CauAIN: Causal Aware Interaction Network for Emotion Recognition in Conversations

Weixiang Zhao , **Yanyan Zhao*** , **Xin Lu**
Harbin Institute of Technology, China
{wxzhao, yyzhao, xlu}@ir.hit.edu.cn

Abstract

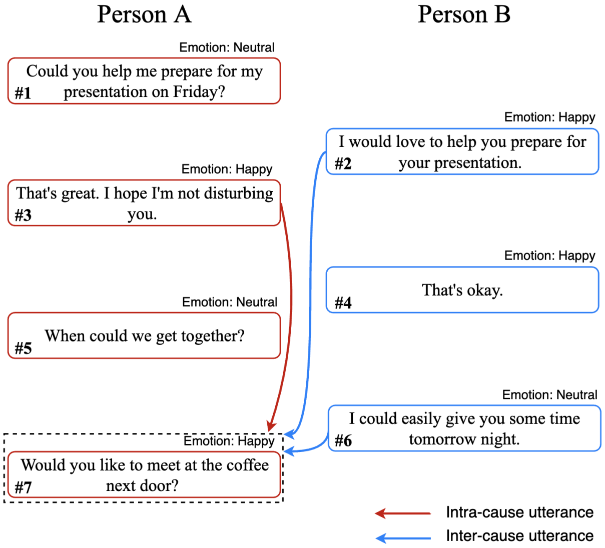
Emotion Recognition in Conversations has attained increasing interest in the natural language processing community. Many neural-network based approaches endeavor to solve the challenge of emotional dynamics in conversations and gain appealing results. However, these works are limited in capturing deep emotional clues in conversational context because they ignore the emotion cause that could be viewed as stimulus to the target emotion. In this work, we propose Causal Aware Interaction Network (CauAIN) to thoroughly understand the conversational context with the help of emotion cause detection. Specifically, we retrieve causal clues provided by commonsense knowledge to guide the process of causal utterance traceback. Both retrieve and traceback steps are performed from the perspective of intra- and inter-speaker interaction simultaneously. Experimental results on three benchmark datasets show that our model achieves better performance over most baseline models.

1 Introduction

Emotion recognition in conversations (ERC) aims at predicting the emotion for each utterance in conversations. Due to its key role to achieve empathetic systems and wide range of applications in opinion mining, social media analysis, health care and other areas, ERC has received increasing attention in the natural language processing (NLP) community.

The key challenge in ERC is posed by emotional dynamics [Poria *et al.*, 2019a], which refers to emotional influences during the interaction between speakers. Early studies have been devoted to cope with this challenge by modeling intra- and inter-speaker dependency with recurrent neural networks (RNN) [Hazarika *et al.*, 2018; Majumder *et al.*, 2019; Ghosal *et al.*, 2020] and graph neural networks (GNN) [Zhang *et al.*, 2019; Ghosal *et al.*, 2019; Ishiwatari *et al.*, 2020; Li *et al.*, 2021].

However, such attempts for intra- and inter-speaker dependency modeling are limited in capturing deeper and richer clues of emotional dynamics, for the reason that the **emotion cause**, which exactly triggers the target emotion, is ignored. We believe that the model can better understand human emotions if it has the ability to associate the emotion cause with the target utterance.



#1 Could you help me prepare for my presentation on Friday?

#2 I would love to help you prepare for your presentation.

#3 That's great. I hope I'm not disturbing you.

#4 That's okay.

#5 When could we get together?

#6 I could easily give you some time tomorrow night.

#7 Would you like to meet at the coffee next door?

Figure 1: An example for intra- and inter-cause utterances triggering the emotion of the target utterance.

For causal utterance that expresses reasons for the speaker to feel such an emotion, we categorize them as either **intra-cause utterance** or **inter-cause utterance**. The former one refers to those appearing in the speaker's own dialogue turns, which means the cause of the emotion is due to a stable mood of the speaker that inherited from previous emotional states, while inter-cause utterance is present in the other speaker's turns and the emotion of the target speaker is influenced by the other speaker. An example is shown in Figure 1 to illustrate emotion recognition with the help of identifying intra- and inter-cause utterances in conversations. The situation is that PersonA seeks PersonB for help about the presentation and we choose #7 as the target utterance to be classified. It is obvious that utterance #7 does not contain any explicit emotion indicator words and almost sound neutral on the surface. Thus, clues for its carrying emotion *happy* could only

*Corresponding author

be inferred from the context. And #2 and #6 are inter-cause utterances that contribute significantly to the evoked emotion *happy* of utterance #7, where PersonB is willing to do PersonA a favor and promise a date tomorrow night. Also, parts of PersonA's *happy* are from the continuous emotional states conveyed by intra-cause utterance #3.

