from $id_e - 10$ to $id_e + 5$. Therefore, it is essential to determine a window within which the classifier extracts EC pairs. An appropriate window size should help to reduce the computational complexity and maintain the performance as well.

However, as described before, because more than 50% of the emotional utterances have more than one cause, it is necessary

ConvECPE	Dialogue	Utterance	Neutral	Pairs	Pairs*	One	Two	Three
ALL	151	7,433 100%	1,708 22.98%	9,473 100%	2,784 29.39%	2,797 48.86%	2,108 36.82%	820 14.32%
Train set	120	5,810 100%	1,324 22.79%	7,558 100%	2,266 29.98%	2,114 47.12%	1,672 37.27%	700 15.60%
Test set	31	1,623 100%	384 23.66%	1,915 100%	518 27.05%	683 55.13%	436 35.19%	120 9.69%

TABLE 2: Statistical information of the ConvECPE dataset. *Neutral* stands for non-emotional utterance. Here *Pair* refers to EC pair and *Pair** refers to the pair where emotion and cause utterances are the same. *One*, *Two* and *Three* are the number of emotional utterance with one, two or three causes, respectively.



(a)

(b)

Fig. 2: The distribution among the EC pairs. (a) denotes the distribution of distances between emotion and cause utterances. (b) denotes the distribution of cause range. Here range is defined as the maximum distance among causes of an emotional utterance.

to know how these causes distribute w.r.t. current utterance. We calculated the distance among these causes (i.e., cause range) with $|id_{c_1} - id_{c_2}|$ and plotted the distribution in Fig. 2b. Apparently, cause range is a power-law distribution, with more than 80% of them smaller than 3. This means we can search the cause within a certain range (e.g., chunk) without losing accuracy. Therefore, we add an auxiliary task emotion-cause-chunk (EC-chunk) pair extraction to the EC pair extraction task to improve the search efficiency, and the details are introduced in section 4.

To facilitate the usage of the dataset and produce comparable results, we follow the original partition of the IEMOCAP dataset [7]. Thus, the training set consists of 120 dialogues(5,810 utterances) and the test set consists of 31 dialogues(1,623 utterances). Generally, the maximum utterance number is 110 and the average is 49.23.

4 Approach

Considering that the number of causes of emotional utterances ranges from one to three, it is hard to directly extract EC pairs in conversations. Therefore, we propose a two-step framework to address this new ECPEC task. In the first step, a multitask model is advanced to jointly detect emotion and cause utterances in conversations. In the second step, an auxiliary task EC-chunk pair extraction is presented to provide informative features which help more accurately extract EC pairs. Here is the detailed definition of the proposed framework:

Step 1 (Joint Emotion and Cause Detection) Here, the goal is to detect whether an utterance is emotional or not and whether it is a cause utterance of any utterance in the conversation. First, we obtain a set of emotional utterances $\boldsymbol{E} = \left\{u_1^e, ..., u_{m_e}^e\right\}$ and cause utterances $\boldsymbol{C} = \left\{u_1^c, ..., u_{m_c}^c\right\}$ for each conversation. To do so, two kinds of multi-task learning networks are presented to jointly extract emotion and cause utterances in a conversation.

Step 2 (EC-Chunk Pair Extraction and EC pair Ectraction) Since there are more than 100 utterances in some conversations, directly pairing emotion set $\boldsymbol{E}$ and cause set $\boldsymbol{C}$ by applying a Cartesian product is not an efficient choice. Based on the statistical information in Fig. 2a and Fig. 2b, the causes of an emotional utterance lie in a small region of the conversation in most situations. In this case, it is easy to differentiate this small cause region from other non-cause regions. Thus, we split $\boldsymbol{C}$ into several cause-chunks and pair the emotional utterance with the cause-chunk instead of the cause. Here cause-chunk is defined as a sequence of contiguous cause utterances in $\boldsymbol{C}$. Then, we train an EC-chunk filter to recognize the candidate EC-chunk pairs, where the correct cause-chunk should contain the cause/causes of the current emotional utterance. For the EC pair extraction task, we pair the emotion set $\boldsymbol{E}$ and the cause set $\boldsymbol{C}$ by applying a Cartesian product. This yields a set of candidate EC pairs. Finally, we train the EC pair filter to select the pairs that contain a causal relationship between emotion and cause utterances. We elaborate each step in the following subsections.

4.1 Step 1: Joint Emotion and Cause Detection

In this ECPEC task, a conversation is composed of a sequence of d utterances. Each utterance in a conversation consists of a number of words which are represented as vectors $\boldsymbol{V_i} = \left[v_i^1, v_i^2, ..., v_i^n\right]$ through the pre-trained GloVe [32] word vectors. Here i represents the ith utterance and n is the number of words in the utterance. As shown in Fig. 3, a two-layer Long Short-Term Memory [33] (LSTM) module is employed as the tool to generate utterance-level embeddings. The LSTM module takes word vectors of an utterance as inputs and outputs the embedding of this utterance. For example, if we input $\boldsymbol{V_i}$, then the utterance embedding u_t can be obtained accordingly. As a result, the input conversation can be represented as a set of d embeddings: $\boldsymbol{U} = [u_1, u_2, ..., u_d]$.

Following the results in [1], we take adjacent utterances u_{i-1} and u_{i+1} of utterance u_i as its context at time step i. Then we employ a linear layer to perform context compositionality for utterance pairs (u_{i-1}, u_i) and (u_i, u_{i+1}) on both tasks, and get context-aware vectors $p^e_{i,f}$ and $p^e_{i,b}$ for emotion detection task and $p^c_{i,f}$ and $p^c_{i,b}$ for cause detection task, respectively. Taking the cause detection task in Fig. 3 as an example, the context-aware vectors $p^c_{i,f}$ and $p^c_{i,b}$ are fed into forward and backward LSTM modules [33] to obtain s^c_i which is the concatenation of the forward and backward LSTM outputs. s^e_i can be obtained analogously.

According to Xia and Ding [2], emotion information contributes to detecting the causes and vice versa. Previous works also proved that soft-parameter sharing can effectively improve the performance of multi-task learning model [34], [35], [36]. Therefore, we propose two multi-task learning variants with different soft information sharing mechanisms to augment the performances of both emotion and cause detection tasks. The one with cross attention mechanism [37] is named as Joint-Xatt and the other with graph convolutional network (GCN) is called Joint-GCN in this paper. The sharing mechanisms play an important role in the joint learning structure and are elaborated in detail in the following part.

As shown in Fig. 3, the features $\boldsymbol{s}^e = \{s^e_1, ..., s^e_{N_e}\}$ of the emotion detection task and $\boldsymbol{s}^c = \{s^c_1, ..., s^c_{N_c}\}$ of the cause detection task are transmitted to the soft information sharing module. N_e and N_c are the number of emotion and cause utterances, respectively. Concretely, in the case of Joint-Xatt, the additional feature for s^e_i can be obtained by the the following equations of cross attention mechanism:

$$\begin{aligned} c_{i,j} &= s^e_i \boldsymbol{W_c} s^c_j \\ \alpha_{i,j} &= \frac{exp(c_{i,j})}{\sum_k exp(c_{i,k})} \\ \hat{s}^e_i &= s^e_i \oplus (\sum\nolimits_j \alpha_{i,j} s^c_j) \end{aligned} \tag{2}$$

where the matrix $\boldsymbol{W}_c$ is the model parameter and $\oplus$ denotes vector concatenation. Here emotion detection is the target task. When cause detection becomes the target task, the calculation of the additional feature for s^c_i is performed in a symmetric manner, where s^c_i and s^e_i exchange their positions in Formulas 2 and $\hat{s}^c_i$ can be obtained accordingly. Finally, $\boldsymbol{s}^e_{xatt} = \{\hat{s}^e_1, ..., \hat{s}^e_{N_e}\}$ and $\boldsymbol{s}^c_{xatt} = \{\hat{s}^c_1, ..., \hat{s}^c_{N_c}\}$ are obtained and fed into a fully connected layer followed by a softmax layer for classification.