In this paper, to extract richer clues for emotional dynamics, we propose Causal Aware Interaction Network (CauAIN) to explicitly model intra- and inter-speaker dependency from a perspective of emotion cause detection. Since there are no ERC datasets with emotion cause information annotated, commonsense knowledge is leveraged as casual clues for emotion cause detection in conversations. To be more specific, six *if-then* relation types from ATOMIC (The Atlas of Machine Commonsense) [Sap *et al.*, 2019] are considered according to their causal relations. $xReact$, $xEffect$ and $xWant$ are viewed as the intra-cause clues, while $oReact$, $oEffect$ and $oWant$ are explained for inter-cause clues. In order to identify causal utterances of a certain emotion expressed by the target utterance, we devise a two-step causal aware interaction consisting of causal clue retrieval and causal utterance traceback. To begin with, different intra- and inter-cause clues are evaluated in terms of effects they have on the target utterance. Then results of the previous step could be viewed as gate values for intra- and inter-cause utterance traceback. In the final classification layer, causal aware representations are mixed for emotion recognition.

To evaluate the performance of the proposed model, we conduct extensive experiments on three datasets: IEMOCAP, DailyDialog and MELD. State-of-the-art performance is achieved on all of the three datasets. Further, we also demonstrate the exact emotion cause discovered by our proposed model.

The main contributions of this work are summarized as follows:

- We propose a novel model CauAIN to fully understand the conversational context with intra- and inter-cause utterance detection. It is the first attempt to explore the emotion cause for emotion recognition in conversations.
- We devise a simple but effective method of automatically detecting causal utterances with commonsense knowledge as causal clues, which does not depend on any annotation of emotion cause.
- Results of extensive experiments on three benchmark datasets demonstrate the effectiveness of the proposed model and our method of detecting exact emotion cause.

2 Methodology

First, we define the problem of the ERC task. Given a conversation with N consecutive utterances $\{u_1, u_2, \cdots, u_N\}$ and M speakers $\{s_1, s_2, \cdots, s_M\}$, the goal of this task is to predict the emotion label e_i of each utterance u_i spoken by s_i. The architecture of our proposed model CauAIN is shown in Figure 3, which consists of four parts: Causal Clue Acquisition, Casual Clue Retrieval, Causal Utterance Traceback and Emotion Recognition. We will elaborate each one of them in the rest of this section.

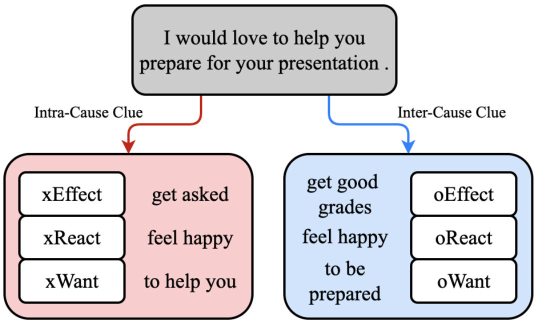


Figure 2: An example of six types of intra- and inter-cause clues.

2.1 Causal Clue Acquisition

Intra- and Inter-Cause Clue Representation

Due to the limitation of annotated emotion cause in existing ERC datasets, we turn to ATOMIC for causal clues offering.

What is ATOMIC? ATOMIC is an atlas of everyday commonsense reasoning and organized through textual descriptions of inferential knowledge, where nine *if-then* relation types are proposed to distinguish causes vs. effects, agents vs. themes, voluntary vs. involuntary events, and actions vs. mental states.

Why we choose ATOMIC? Previous works [Sap *et al.*, 2019; Turcan *et al.*, 2021] demonstrate that neural networks are capable of anticipating the likely causes and effects of previously unseen events with the help of rich inferential knowledge provided by ATOMIC. According to this, we extend such unseen events under the circumstance of dialogue and explore six relation types which are all categorized as "effects" based on their casual relations. To be more specific, *xReact*, *xEffect* and *xWant* provide the intra-cause clues that represent influences or results generated by utterances from the speaker him/herself. In addition, *oReact*, *oEffect* and *oWant* mean what effects exert on others or what others would like to do and feel after receiving the current utterance. Thus, rich inter-cause clues could be revealed if we take such three relation types into consideration. Figure 2 illustrates causal clues corresponding to the aforementioned six relation types.