For Joint-GCN model, we model the interactions between $\boldsymbol{s}^e = \{s^e_1, ..., s^e_{N_e}\}$ and $\boldsymbol{s}^c = \{s^c_1, ..., s^c_{N_c}\}$ via a directed graph. Each node in the graph represents a vector in these two sets. Edges represent dependency between vectors from the sets $\boldsymbol{s}^e$ and $\boldsymbol{s}^c$. In other words, we focus on the inter-dependency instead of intra-dependency between these two sets of vectors. By feeding this graph to a GCN [38] consisting of one convolution operation, the information is propagated through vectors in $\boldsymbol{s^e}$ and $\boldsymbol{s^c}$. Here we omit the details of GCN, and readers may refer to [39] and [40]. Similarly, the features $\boldsymbol{s}^e_{gcn} = \{\hat{s}^e_1, ..., \hat{s}^e_{N_e}\}$ and $\boldsymbol{s}^c_{gcn} = \{\hat{s}^c_1, ..., \hat{s}^c_{N_c}\}$ are obtained and fed to the output layer. Here we employ $\boldsymbol{s}^e_{gcn}$ and $\boldsymbol{s}^c_{gcn}$ to introduce the equations for emotion and cause detections:

$$\begin{aligned} \hat{\boldsymbol{y}}^e &= softmax(\boldsymbol{W}^e \boldsymbol{s}^e_{gcn} + \boldsymbol{b}^e) \\ \hat{\boldsymbol{y}}^c &= softmax(\boldsymbol{W}^c \boldsymbol{s}^c_{gcn} + \boldsymbol{b}^c) \end{aligned} \tag{3}$$

where $\boldsymbol{W}^e$, $\boldsymbol{b}^e$, $\boldsymbol{W}^c$ and $\boldsymbol{b}^c$ are the model parameters. The loss of Joint-GCN is a weighted sum of two parts:

$$\boldsymbol{L} = \lambda \boldsymbol{L}^e + (1 - \lambda) \boldsymbol{L}^c \tag{4}$$

where $\boldsymbol{L}^e$ and $\boldsymbol{L}^c$ are the cross-entropy error of emotion and cause detections, respectively. Here λ is a trade-off parameter. The loss of Joint-Xatt can be computed in the same way as Joint-GCN.

4.2 Step 2: EC-Chunk Pair Extraction and EC Pair Extraction

After *step 1*, we get a set of emotions $E = \{u^e_1, ..., u^e_{m_e}\}$ and cause utterances $C = \{u^c_1, ..., u^c_{m_c}\}$. As mentioned before, instead of directly filtering candidate EC pairs, we propose an auxiliary task EC-chunk pair extraction to enhance the performance of the EC pair extraction task. Then these two tasks are combined into a unified multi-task model, named Joint-EC.

In the Joint-EC model, we follow the same procedure as in *step 1* models to generate utterance-level embeddings from pre-trained GloVe word vectors. As shown in Fig. 4, The EC-chunk filter and EC pair filter are combined in a parallel structure. Both of them take the emotion utterance vectors u^e and cause utterance vectors u^c as inputs, except that the output of EC-chunk filter $\hat{y}^{ck}$ is fed into the EC pair filter as well, which makes this structure an interactive multi-task learning framework. The EC-chunk filter is designed to extract EC-chunk pairs. The equation of EC-chunk filter is as follows:

$$\begin{aligned} &\boldsymbol{\phi}_{i,j} = \boldsymbol{u}^e_i \oplus \boldsymbol{u}^{ck}_j \oplus (\|\boldsymbol{u}^e_i - \boldsymbol{u}^{ck}_j\|) \oplus (\boldsymbol{u}^e_i \odot \boldsymbol{u}^{ck}_j) \oplus \boldsymbol{p}^{ck}_j \\ &\boldsymbol{g}_{i,j} = LeakyReLU(\boldsymbol{W}^g \boldsymbol{\phi}_{i,j} + \boldsymbol{b}^g) \\ &\hat{\boldsymbol{y}}^{ck}_{i,j} = softmax(\boldsymbol{W}^{ck} \boldsymbol{g}_{i,j}) \end{aligned} \tag{5}$$

where $\boldsymbol{W}^g$, $\boldsymbol{b}^g$ and $\boldsymbol{W}^{ck}$ are model parameters, $\odot$ denotes element-wise multiplication. u^{ck}_j is the embedding of the jth cause-chunk. Cause-chunk u^{ck} is obtained by splitting cause vectors u^c and attention mechanism [41] is applied to obtaining chunk-level embedding. $\boldsymbol{p}^{ck}_j$ is a fixed position embedding of the jth cause-chunk. $\boldsymbol{g}_{i,j}$ is the final feature vector and $\hat{\boldsymbol{y}}^{ck}_{i,j}$ the prediction result.

A Cartesian product is applied to pairing all the possible EC pairs. Then we apply EC pair filter composed of the classification module [37] to extract EC pairs. The equation of EC pair filter is as follows:

$$\begin{aligned} &\boldsymbol{\delta}_{i,j} = \boldsymbol{u}^e_i \oplus \boldsymbol{u}^c_j \oplus (\|\boldsymbol{u}^e_i - \boldsymbol{u}^c_j\|) \oplus (\boldsymbol{u}^e_i \odot \boldsymbol{u}^c_j) \oplus \boldsymbol{p}^p_j \\ &\boldsymbol{f}_{i,j} = LeakyReLU(\boldsymbol{W}^f \boldsymbol{\delta}_{i,j} + \boldsymbol{b}^f) \\ &\hat{\boldsymbol{y}}_{i,j} = softmax(\boldsymbol{W}^p \boldsymbol{f}_{i,j}) \end{aligned} \tag{6}$$

where $\boldsymbol{W}^f$, $\boldsymbol{b}^f$ and $\boldsymbol{W}^p$ are model parameters, $\odot$ denotes element-wise multiplication. $\boldsymbol{p}^p_j$ is a fixed position embedding of the j-th cause u_j. $\boldsymbol{f}_{i,j}$ is the final feature vector and $\hat{\boldsymbol{y}}^p_{i,j}$ the prediction result. Motivated by teacher forcing mechanism [42] used in machine translation, with a certain probability, we feed ground truth labels of textitstep 1 into *step 2* models to further enhance the performance in *step 2* and reduce overfitting [43] to some extent. Besides, we propose the window-restricted Joint-EC(named as Joint-ECW) which is a standard Joint-EC model taking $u^e_i - u^c_j$ as input where $i - j \in [-window, window]$. Window-restricted mechanism is proved to be computation-efficient [4]. In addition, it is capable of alleviating unbalanced sample issue.