How to acquire causal clue from ATOMIC? To acquire intra- and inter-cause clues contained in ATOMIC, we adopt the generative commonsense transformer model COMET [Bosselut *et al.*, 2019] which is trained on ATOMIC. Given the input event (which is referred as an utterance u_i under the circumstance of dialogue) and the selected relation type, COMET would generate descriptions of "then" with the format of *if-then* reasoning. For example, taking u_i and the relation type *oReact* as inputs, a reasoning sequence "If u_i, then others would feel" could be derived from COMET. We concatenate u_i and a relation with mask tokens such as (u_i $[MASK]$ $oReact$) to construct the input of COMET. Following [Ghosal *et al.*, 2020], the hidden state representation from the last encoder layer of COMET are taken as causal clue. Thus, for each u_i in this work, three pieces of intra-cause clue generated from COMET are concatenated and mapped to the dimension of $2d_h$ with a linear unit. So do the other three pieces of inter-cause clue. We denote them

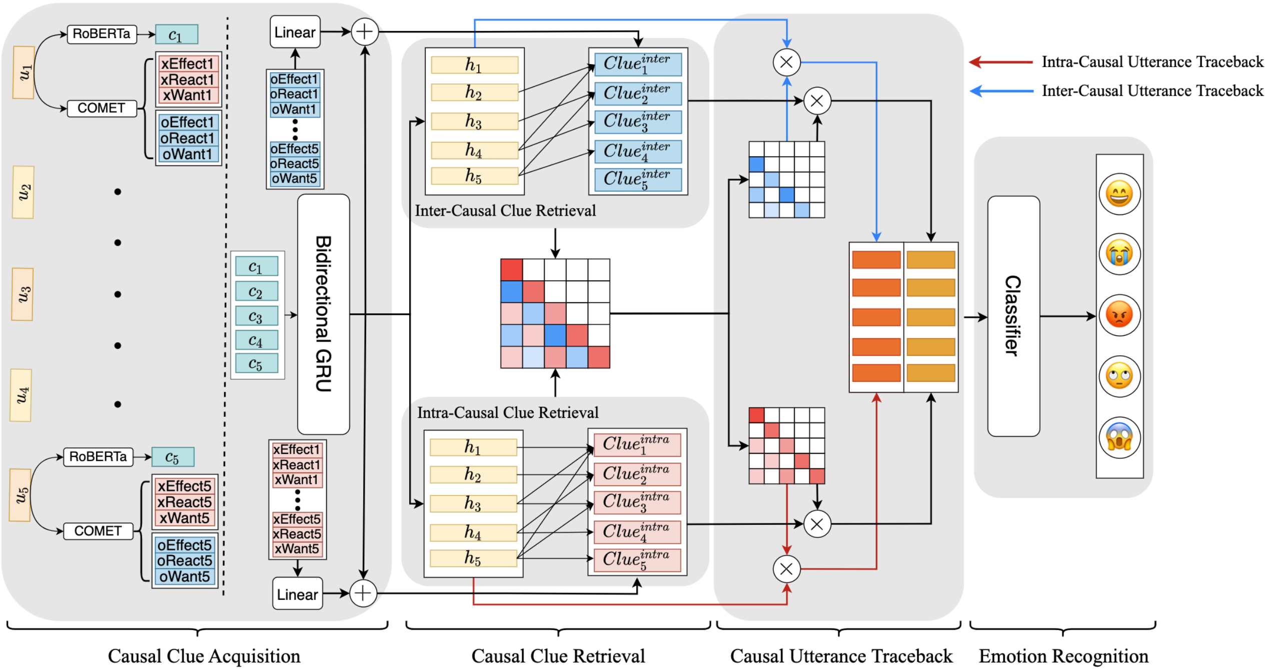


Figure 3: The overall architecture of our proposed model.

with $Clue_i^{intra} \in \mathbb{R}^{2d_h}$ and $Clue_i^{inter} \in \mathbb{R}^{2d_h}$.

Utterance-level Representation

We employ the widely-used pretrained model RoBERTa [Liu *et al.*, 2019] to extract utterance-level feature vectors. Specifically, for each utterance $u_i = \{w_1, w_2, \cdots, w_L\}$, we concatenate a special token $[CLS]$ to the beginning of the utterance. Then the sequence $\{[CLS], w_1, w_2, \cdots, w_L\}$ are fed to fine-tune the pretrained RoBERTa model with an utterance-level emotion classification task and the $[CLS]$ token from the last layer is passed through a pooling layer to classify it into its emotion class.

After the process of fine-tuning, to derive the utterance-level feature vector c_i corresponding to the $[CLS]$ token, we pass each utterance in the same input format as $\{[CLS], w_1, w_2, \cdots, w_L\}$:

$$c_i = RoBERTa(\,[CLS], w_1, w_2, \cdots, w_L\,) \tag{1}$$

where $c_i \in \mathbb{R}^{d_m}$ and d_m is the dimension of hidden states of tokens in RoBERTa. Following [Ghosal *et al.*, 2020], $[CLS]$ tokens from final four layers are averaged to obtain the utterance-level feature vector for each utterance.

Conversational Representation

Under the circumstance of conversational setting, the emotion of an utterance usually depends on the context of the whole conversation. Thus, based on utterance-level features c_i, we apply a bi-directional Gated Recurrent Unit (GRU) to model sequential dependencies between adjacent utterances and the conversational representation h_i can be computed as:

$$h_i = \overleftrightarrow{GRU}(\,c_i,\, h_{i-1}\,) \tag{2}$$

where $h_i \in \mathbb{R}^{2d_h}$ represents the hidden state vector at time step i and d_h is the dimension of a GRU cell output.

2.2 Causal Aware Interaction

To attain richer clues for emotional dynamics in conversations and explicitly interact intra- and inter-speaker dependencies, we devise the two-step causal aware interaction, including causal clue retrieval and causal utterance traceback, to enrich context representation with emotion cause.

Causal Clue Retrieval

To explore how much the emotion cause of the target utterance depends on intra- or inter-cause utterances, we should retrieve intra- and inter-cause clue and assign weighted scores to them. For intra-cause clue retrieval, we focus on influences or effects from the same speaker and the retrieval score could be computed as:

$$scores_{i,j}^{intra} = \frac{[f_q(h_i)(f_k(h_j) + f_e(Clue_j^{intra}))]mask_{i,j}^{intra}}{\sqrt{d_h}} \tag{3}$$

where $f_q(x)$; $f_k(x)$; $f_e(x)$ are all linear transformations. $mask_{i,j}^{intra}$ makes sure the target utterance h_i retrieve utterances from the same speaker as it to perform intra-cause clue retrieval. It is also worth noting that $mask_{i,j}^{intra}$ guarantee the correct temporal order of retrieval process, which meets the nature of causality that cause could not be found according to causal clues from the future.

$$mask_{i,j}^{intra} = \begin{cases} 1, & if\ j <= i\ and\ \phi(h_i) = \phi(h_j) \\ 0, & otherwise \end{cases} \tag{4}$$

where ϕ maps the index of the utterance into that of the corresponding speaker.

The process of inter-cause clue retrieval pay attention to clues contained in utterances from other speakers.

$$scores_{i,j}^{inter} = \frac{[f_q(h_i)(f_k(h_j) + f_e(Clue_j^{inter}))]mask_{i,j}^{inter}}{\sqrt{d_h}} \tag{5}$$

$$mask_{i,j}^{inter} = \begin{cases} 1, & if\ j < i\ and\ \phi(h_i) \neq \phi(h_j) \\ 0, & otherwise \end{cases} \tag{6}$$

Once retrieval scores from intra- and inter-cause clue are obtained, we should consider them comprehensively on the same scale. The joint value that controls how much information should be gathered from intra- or inter-cause utterance can be computed by:

$$\alpha_{i,j}^{joint} = softmax(scores_{i,j}^{intra} + scores_{i,j}^{inter}) \tag{7}$$

Causal Utterance Traceback

In the step of causal utterance traceback, the model could be aware of different weights to focus more on utterances related to the emotion cause according to results derived from causal clue retrieval. The joint value is decomposed into two parts from the turns of intra- and inter-speaker according to Eq. 4 and Eq. 6, which is denoted as α^{intra} and α^{inter}.

Then causal-aware context representations incorporated with intra-cause utterance and inter-cause utterance could be obtained by:

$$\tilde{h_i} = \sum_{j \in S(i)} \alpha_{i,j}^{intra} f_q(h_j) + \sum_{j \in O(i)} \alpha_{i,j}^{inter} f_q(h_j) \tag{8}$$

where $S(i)$ is the set of utterances with the same speaker as utterance u_i and $O(i)$ stands for the set of utterances with the speaker different from utterance u_i. Further, emotional information included in the causal clue should also be taken into consideration:

$$\tilde{c_i} = \sum_{j \in S(i)} \alpha_{i,j}^{intra} C_j^{intra} + \sum_{j \in O(i)} \alpha_{i,j}^{inter} C_j^{inter} \tag{9}$$

$$C_j^{intra} = f_k(h_j) + f_e(Clue_j^{intra}) \tag{10}$$

$$C_j^{inter} = f_k(h_j) + f_e(Clue_j^{inter}) \tag{11}$$

where $f_k(x)$ is a linear transformation. And final causal-aware representation is concatenated by:

$$h_i^f = \tilde{h_i} \oplus \tilde{c_i} \tag{12}$$

2.3 Emotion Recognition

Finally, taking the above causal-aware representation as input, an emotion classifier is applied to predict the emotion of the utterance.