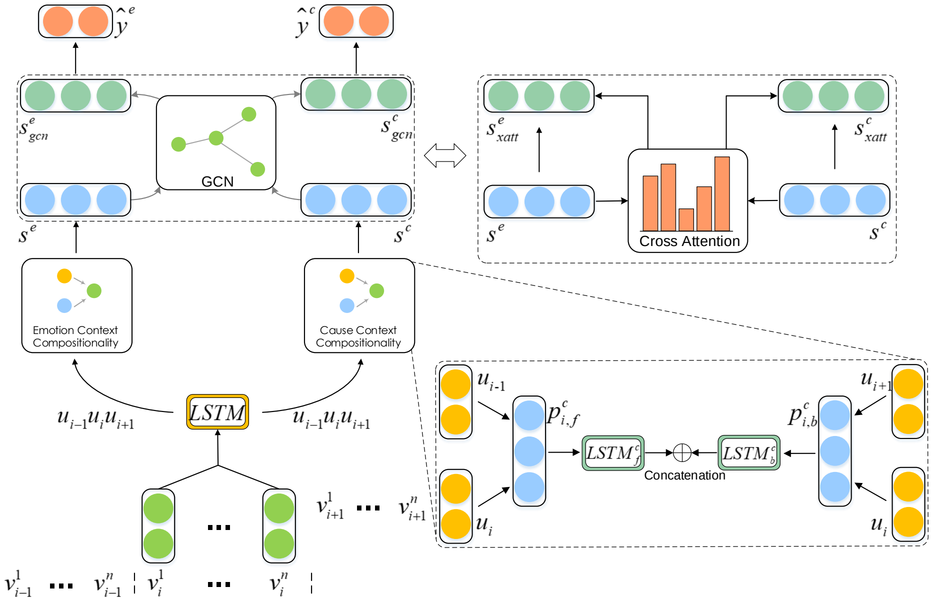

GCN
Cross Attention
Emotion Context Compositionality
Cause Context Compositionality
LSTM
Concatenation

Fig. 3: The proposed two models for joint emotion and cause utterances detection. Joint-GCN is illustrated in the left side and it uses graph convolutional network as the soft information sharing method to enhance the emotion and cause detection tasks. The structure of Joint-Xatt model is similar to Joint-GCN and the difference is that Joint-Xatt model employs cross attention as the soft information sharing method to improve the performances of both tasks. Take Joint-GCN as an example, it has emotion and cause context compositionality modules for emotion and cause detection tasks, respectively. These two modules have the same structure, hence we only display the cause context compositionality module in this figure.

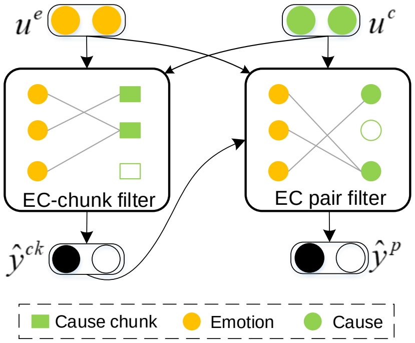


Fig. 4: Illustration of Joint-EC

The loss of Joint-EC model is a sum of two parts:

$$\boldsymbol{L} = \boldsymbol{L}_{ck} + \boldsymbol{L}_{p} \tag{7}$$

where $\boldsymbol{L}_{ck}$ and $\boldsymbol{L}_{p}$ are the cross-entropy error of EC-chunk pair and EC pair extraction tasks, respectively.

5 Experiments

5.1 Metrics

Similarly to previous work on ECPE [2], we evaluate our models on precision, recall and F1 score. We report the definition of these metrics:

$$Precision = \frac{\sum correct\_pairs}{\sum proposed\_pairs} \tag{8}$$

$$Recall = \frac{\sum correct\_pairs}{\sum annotated\_pairs} \tag{9}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{10}$$

where *correct_pairs* is the number of emotion, cause, or EC pairs (depending on the task) correctly identified by the model. *proposed_pairs* is the total number of emotion, cause, or EC pairs predicted by the model, including the false positives. Lastly, *annotated_pairs* is the total number of emotion, cause, or EC pairs present in the dataset. Precision measures the ability of a classifier to correctly detect emotion(cause) utterances from proposed emotion(cause) utterances. Recall is to measure how many annotated emotion(cause) utterances are found. F1 score is the harmonic mean of precision and recall and it is a consolidated indicator to evaluate the performance of a classifier.

5.2 Baseline and Setting

5.2.1 *BERT baseline*

To assess how well our model fares compared to other solutions, we also experiment with a BERT-based baseline [44], where we apply it to the emotion and cause detection tasks separately, as well as to the EC pair extraction. More specifically, we employ the pre-trained BERT base model followed by a layer to perform the binary classification task. We use a maximum sequence length of 128 and train the model using the Adam optimizer with a learning rate of $1e-6$ and minimize the weighted binary cross-entropy loss function. For the EC pair prediction task we train and evaluate

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2022.3216551

the model on all of the possible pairs within a conversation. In order to address the class imbalance problem, as well as to reduce the training time, we use negative sampling, resulting in roughly the same number of negative and positive samples at each training iteration. Moreover, we experiment with a windowed version of this model, taking into account only the utterances that lie within 10 utterances of each emotional utterance.

5.2.2 ECPE baselines

The models (e.g., Inter-CE, Inter-EC) from [2] are currently the most relevant to our proposed task. Therefore, it is necessary to make such comparisons. The code we use is publicly available [2], and it is the default setting applied during the model training. Specifically, the maximum sequence length is set to 128, the learning rate is 0.005 and the batch size is set to 32.

5.2.3 Joint-Xatt, Joint-GCN, Indep and Joint-EC

As mentioned in subsection 4.1, we use the pre-trained 100-dimensional GloVe word vectors[3] for Joint-GCN, Joint-Xatt and Indep. For Joint-EC model, we also use the pre-trained 100D GloVe word vectors and set chunk size as 2. For training details, we use Adam [45] as the optimizer. Batch size and learning rate are set to 1 and $1e-3$, respectively. To alleviate overfitting [43], dropout is applied and set to 0.4. Indep is a variant of *step 1* model, where the emotion and cause utterances are detected individually. Window size of Joint-ECW is set to 8 in this experiment.

5.3 Evaluation on ECPEC task

The emotion detection task is defined as a binary classification by predicting whether the utterance carries emotion, and it is the same case in the cause detection task. As mentioned earlier, each utterance may have one or more causes, therefore, all of these causes are treated as true labels in the prediction. Table 3 shows the results for our proposed *step 1* models and other baselines. Here we use the average F1 score (**AVG**) on emotion detection and cause detection tasks to evaluate overall performances of *step 1* models since a high **AVG** corresponds to more emotion-cause pairs, which would benefit *step 2* models. From the overall perspective of AVG, we observe that our proposed *step 1* models get better performances than ECPE related models (i.e., Inter-CE, Inter-EC), and even outperform the BERT model by a large margin. These results validate the effectiveness of the proposed *step 1* models.