$$\hat{y} = softmax(W_e h^f + b_e) \tag{13}$$

where W_e and b_e are trainable parameters.

Cross entropy loss is utilized to train the model and the loss function is defined as:

$$L = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{E} \hat{y_i}^j \cdot log(y_i^j) \tag{14}$$

where E is the number of emotion class and y_i^j stands for the ground-truth emotion label of the utterance i.

Dataset	Dialogues			Utterances		
	Train	Val	Test	Train	Val	Test
IEMOCAP	120		31	5,810		1,623
DailyDialog	11,118	1,000	1,000	87,170	8,069	7,740
MELD	1,039	114	280	9,989	1,109	2,610

Table 1: Dataset statistics

3 Experiments

3.1 Dataset

We conduct experiments on three benchmark datasets from IEMOCAP [Busso *et al.*, 2008], DailyDialog [Li *et al.*, 2017] and MELD [Poria *et al.*, 2019b]. Statistics of the three datasets are shown in Table 1.

IEMOCAP is a dyadic conversation dataset between ten speakers. Each utterance is annotated with one of the following six emotion labels: *happy*, *sad*, *neutral*, *angry*, *excited* and *frustrated*.

DailyDialog contains two-way dialogues covering topics about the daily life. There are seven emotion labels in this dataset: *anger*, *disgust*, *fear*, *joy*, *neutral*, *sadness* and *surprise*. The dataset has over 83% *neutral* labels.

MELD is a multimodal dataset with multi-speaker conversations. It is collected from the TV show *Friends* and the emotion classes belong to *anger*, *disgust*, *fear*, *joy*, *neutral*, *sadness* and *surprise*.

3.2 Baselines and Comparison Models

We compare our proposed model with the following methods:

ICON [Hazarika *et al.*, 2018] adopts two GRUs to attain utterance representations between two parties. Then a global GRU is utilized for inter-speaker dependency modeling.

DialogueRNN [Majumder *et al.*, 2019] devises three states including global state, party state and emotion state to model intra- and inter-speaker dependencies with GRUs.

DialogueGCN [Ghosal *et al.*, 2019] uses graph convolutional network to model intra- and inter-speaker dependency over the graph structure.

IEIN [Lu *et al.*, 2020] argues that emotions of the utterance are interactive and utilizes the predicted emotion labels to explicitly guide the interaction of emotional dynamics among utterances.

RGAT [Ishiwatari *et al.*, 2020] enhances relation modeling ability of relation-aware graph attention network with the proposed relational position encoding.

COSMIC [Ghosal *et al.*, 2020] models more refined speaker states and utilizes commonsense knowledge to understand emotional dynamics better.

DialogXL [Shen *et al.*, 2021] devises four different types of attention to make the model aware of intra- and inter-speaker dependency.

DialogueCRN [Hu *et al.*, 2021] proposes to fully understand the conversational context from a cognitive perspective and designs reasoning modules to integrate emotional clues.

SKAIG [Li *et al.*, 2021] constructs a novel graph to explore psychological states of speakers and graph transformer is used to propagate the interactive information over the graph.

Model	**IEMOCAP**	**DailyDialog**		**MELD**
	weighted-F1	micro-F1	macro-F1	weighted-F1
ICON	58.54	-	-	-
DialogueRNN	62.57	55.95	41.8	57.03
DialogueGCN	64.18	-	-	58.1
IEIN	64.37	-	-	60.72
DialogueCRN	66.2	-	-	58.39
RGAT	65.22	54.31	-	60.91
COSMIC	65.28	58.48	51.05	65.21
DialogXL	65.94	54.93	-	62.41
KI-Net	66.98	57.3	-	63.24
SKAIG	66.96	**59.75**	51.95	65.18
CauAIN (Ours)	**67.61**	58.21	**53.85**	**65.46**
w/o Inter Cause	64.61	54.23	49.53	62.83
w/o Intra Cause	64.66	55.24	48.7	59.52
w/o Causal Clue	63.77	57.2	51.73	65.2

Table 2: Comparison of our model against state-of-the-art baselines. Intra Cause and Inter Cause are the process of intra- and inter cause detection, respectively and Causal Clue refers to causal clue generated from COMET.

KI-Net [Xie *et al.*, 2021] concentrates on direct utterance-knowledge interaction and involves additional affective information with an auxiliary task.

3.3 Implementation Details

For utterance-level feature extraction, we fine-tune RoBERTa Large model for a batch size of 32 and Adam optimizer is adopted with learning rate of 1e-5. Thus, the dimension of utterance-level feature vector d_m is 1024. For all representations in the following parts of CauAIN, d_h is set to 300. We train CauAIN with Adam optimizer in a learning rate of 1e-4.

The weighted F1 score is selected as the evaluating metric for IEMOCAP and MELD. Following previous works, we report the micro F1 score excluding utterances annotated with *neutral* and macro F1 score for DailyDialog. All of our results are averaged on 5 runs.

4 Results and Analysis

4.1 Overall Results

Illustrated in Table 2, our proposed model achieves state-of-the-art results on all three datasets.

IEMOCAP and DailyDialog. IEMOCAP contains rich information of conversational context with the average conversation length up to 50 turns and the phenomena of emotional dynamics is more frequently observed in this dataset. Benefiting from emotion cause detection to extract deep and rich clues for emotional dynamics, CauAIN achieves new state-of-the-art scores of 67.61 on IEMOCAP. Even if some similar types of external knowledge from COMET is used, we think the reason why CauAIN performs better than COSMIC is that the step-by-step sequential interaction between utterances and lack of considering emotion cause hinder the understanding of rich emotional clues contained in the conversational context. In addition, the improvement of performance over SKAIG demonstrates that emotion cause could provide more direct and richer emotional clues for emotion recognition than psychological information does. Also, our model achieve the best performance on DailyDialog in terms of the macro-F1 score, which demonstrates CauAIN could partly alleviate the influence of data imbalance.

MELD. We observe that CauAIN gains slight improvement over recent baseline models on MELD. The reason may be that MELD is a multiple-speaker dataset with short conversations. Thus, utterances that are not spoken to the current speaker would bring noise and useless information for emotion cause detection. It reminds us to further incorporate the discourse structure of conversations and explicitly model the conversation threads in the setting of multi-party conversations.

4.2 Ablation Study

We conduct ablation studies to verify the effectiveness of different components proposed in CauAIN. Results are shown in Table 2.

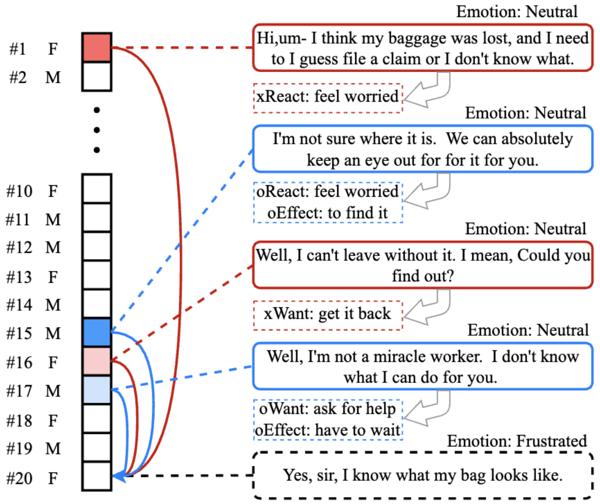
Effect of Emotion Cause

To investigate the impact of emotion cause for emotion recognition, we remove either the intra- and inter-cause related parts in the model. Specifically, causal clue from ATOMIC is discarded and we do not perform the two-step causal aware interaction. The performance has a certain degree of decline on all three datasets, which is displayed in the second-to-last row and the third-to-last row in Table 2. This suggests that both intra- and inter cause play a significant role in offering rich emotional clues for emotion recognition. Also, it manifests the effectiveness of explicit and thorough intra- and inter-speaker dependency modeling with the aid of intra- and inter cause detection. Besides, it is worth mentioning that the margin of dropped results perform a notable difference on MELD, where only considering inter cause would result in attending more useless contextual information that confuses the model. It is in accordance with the analysis in section 4.1 that additional information of conversation thread should be considered for multi-party conversations.