Among our proposed *step 1* models, Joint-GCN and Joint-Xatt achieve almost the same average F1 score on both tasks, which shows that both GCN and cross attention are effective soft information sharing mechanisms for joint learning of emotion and cause detection tasks. In particular, compared with Indep, Joint-GCN and Joint-Xatt improve the performance of cause detection task by a large margin without reducing the performance of emotion detection too much. Compared with the interactive multi-task learning modules of Inter-EC and Inter-CE which directly employ the prediction results of one task as additional features to improve the performance of another task, learnable soft information sharing modules GCN and cross attention are more helpful and powerful. If the prediction results of one task is incorrect, the error would broadcast as the prediction is directly used as part of the input features of another task. Our learnable soft information sharing modules do not directly share the prediction results. On the contrary, they take the coupling of two tasks into consideration to share useful features.

The performances of BERT baseline are far behind the other models including Inter-CE, Inter-EC and our *step 1* models. However, BERT model adopts a different detection strategy which is noteworthy. To be specific, BERT is inclined to detect emotion or cause utterances as precisely as possible instead of detecting emotion or cause utterances as many as possible. In contrast, Inter-EC and Inter-CE tend to achieve a high recall but are not good at correctly detecting emotion or cause utterances. The possible reason is that the emotion and cause clauses in the ECPE documents are sparse (most documents contain only one EC pair [2]); whereas, there are around 38 emotional utterances on average in a conversation, and each of them has one or more causes. This misleads ECPE related models (i.e., InterCE, InterEC) into predicting that almost all utterances are the emotional utterances and most utterances are cause utterances. Our *step 1* models have balanced performances over two metrics recall and precision, and thus get better F1 scores than ECPE related models.

By summarizing the results in Table 3, we find that the performances of all the models on emotion detection task are far better than those on cause detection task. According to the results in ECPE task [2], the performance gap between emotion and cause detection tasks is even larger(around 20%). One possible reason is that cause detection is a more challenging task. In most cases, an utterance itself can provide sufficient information for detecting whether it is emotional or not. However, this is not applicable to the detection of cause utterances since cause utterances are highly related to their corresponding emotional utterances. In future research, the unique properties of cause utterances should be taken into account to design cause utterance-specific structure.

As mentioned before, the Joint-EC model is applied to extract EC pairs from the detected emotion and cause utterances. To verify the feasibility of the ECPEC task and evaluate the performance of the proposed Joint-EC model, we introduce a series of Joint-EC based models and compare these models with baselines. To be specific, Joint-EC based models are as follows: Joint-EC(BERT_Indep), Joint-EC(Inter-EC), Joint-EC(Inter-CE), Joint-EC(Indep), Joint-EC(Joint-Xatt), Joint-EC(Joint-GCN), Joint-ECW(BERT_Indep), Joint-ECW(Inter-EC), Joint-ECW(Inter-CE), Joint-ECW(Indep), Joint-ECW(Joint-Xatt), Joint-ECW(Joint-GCN), where Joint-EC is the standard Joint-EC model and Joint-ECW is window-restricted Joint-EC model. The names between brackets refer to the *step 1* models, and Joint-EC or Joint-ECW is the name for *step 2* model. Take Joint-EC(BERT_Indep) model as an example, it employs BERT in the first step to extract the emotion and cause utterances independently and then applies our Joint-EC framework to extract EC pairs. BERT is employed as the baseline model in the experiment. The results are listed in Table 4.

Applying BERT to extract target EC pairs from all candidate pairs is time consuming with low extraction accuracy, since it deals with tens of thousands of pairs. This is validated in the results of BERT. BERT model tends to classify most of the candidate EC pairs as true EC pairs and this yields a high recall. However, considering that true EC pairs are very rare among candidate EC pairs, it is hard for BERT to precisely extract true EC pairs in this

[2] https://github.com/NUSTM/ECPE

[3] http://nlp.stanford.edu/data/glove.6B.zip

[1] The window size is set to 12 for Joint-ECW(BERT_Indep) to run this experiment.

[2] We use Joint-GCN as the input for Joint-EC-Bound and Joint-EC-Bound.

	Emotion Detection			Cause Detection			AVG
	Precision	Recall	F1	Precision	Recall	F1	
Joint-GCN	83.05%	95.32%	**88.76**%	71.47%	86.35%	78.21%	**83.49**%(*8e-4*)
Joint-Xatt	83.32%	94.35%	88.49%	69.68%	**89.42**%	**78.33**%	83.41%(*3e-4*)
Indep	80.70%	96.21%	87.78%	72.76%	80.55%	76.46%	82.12%(*5e-4*)
Inter-CE	76.34%	**100.00**%	86.58%	67.79%	86.92%	76.17%	81.38%(*0*)
Inter-EC	76.34%	**100.00**%	86.58%	68.55%	85.55%	76.11%	81.35%(*0*)
BERT	**90.71**%	67.80%	77.60%	**75.64**%	74.18%	74.90%	76.25%(*3e-3*)

TABLE 3: Results on emotion detection and cause detection tasks, AVG denotes the average F1 score of the two tasks, the higher the better. The reported results are the average of 3 runs. The numbers in brackets are the standard deviation of the average F1 scores.

Combinations	PR	REC	F1
BERT	9.81%	60.57%	16.89%(*2e-3*)
BERT-Window	13.80%	**68.72**%	22.98%(*3e-3*)
Joint-EC(BERT_Indep)	35.68%	24.39%	28.97%(*4e-3*)
Joint-ECW[1](BERT_Indep)	37.31%	35.93%	36.61%(*1e-3*)
Joint-EC(Inter-EC)	30.91%	37.34%	33.82%(*2e-3*)
Joint-ECW(Inter-EC)	30.76%	57.13%	39.99%(*7e-3*)
Joint-EC(Inter-CE)	30.05%	37.75%	33.46%(*5e-3*)
Joint-ECW(Inter-CE)	31.24%	58.06%	40.62%(*1e-3*)
Joint-EC(Indep)	37.38%	31.23%	34.03%(*4e-3*)
Joint-ECW(Indep)	34.79%	52.95%	41.99%(*3e-3*)
Joint-EC(Joint-Xatt)	38.23%	37.08%	37.65%(*3e-3*)
Joint-ECW(Joint-Xatt)	37.12%	56.09%	44.67%(*3e-3*)
Joint-EC(Joint-GCN)	**42.79**%	35.35%	38.72%(*4e-3*)
Joint-ECW(Joint-GCN)	38.16%	59.27%	**46.43**%(*4e-3*)
Joint-EC-Bound[2]	#52.67%	#39.63%	#45.23%(*2e-3*)
Joint-ECW-Bound	#49.27%	#59.66%	#53.97%(*3e-3*)

TABLE 4: The results of the two-step framework and baseline models. PR and REC are the abbreviation of precision and recall, respectively. The reported results are the average of 3 runs. The numbers in brackets are the standard deviation of F1 scores.

case. If applying window restriction to the BERT model, then the BERT-Window model achieves a better precision and recall in the mean time. Nevertheless, the performance of BERT-Window is still poor in terms of precision. The proposed twelve models, by contrast, show a different scheme which is to reduce the *proposed_pairs* without decreasing *correct_pairs* too much. In this case, *precision* increases a lot while *recall* slightly decreases. Among these models, our proposed Joint-ECW(Joint-GCN) and Joint-ECW(Joint-Xatt) get the best performances in terms of F1 score. In addition, Joint-ECW based models surpass the corresponding Joint-EC based models by 7.28% on average. Take Joint-ECW(Joint-GCN) and Joint-EC(Joint-GCN) as an example, there are 145,228 proposed EC pairs in the training set of Joint-EC(Joint-GCN) among which 7,558 pairs are true EC pairs. If we apply window restriction to the Joint-EC, then the proposed pairs reduce to 47,049 and the true EC pairs are 7,424 in this case. With the help of the window restriction, we obtain a more balanced training set, which would benefit the filter. Meanwhile, this also demonstrates the effectiveness of the window restriction mechanism.