Effect of Causal Clue

To verify the effectiveness of the generated causal clue, we first analyze its impact on emotion cause detection. Since there is no emotion cause label annotated on IEMOCAP, we randomly select 100 utterances from the test set for human evaluation. Utterances corresponding to top three weighted values from the output of causal clue retrieval α^{joint} are chosen as the candidate of emotion cause. And we ask human annotators to judge whether true casual utterances belong to the candidate set. Accuracy is adopted as the evaluation metrics. With the aid of causal clue, the emotion cause related modules of our proposed model finally achieve an accuracy of 60%. When discarding the generated causal clue, the process of causal aware interaction degrades to the plain interaction between conversational utterances and the accuracy of emotion cause detection is reduced to 53%. Although the performance of modules related to emotion cause are not significant enough, the introduction of causal clue does help detect exact causal utterances in conversations and the whole process does not rely on any emotion cause annotation.

Further, the impact of causal clue on emotion recognition is shown in the last row in Table 2. The model without causal



Emotion: Neutral
Hi,um- I think my baggage was lost, and I need to I guess file a claim or I don't know what.

xReact: feel worried

Emotion: Neutral
I'm not sure where it is. We can absolutely keep an eye out for for it for you.

oReact: feel worried
oEffect: to find it

Emotion: Neutral
Well, I can't leave without it. I mean, Could you find out?

xWant: get it back

Emotion: Neutral
Well, I'm not a miracle worker. I don't know what I can do for you.

oWant: ask for help
oEffect: have to wait

Emotion: Frustrated
Yes, sir, I know what my bag looks like.

Figure 4: A case that our model gives the correct prediction. The most two relevant intra- and inter-cause utterances are illustrated through the process of Causal Utterance Traceback.

clue means that we do not introduce the generated causal clue from COMET and utterance interaction is just performed between pure context. From the previous analysis, it could be observed that the introduced causal clue help the model detect more accurate emotion cause. And the dropped results on ERC datasets not only demonstrates that the performance of emotion recognition is benefit from incorporating exact emotion cause, but also shows the effectiveness of enriching context representation with latent emotional information contained in the causal clue. And we believe that both emotion recognition and emotion cause detection would perform better if the golden label of emotion cause is available.

4.3 Case Study

In Figure 4, we exemplify a case from the test set of IEMOCAP to demonstrate the deep and rich emotional information brought by casual utterances when it comes to detect the emotion of the target utterance. The situation is that the woman lose her baggage in the airport and ask an official staff for help. It could be easily observed that no direct emotion descriptor words are contained in the target utterance #20. Thus, its carrying emotion *Frustrated* should be inferred from the conversational contexts. By performing intra-cause clue retrieval and intra-cause utterance traceback, utterances #1 and #16 are located as reasonable intra-cause utterances, where utterance #1 sets the initial emotional state of worry and #16 emphasizes the importance of baggage for the woman. Meanwhile, the aggravation of the woman's frustration is caused by inter-cause utterances #15 and #17 that the man has no idea about the lost baggage and could not offer effective help for the woman. These exact causal utterances and latent emotional information contained in causal clue provide our model with deeper and richer clues for right predictions.

5 Related Work

Emotion Recognition in Conversations. Recent works on ERC devote to model intra- and inter-speaker dependency to cope with emotional dynamics. We summarize them into two categories according to whether external resources such as commonsense knowledge or auxiliary tasks are incorporated.

For those without introducing external resources, various types of deep neural network are leveraged for context modeling. 1) **RNN**. Majumder *et al.* [2019] consider global state, party state and emotion state of speakers and utilize three GRUs to model intra- and inter-speaker dependency. Hu *et al.* [2021] explore cognitive factors of speakers and devise a cognitive reasoning module to iteratively capture emotional clues contained in the context. 2) **GNN**. Ghosal *et al.* [2019] model the interaction among speakers with information of nodes and edges propagating over a graph. Further, Ishiwatari *et al.* [2020] refine types of edges considered in the graph and propose relational position encoding to enhance relation-aware graph attention network.

More recently, many works turn to external resources for additional knowledge to guide contextual modeling. Ghosal *et al.* [2020] explore commonsense knowledge to better understand aspects of conversation such as personality, events, mental states and intents. Based on this, Li *et al.* [2021] utilize external knowledge to model psychological interactions between utterances. Besides, an auxiliary task named sentiment polarity intensity prediction is introduced to involve more affective information directly [Xie *et al.*, 2021].

Emotion Cause in Conversations. Poria *et al.* [2021] define two sub-tasks of recognizing emotion cause in conversations and provide a corresponding dialogue-level dataset named RECCON.