Combining the results in Table 3 and Table 4, we observe that the performances of EC pair extraction are positively correlated to the *step 1* performances. The higher the performance in *step 1* is achieved, the higher the performance in EC pair extraction is observed. For instance, Joint-EC(GCN) performs the best in terms of average F1 score in *step 1*, and its performance surpasses the other models in *step 2*. In contrast, BERT has a low average F1 score in *step 1*. Thus its F1 score in *step 2* is far behind the other models. Besides, we find that Joint-EC(Inter-CE) is slightly behind the Joint-EC(Inter-EC) while Joint-ECW(Inter-CE) is slightly better than Joint-ECW(Inter-EC) in *step 2*, although the Inter-CE and Inter-EC have almost the same performance in *step 1*. One possible reason is that the extracted emotion and cause utterances by Inter-EC and Inter-CE may contain different numbers of true EC pairs under different conditions. For instance, both Inter-EC and Inter-CE extract 100 emotion utterances and 100 cause utterances in *step 1*, where there are 50 true EC pairs in Inter-EC and 40 true EC pairs in Inter-CE. Joint-EC(Inter-EC) may obtain a better performance than Joint-EC(Inter-CE) in this case. Nonetheless, Inter-CE may have more true EC pairs than Inter-EC under the restriction of window mechanism, which may account for the better performance of Joint-ECW(Inter-CE). In addition, the above analysis can also be used to analyze the performance gap between Joint-ECW(Joint-Xatt) and Joint-ECW(Joint-GCN). The F1 score of the latter outperforms the former by a large margin, even though the AVG F1 scores of the two models are comparable. In the training set of Joint-ECW(Joint-Xatt), there are 52,660 EC pairs among which 6,937 are true EC pairs. As mentioned before, 7,424 out of the 47,049 EC pairs are true in the training set of Joint-ECW(Joint-GCN). In fact, Joint-GCN is a better filter than Joint-Xatt, which cannot be reflected by the current metrics. In future research, we need to design a more accurate measurement to evaluate the performance of the *step 1* filter.

To explore the upper bound of the performance, we use the annotated label directly rather than the output from the first step as the input of Joint-EC. However, this cannot be fairly compared with other models and is marked by a "#". Nevertheless, we can still get some insights from the results of Joint-EC-Bound and Joint-ECW-Bound. With the window restriction, both Joint-EC-Bound and Joint-ECW-Bound dramatically improve the F1 score by more than 6%. The results of Joint-EC-Bound and Joint-ECW indicate that it is essential to keep improving the performances in *step 1*. Additionally, in future research, if prior knowledge like the interaction pattern [46] is incorporated in a conversation into a model like Joint-ECW, we may extract more EC pairs and get better performance on this task.

5.4 Ablation Study

To further explore the proposed Joint-EC model, we perform an ablation study to investigate the influence of the proposed auxiliary task EC-chunk pair extraction task. To be specific, we employ Joint-EC(Joint-GCN) and Joint-ECW(Joint-GCN) to conduct experiments on ConvECPE dataset. The results are shown in Table 5, where EC is a standard Joint-EC model without EC-chunk extraction task; ECW is a standard Joint-ECW model without EC-chunk extraction task. The auxiliary task EC-chunk pair extraction is proved to be effective. Comparing EC with Joint-EC, the EC-chunk task increases the F1 score from 32.53% to 38.72%. Moreover, with the help of the auxiliary task, Joint-ECW

Model	EC-chunk extraction	Window restriction	Precision	Recall	F1
EC	-	-	**43.04**%	26.15%	32.54%(3e-3)
ECW	-	+	38.07%	29.71%	33.38%(2e-3)
Joint-EC	+	-	42.79%	35.35%	38.72%(4e-3)
Joint-ECW	+	+	38.16%	**59.27**%	**46.43**%(4e-3)

TABLE 5: Results of ablated Joint-EC models on the ConvECPE dataset. The reported results are the average of 3 runs. The numbers in brackets are the standard deviation of F1 scores.

outperforms ECW by 13.05% in terms of the F1 score. Similarly, ECW achieves a better F1 score than EC model by means of window restriction mechanism. If we combine EC-chunk extraction and window restriction mechanism, the F1 score of Joint-ECW improves by 13.89% compared with that of the EC model, which indicates a good coupling between these two parts.

Window restriction can partially alleviate the unbalanced sample issue, which in turn benefits the EC pair extraction task. We also notice that the window restriction mechanism slightly reduces the precision. One possible reason is that the number of proposed pairs increases more than correct pairs within the window scope. Specifically, the mechanism may lead the classifier to identify more potential pairs. Meanwhile, we observe significant performance gain in terms of recall as the ratio of correct pairs increases dramatically within the window area. EC-chunk pair extraction focuses on a small region rather than a specific cause utterance, which is more stable and robust compared to EC pair extraction. Therefore, EC-chunk extraction may provide informative features enhancing the EC pair extraction task. In particular, the auxiliary task may slightly increase or reduce the precision while increasing the recall by a large margin. One possible reason is that EC-chunk extraction may have a similar function to the window mechanism. In other words, the auxiliary task may determine one or several chunk regions that are highly related to the current utterance, which provides additional features enabling the EC pair filter to concentrate on the pairs within the chunk regions.

5.5 Effect of Chunk Size

In this subsection, we employ Joint-EC and Joint-ECW to further study the influence of chunk size of the auxiliary task. Chunk size reflects the span of the cause utterance region, which is of vital importance for the EC-chunk pair extraction task and the subsequent EC pair extraction task. As shown in Fig. 5, the optimal value of chunk size is 2 on both models. The value of F1 score declines slightly when chunk size changes from 2 to 6. When chunk size increases from 6 to 10, F1 score slightly fluctuates up and down. There are some practical explanations about the results in Fig. 5. In general, a small chunk could provide more informative features for the main task EC pair extraction since a small adjacent cause utterance region is closely related to the corresponding emotional utterance. The only exception is when chunk size equals 1. In this case, the auxiliary task is equivalent to the EC pair extraction task. However, it cannot provide enough complementary information for the main task. A large chunk contains a lot of irrelevant information which is useless to the main task. The performance of Joint-ECW decreases to around 35.5% which is close to the model without the auxiliary task (ECW model). The performance of Joint-EC is even slightly worse than the one without the auxiliary task (EC model). In conclusion, the experimental results coincide with the theoretical analysis of the proposed auxiliary task.

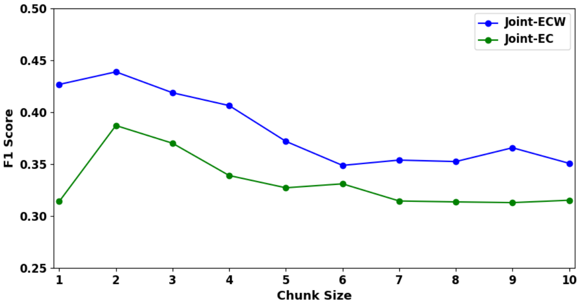


Fig. 5: The influence of chunk size of the auxiliary task

5.6 Effect of Teacher Forcing Rate

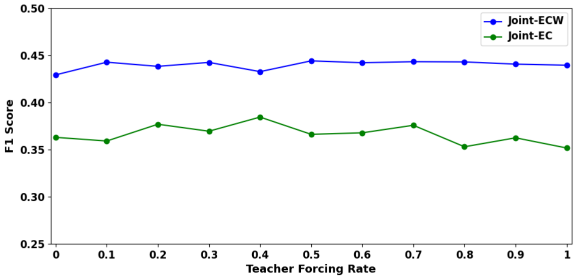
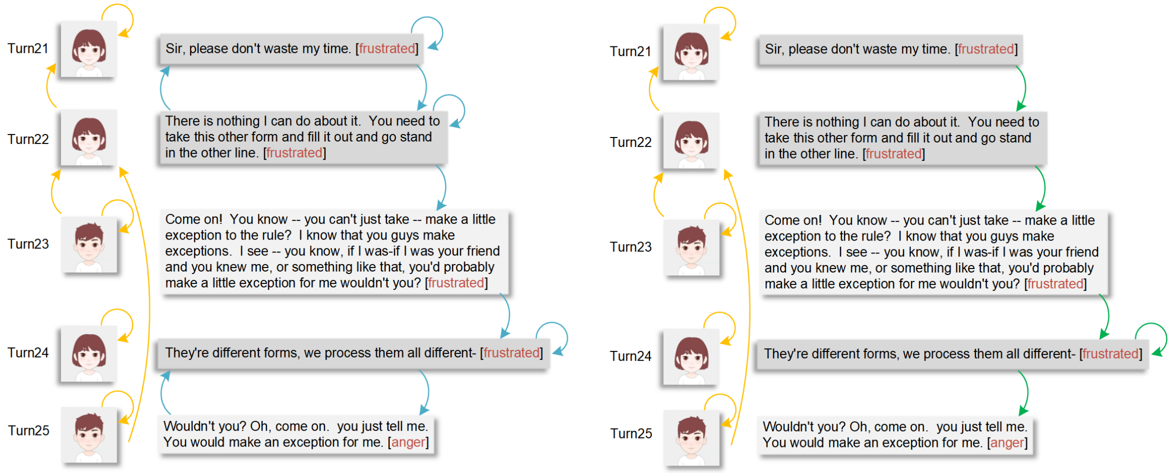


Fig. 6: The influence of teacher forcing rate on EC pair extraction

We also conduct experiments to investigate the effect of teacher forcing rate. As shown in Fig. 6, The optimal values for Joint-EC and Joint-ECW are 0.4 and 0.5, respectively. Hence, there is no one-size-fits-all optimal teacher forcing rate for different models. In general, teacher forcing mechanism has a limited influence on EC pair extraction task. One possible reason is that the predicted distribution in *step 1* is close to the distribution of the ground truth. Besides, Joint-ECW is less sensitive to the teacher forcing rate compared with Joint-EC since it is a more stable and robust filter. On the one hand, predicted labels in the training set have a very similar distribution to the predicted labels in the test set. On the other hand, there are a number of errors in the predicted labels of the training set, which may reduce the accuracy of the filter. Therefore, introducing the ground truth labels may improve the performance of the filter and alleviate overfitting to some extent.

5.7 Case Study

In Fig. 7, we illustrate the EC pair extraction result of a conversation snippet. There are five utterances in the snippet. Three are from speaker A and two are from speaker B. According to the ground truth label (the yellow arrows), there are seven EC pairs in



Sir, please don't waste my time. [frustrated]

There is nothing I can do about it. You need to take this other form and fill it out and go stand in the other line. [frustrated]

Come on! You know -- you can't just take -- make a little exception to the rule? I know that you guys make exceptions. I see -- you know, if I was-if I was your friend and you knew me, or something like that, you'd probably make a little exception for me wouldn't you? [frustrated]

They're different forms, we process them all different- [frustrated]

Wouldn't you? Oh, come on. you just tell me. You would make an exception for me. [anger]

Sir, please don't waste my time. [frustrated]

There is nothing I can do about it. You need to take this other form and fill it out and go stand in the other line. [frustrated]

Come on! You know -- you can't just take -- make a little exception to the rule? I know that you guys make exceptions. I see -- you know, if I was-if I was your friend and you knew me, or something like that, you'd probably make a little exception for me wouldn't you? [frustrated]

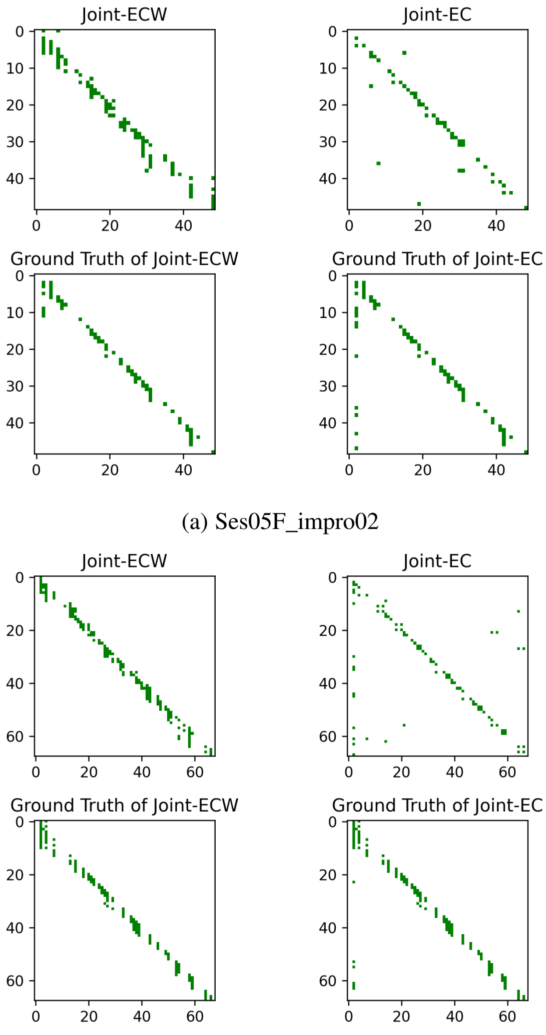
They're different forms, we process them all different- [frustrated]

Wouldn't you? Oh, come on. you just tell me. You would make an exception for me. [anger]

Fig. 7: Case study of the ECPEC task. In this figure, the EC pair extraction results of a conversation snippet is illustrated. The one on the left is the result of the Joint-ECW model while the one on the right is from the Joint-EC model. The arrow points from an emotional utterance to its corresponding cause utterance. The yellow arrows are the ground truth. The blue arrows and green arrows are predicted by Joint-ECW and Joint-EC, respectively.

this snippet, i.e., $(turn~21, turn~21)$, $(turn~22, turn~21)$, $(turn~23, turn~22)$, $(turn~23, turn~23)$, $(turn~24, turn~24)$, $(turn~25, turn~25)$, $(turn~25, turn~22)$. As mentioned before, an utterance may contain adequate information where emotion and corresponding cause are included at the same time. Take $turn$ 21 as an example, speaker A thought that speaker B wastes her time, which makes her feel frustrated. The emotion of $turn$ 25 is anger and the cause is that speaker A cannot help him and makes an exception for him.