6 Conclusion

In this paper, in order to capture deeper and richer clues for emotion dynamics and explicitly model intra- and inter-speaker dependency , we propose novel Causal Aware Interaction Network (CauAIN) for emotion recognition in conversations. To be more specific, we explore the effectiveness of incorporating emotion cause when recognizing the emotion of the target utterance. Commonsense knowledge is leveraged as causal clues to help automatically extract causal utterances and alleviate the limitation brought by lack of emotion cause annotation. Then two-step causal aware interaction including causal clue retrieval and causal utterance traceback is devised to detect the intra and inter emotion cause corresponding to the target utterance. And causal aware contextual representation is obtained for the emotion recognition. Experimental results on three benchmark datasets demonstrate the effectiveness of proposed CauAIN and its ability to detect exact emotion cause.

For future work, we would like to improve the performance of causal aware model under the circumstance of multi-party conversations because useless information from multi speakers is not effectively filtered in our current method. Moreover, conversation resources with emotion and emotion cause annotated would be explored to jointly perform recognition of emotion and emotion cause at the same time.

Acknowledgments

We thank the anonymous reviewers for their insightful comments and suggestions. This work was supported by the National Key RD Program of China via grant 2020AAA0106501 and the National Natural Science Foundation of China (NSFC) via grant 62176078.

References

[Bosselut *et al.*, 2019] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. COMET: commonsense transformers for automatic knowledge graph construction. In *Proc. of ACL*, 2019.

[Busso *et al.*, 2008] Carlos Busso, Murtaza Bulut, Chi-Chun Lee, Abe Kazemzadeh, Emily Mower, Samuel Kim, Jeannette N. Chang, Sungbok Lee, and Shrikanth S. Narayanan. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation*, 2008.

[Ghosal *et al.*, 2019] Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander Gelbukh. DialogueGCN: A graph convolutional neural network for emotion recognition in conversation. In *Proc. of EMNLP*, 2019.

[Ghosal *et al.*, 2020] Deepanway Ghosal, Navonil Majumder, Alexander F. Gelbukh, Rada Mihalcea, and Soujanya Poria. COSMIC: commonsense knowledge for emotion identification in conversations. In *Proc. of ACL*, 2020.

[Hazarika *et al.*, 2018] Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. ICON: Interactive conversational memory network for multimodal emotion detection. In *Proc. of EMNLP*, 2018.

[Hu *et al.*, 2021] Dou Hu, Lingwei Wei, and Xiaoyong Huai. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. In *Proc. of ACL*, 2021.

[Ishiwatari *et al.*, 2020] Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Proc. of EMNLP*, 2020.

[Li *et al.*, 2017] Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, 2017.

[Li *et al.*, 2021] Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang. Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In *Proc. of ACL*, 2021.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, 2019.

[Lu *et al.*, 2020] Xin Lu, Yanyan Zhao, Yang Wu, Yijian Tian, Huipeng Chen, and Bing Qin. An iterative emotion interaction network for emotion recognition in conversations. In *Proc. of COLING*, 2020.

[Majumder *et al.*, 2019] Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. Dialoguernn: An attentive rnn for emotion detection in conversations. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2019.

[Poria *et al.*, 2019a] S. Poria, N. Majumder, R. Mihalcea, and E. Hovy. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 2019.

[Poria *et al.*, 2019b] Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proc. of ACL*, 2019.

[Poria *et al.*, 2021] Soujanya Poria, Navonil Majumder, Devamanyu Hazarika, Deepanway Ghosal, Rishabh Bhardwaj, Samson Yu Bai Jian, Pengfei Hong, Romila Ghosh, Abhinaba Roy, Niyati Chhaya, Alexander F. Gelbukh, and Rada Mihalcea. Recognizing emotion cause in conversations. *Cogn. Comput.*, 13(5):1317–1332, 2021.

[Sap *et al.*, 2019] Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *Proc. of AAAI*, 2019.

[Shen *et al.*, 2021] Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Proc. of AAAI*, 2021.

[Turcan *et al.*, 2021] Elsbeth Turcan, Shuai Wang, Rishita Anubhai, Kasturi Bhattacharjee, Yaser Al-Onaizan, and Smaranda Muresan. Multi-task learning and adapted knowledge models for emotion-cause extraction. In *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, 2021.

[Xie *et al.*, 2021] Yunhe Xie, Kailai Yang, Chengjie Sun, Bingquan Liu, and Zhenzhou Ji. Knowledge-interactive network with sentiment polarity intensity-aware multi-task learning for emotion recognition in conversations. In *Proc. of ACL*, 2021.

[Zhang *et al.*, 2019] Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, and Guodong Zhou. Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *Proc. of IJCAI*, 2019.