In this snippet, the Joint-ECW model extracts nine EC pairs while the Joint-EC model extracts five EC pairs, which coincides with the performances of the two filters. The Joint-ECW has a slightly lower $Precision$ compared with Joint-EC. Hence, the Joint-ECW is inclined to extract more candidate EC pairs than the Joint-EC. Besides, it is hard for both models to extract the EC pair that has a long distance between emotion and cause utterances. Moreover, both models have a common mistake. Take utterances $turn$ 23 and $turn$ 24 as an example, the cause of $turn$ 23 cannot be $turn$ 24 since the response $turn$ 24 happens after $turn$ 23 and cannot evoke the emotion of $turn$ 23 in this case. One possible reason for such a mistake is that, even though an utterance cannot evoke emotion, it may be still partially related to the emotional utterance.



(a) Ses05F_impro02

(b) Ses05F_impro07

Fig. 8: Visualization of EC pairs in conversations

5.8 Visualization

In this subsection, we use visualization to illustrate EC pairs in conversations. In Fig. 8a and Fig. 8b, we display the EC pair extraction results given by Joint-ECW and Joint-EC models, respectively. The x- and y-axes in Fig. 8 represent the index of each utterance in a conversation. Taking Fig. 8a as an example, the figure in the left upper corner shows the extracted EC pairs from conversation $Ses05F\_impro02$, where each green point is an EC pair. The ground truth label of the Joint-ECW model is shown in the bottom-left corner. The extracted results and ground truth label of the Joint-EC model are in the right side. In general, most of the EC pairs lie around the diagonal of the figure, which is in accord with the theoretical analysis in 3.3. Meanwhile, we observe the same result as in the case study, that is, the Joint-ECW model tends to extract more candidate EC pairs than the Joint-EC model. Besides, compared with the ground truth of Joint-EC, the ground truth of Joint-ECW removes some points that are far from the diagonal. On the one hand, this indicates the effectiveness of the window restriction mechanism. On the other hand, these points have a negative impact on the Joint-EC filter and would lead to more long-range EC pairs. It is hard to extract long-range EC pairs and the accuracy of long-range EC pairs of the Joint-EC model is low according to Fig. 8a and Fig. 8b. With the help of window restriction, the Joint-ECW filter focuses more on short-range EC pairs than long-range EC pairs and thus achieves a better performance.

6 Conclusion

In this paper, we propose a new task, termed ECPEC, which aims to extract all possible EC pairs in conversations. Since there is no existing dataset for the new task, we introduce a high-quality conversational emotion-cause pair extraction dataset ConvECPE. To deal with this task, we propose a two-step framework taking the properties of conversations like context-dependence and interactivity into account. We first employ multi-task learning to combine emotion detection and cause detection into a unified model. Then, instead of directly pairing all the detected emotions and causes, we propose a multi-task model which pairs EC-chunk and EC pairs at the same time. Experimental results demonstrate the feasibility of ECPEC task and the effectiveness of our models. Furthermore, the experiments on the ConvECPE dataset enable us to have a deep understanding of this new task.

Our two-step framework is the first successful attempt on this new task and new dataset. Nevertheless, it is not the final solution for this challenging task. In future work, we plan to build a model combining the two steps into a unified framework to further improve the performance of this ECPEC task. Moreover, information about speakers should be taken into account in future research to enhance the filter. In addition, it is also feasible to perform a multi-modal EC pair extraction task through the ConvECPE dataset since it also contains visual and audio features.

Acknowledgments

This research is supported by the Agency for Science, Technology and Research (A*STAR) under its AME Programmatic Funding Scheme (Project # A18A2b0046). We would like to thank Dr. Mao and Dr. Sooji for their useful comments on this paper.

References

[1] W. Li, W. Shao, S. Ji, and E. Cambria, "Bieru: Bidirectional emotional recurrent unit for conversational sentiment analysis," *Neurocomputing*, vol. 467, pp. 73–82, 2022.

[2] R. Xia and Z. Ding, "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*. Association for Computational Linguistics (ACL), 2019, pp. 1003–1012.

[3] C. Fan, C. Yuan, J. Du, L. Gui, M. Yang, and R. Xu, "Transition-based directed graph construction for emotion-cause pair extraction," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3707–3717.

[4] Z. Ding, R. Xia, and J. Yu, "Ecpe-2d: Emotion-cause pair extraction based on joint two-dimensional representation, interaction and prediction," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3161–3170.

[5] I. Hutchby, "Conversation analysis," *The Wiley-Blackwell Encyclopedia of Social Theory*, pp. 1–9, 2017.

[6] R. Wooffitt, *Conversation analysis and discourse analysis: A comparative and critical introduction*. Sage, 2005.

[7] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, no. 4, p. 335, 2008.

[8] K. Crawford, *The Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence*. Yale University Press, 2021.

[9] Y. Li, H. Su, X. Shen, W. Li, Z. Cao, and S. Niu, "Dailydialog: A manually labelled multi-turn dialogue dataset," in *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2017, pp. 986–995.

[10] S. Y. M. Lee, Y. Chen, and C.-R. Huang, "A text-driven rule-based system for emotion cause detection," in *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, 2010, pp. 45–53.

[11] A. Neviarouskaya and M. Aono, "Extracting causes of emotions from text," in *Proceedings of the Sixth International Joint Conference on Natural Language Processing*, 2013, pp. 932–936.

[12] W. Li and H. Xu, "Text-based emotion classification using emotion cause extraction," *Expert Systems with Applications*, vol. 41, no. 4, pp. 1742–1749, 2014.

[13] K. Gao, H. Xu, and J. Wang, "Emotion cause detection for chinese micro-blogs based on ecocc model," in *Pacific-Asia Conference on Knowledge Discovery and Data Mining*. Springer, 2015, pp. 3–14.

[14] S. Yada, K. Ikeda, K. Hoashi, and K. Kageura, "A bootstrap method for automatic rule acquisition on emotion cause extraction," in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. IEEE, 2017, pp. 414–421.

[15] Y. Chen, S. Y. M. Lee, S. Li, and C.-R. Huang, "Emotion cause detection with linguistic constructions," in *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, 2010, pp. 179–187.

[16] X. Cheng, Y. Chen, B. Cheng, S. Li, and G. Zhou, "An emotion cause corpus for chinese microblogs with multiple-user structures," *ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP)*, vol. 17, no. 1, pp. 1–19, 2017.

[17] S. Song and Y. Meng, "Detecting concept-level emotion cause in microblogging," in *Proceedings of the 24th International Conference on World Wide Web*, 2015, pp. 119–120.

[18] I. Russo, T. Caselli, F. Rubino, E. Boldrini, P. Martínez-Barco *et al.*, "Emocause: an easy-adaptable approach to emotion cause contexts," in *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis-ACL*. Association for Computational Linguistics (ACL), 2011, pp. 153–160.

[19] L. Gui, L. Yuan, R. Xu, B. Liu, Q. Lu, and Y. Zhou, "Emotion cause detection with linguistic construction in chinese weibo text," in *Natural Language Processing and Chinese Computing*. Springer, 2014, pp. 457–464.

[20] E. Cambria, Q. Liu, S. Decherchi, F. Xing, and K. Kwok, "Senticnet 7: a commonsense-based neurosymbolic ai framework for explainable sentiment analysis," *Proceedings of LREC*, pp. 3829–3839, 2022.

[21] L. Gui, J. Hu, Y. He, R. Xu, Q. Lu, and J. Du, "A question answering approach to emotion cause extraction," in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics (ACL), 2017, pp. 1593–1602.

[22] Y. Kim, "Convolutional neural networks for sentence classification," in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics (ACL), 2014, pp. 1746–1751.

[23] 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 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 4752–4757.

[24] I. Maks and P. Vossen, "A lexicon model for deep sentiment analysis and opinion mining applications," *Decision Support Systems*, vol. 53, no. 4, pp. 680–688, 2012.

[25] K. Denecke and Y. Deng, "Sentiment analysis in medical settings: New opportunities and challenges," *Artificial intelligence in medicine*, vol. 64, no. 1, pp. 17–27, 2015.

[26] Y. Li, S. Wang, Q. Pan, H. Peng, T. Yang, and E. Cambria, "Learning binary codes with neural collaborative filtering for efficient recommendation systems," *Knowledge-Based Systems*, vol. 172, pp. 64–75, 2019.

[27] S. Rani and P. Kumar, "A sentiment analysis system to improve teaching and learning," *Computer*, vol. 50, no. 5, pp. 36–43, 2017.

[28] S. Poria, E. Cambria, D. Hazarika, N. Majumder, A. Zadeh, and L.-P. Morency, "Context-dependent sentiment analysis in user-generated

videos," in *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*, 2017, pp. 873–883.

[29] D. Hazarika, S. Poria, R. Mihalcea, E. Cambria, and R. Zimmermann, "Icon: Interactive conversational memory network for multimodal emotion detection," in *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2018, pp. 2594–2604.

[30] N. Majumder, S. Poria, D. Hazarika, R. Mihalcea, A. Gelbukh, and E. Cambria, "Dialoguernn: An attentive rnn for emotion detection in conversations," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33, 2019, pp. 6818–6825.

[31] S. Poria, N. Majumder, D. Hazarika, D. Ghosal, R. Bhardwaj, S. Y. B. Jian, P. Hong, R. Ghosh, A. Roy, N. Chhaya *et al.*, "Recognizing emotion cause in conversations," *Cognitive Computation*, vol. 13, no. 5, pp. 1317–1332, 2021.

[32] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, 2014, pp. 1532–1543.

[33] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural computation*, vol. 9, no. 8, pp. 1735–1780, 1997.

[34] I. Misra, A. Shrivastava, A. Gupta, and M. Hebert, "Cross-stitch networks for multi-task learning," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016, pp. 3994–4003.

[35] Z. Chen and T. Qian, "Relation-aware collaborative learning for unified aspect-based sentiment analysis," in *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 3685–3694.

[36] R. Mao and X. Li, "Bridging towers of multi-task learning with a gating mechanism for aspect-based sentiment analysis and sequential metaphor identification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 15, 2021, pp. 13 534–13 542.

[37] J. Li, M. Zhang, D. Ji, and Y. Liu, "Multi-task learning with auxiliary speaker identification for conversational emotion recognition," *arXiv e-prints*, pp. arXiv–2003, 2020.

[38] M. Schlichtkrull, T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling, "Modeling relational data with graph convolutional networks," in *European semantic web conference*. Springer, 2018, pp. 593–607.

[39] L. Yao, C. Mao, and Y. Luo, "Graph convolutional networks for text classification," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 33(01), 2019, pp. 7370–7377.

[40] D. Ghosal, N. Majumder, S. Poria, N. Chhaya, and A. Gelbukh, "Dialoguegcn: A graph convolutional neural network for emotion recognition in conversation," in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, 2019.

[41] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances in neural information processing systems*, 2017, pp. 5998–6008.

[42] R. J. Williams and D. Zipser, "A learning algorithm for continually running fully recurrent neural networks," *Neural computation*, vol. 1, no. 2, pp. 270–280, 1989.

[43] S. Ruder, "An overview of multi-task learning in deep neural networks," *arXiv preprint arXiv:1706.05098*, 2017.

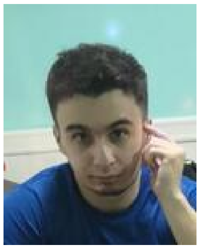
[44] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[45] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *Computer Science*, 2014.

[46] Y. Zhang, P. Tiwari, D. Song, X. Mao, P. Wang, X. Li, and H. M. Pandey, "Learning interaction dynamics with an interactive lstm for conversational sentiment analysis," *Neural Networks*, vol. 133, pp. 40–56, 2021.

Wei Li received his Bachelor of Industrial Engineering from Shanghai Jiao Tong University in 2015. After that, he obtained his Master of Management Science and Engineering from University of Chinese Academy of Sciences in 2018. He enrolled as a PhD student under the supervision of Erik Cambria at NTU SCSE in 2019. His main research interests include sentiment analysis, natural language processing and deep learning. In particular, he is working on the construction of advanced deep learning models for sentiment analysis.

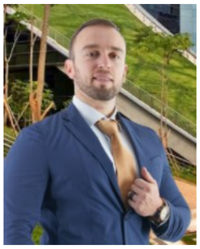
Yang Li is an associate professor with the school of automation at Northwestern Polytechnical University. After receiving his bachelor's and doctoral degrees from Northwestern in 2014 and 2018 respectively, he worked as a research fellow in SenticTeam under Professor Erik Cambria at Nanyang Technological University in Singapore. His research goal is to build a trustworthy AI system in the real application, and his research interests are in Natural Language Processing, Machine Learning, Recommender System, Explainable Artificial Intelligence, etc. He has published several papers on these topics at international conferences and peer-reviewed journals.



Vlad Pandelea received his Bachelor and Master of Science in Computer Science from the University of Pisa in 2017 and 2019, respectively. Since 2020, he is a PhD student at NTU SCSE under the supervision of Erik Cambria. His research interest is in dialogue systems and sentiment analysis. Currently, he is working on methods to efficiently deploy competitive solutions on resource-constrained devices. He is interested in both deep learning techniques that focus on improving performance and on more efficient methods such as randomized neural networks that aim to reduce computational requirements.

Mengshi Ge received her Bachelor of Science in Mathematics and Applied Mathematics from Dalian University of Technology (China) in 2017. She later moved on to get a Master of Science in Statistics at George Washington University (USA) in 2019. She is currently a part-time PhD student at Nanyang Technological University, School of Computer Science and Engineering under the tutelage of Prof Erik Cambria. Her research interests include natural language processing, metaphor understanding, and gender bias in text.

Luyao Zhu received her Bachelor of Science in Mathematics and Applied Mathematics and Bachelor of Science in Economics from China University of Political Science and Law, in 2016. Subsequently, she went on to pursue a Master of Science in Management Science and Engineering at University of Chinese Academy of Sciences, in 2019. After joining the Sentic Team as an intern, she is now a PhD student under the supervision of Erik Cambria at NTU SCSE. Her research interests include natural language understanding and deep reinforcement learning, which she applies to the development of persona-specific dialogue systems.



Erik Cambria is currently an Associate Professor with Nanyang Technological University, Singapore. His main research interests are AI and affective computing. He received the Ph.D. degree in computing science and mathematics through a joint programme between the University of Stirling, Stirling, U.K., and MIT Media Lab, Cambridge, MA, USA. He is the corresponding author of this article. Contact him at cambria@ntu.edu.sg.