MuCDN: Mutual Conversational Detachment Network for Emotion Recognition in Multi-Party Conversations

Weixiang Zhao, Yanyan Zhao*, Bing Qin
Research Center for Social Computing and Information Retrieval
Harbin Institute of Technology, China
{wxzhao, yyzhao, qinb}@ir.hit.edu.cn

Abstract

As an emerging research topic in natural language processing community, emotion recognition in multi-party conversations has attained increasing interest. Previous approaches that focus either on dyadic or multi-party scenarios exert much effort to cope with the challenge of emotional dynamics and achieve appealing results. However, since emotional interactions among speakers are often more complicated within the entangled multi-party conversations, these works are limited in capturing effective emotional clues in conversational context. In this work, we propose Mutual Conversational Detachment Network (MuCDN) to clearly understand the conversational context by separating conversations into detached threads. Specifically, two detachment ways are devised to perform context and speaker-specific modeling within detached threads and they are bridged through a mutual module. Experimental results on two datasets show that our model achieves better performance over the baseline models.

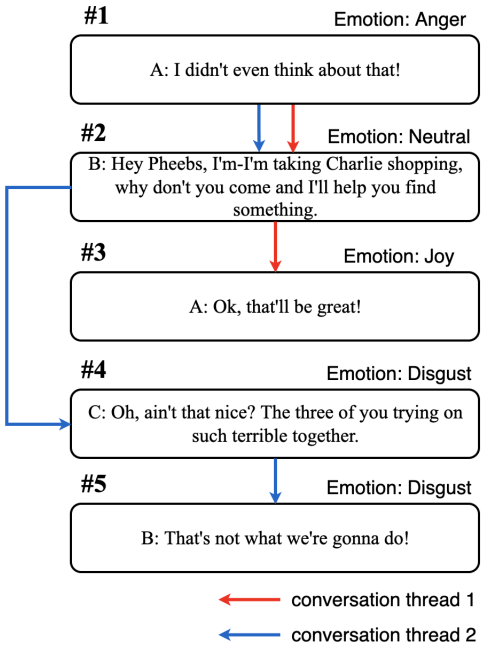

Figure 1: An example of a multi-party conversation from MELD dataset.

1 Introduction

Emotion recognition in conversations (ERC) is a task that predicts the emotion for each utterance in conversations. Since it plays a significant role in achieving empathetic systems and is of great values to be applied in the fields of opinion mining in conversations, social media analysis, mental health care and other areas, ERC has received more and more attention in the natural language processing (NLP) community. We focus on emotion recognition in multi-party conversations (ERMC) where three or more speakers are involved.

The challenge of ERC, especially for that of ERMC, lies in a complicated emotional interaction referred to as emotional dynamics (Poria et al., 2019b) that utterances from one speaker would have impact on expressions of others. To cope with the challenge, many previous approaches exploit recurrent neural network (RNN) (Majumder et al., 2019; Ghosal et al., 2020) and graph neural network (GNN) (Ghosal et al., 2019; Shen et al., 2021b; Zhang et al., 2022) to perform context and speaker-specific modeling. However, they are not effective enough to capture exact emotional clues for emotion recognition because the salient characteristic of multi-party conversations is ignored.

Multiple conversation threads are intermingled in the same dialog history simultaneously in the multi-party scenario (Ma et al., 2021; Liu et al., 2021a), which makes it hard to capture emotional clues in such a complex and confusing interaction. As shown in Figure 1, there are two conversation threads in a multi-party conversation consisting of three speakers A, B and C. Targeting at the same utterance #2 from speaker B to reply, speaker A is consoled and becomes joyful to go shopping, while

*Corresponding author

Proceedings of the 29th International Conference on Computational Linguistic, pages 7020–7030
October 12–17, 2022.

speaker C is unbearable towards the outfit of B and feels disgusted. Since speaker C and A originally focus on different aspects of the same dialog history, two individual threads are naturally developed. It is of great necessity to figure out the intra- and inter-personal dependencies of utterances within distinct conversation threads, for the convenience of understanding the exact emotional interaction exerted on different speakers.

In this paper, in order to clearly comprehend emotional clues for emotion recognition, we propose Mutual Conversational Detachment Network (MuCDN) to separate a multi-party conversation into detached threads and effectively perform context and speaker-specific modeling within them.

To be more specific, two detachment ways, named explicit detachment and implicit detachment are devised to separate a multi-party conversation into several threads and jointly carry out context and speaker-specific modeling within them. For the former one, we detach a multi-party conversation with the help of a dialog discourse parser. Along with paths of a discourse tree, detached conversation threads could be explicitly attained. Then two speaker-aware gated neural networks (GRU) are adopted for conversational information propagating. Further, implicit detachment aims at capturing the latent and global interaction in the conversation. We construct two speaker-specific implicit detachment graphs and utilize self-attention mechanism to obtain detached threads implicitly. In addition, a mutual module is designed to create the interaction between explicit detachment and implicit detachment. To evaluate the performance of MuCDN, we conduct extensive experiments on two ERMC datasets and new state-of-the-art performance is achieved on both of them.

The main contributions of this work are summarized as follows:

- In order to cope with the complicated emotional dynamics in ERMC, we propose MuCDN with the notion of conversation detachment.
- We devise two detachment ways to separate multi-party conversations into distinct threads and clearly perform context and speaker-specific modeling within them. A mutual module is also designed for interaction.
- Results of extensive experiments on two benchmark ERMC datasets demonstrate the effectiveness of the proposed model. Our source code is publicly available at https://github.com/circle-hit/MuCDN.

2 Related Work

2.1 Emotion Recognition in Conversations

Recent works that focus on dyadic conversations also perform experiments on ERMC datasets. They either plainly perform context modeling or incorporate external resources, which can be divided into two categories.

For those without incorporating external resources, two types of deep neural network are adopted for context modeling. (1) **RNN**. Majumder et al. (2019) take global state, personal state and emotion state of speakers into account and utilize three GRUs to model emotional dependency among speakers. Hu et al. (2021) devise a cognitive reasoning module to iteratively capture emotional clues and fully understand conversational context with cognitive factors. (2) **GNN**. Ghosal et al. (2019) perform context and speaker-specific modeling upon a graph and relational graph convolutional network is adopted for message passing through various types of edges. Further, Ishiwatari et al. (2020) simplify the selection of some types of edges in the graph and propose relational position encoding to enhance relation-aware graph attention network. Shen et al. (2021b) abstract a conversation into a directed acyclic graph and extend the directed acyclic graph neural network to be suitable under the conversation setting.

More recently, many works leverage external resources to enrich contextual representations. Zhong et al. (2019) utilize commonsense knowledge and emotion lexicon to guide context modeling and dynamically retrieve context-aware and emotion-related knowledge. Ghosal et al. (2020) explore several types of commonsense knowledge for a comprehensive understanding of various aspects of conversations such as personality, events, mental states and intents. Based on this, Li et al. (2021) concentrate more on psychological interactions between utterances. Besides, an auxiliary task named sentiment polarity intensity prediction is introduced to involve direct affective information from the emotion lexicon (Xie et al., 2021). And Zhao et al. (2022) leverage commonsense knowledge as causal clues to aid the emotion recognition with emotion cause detection.

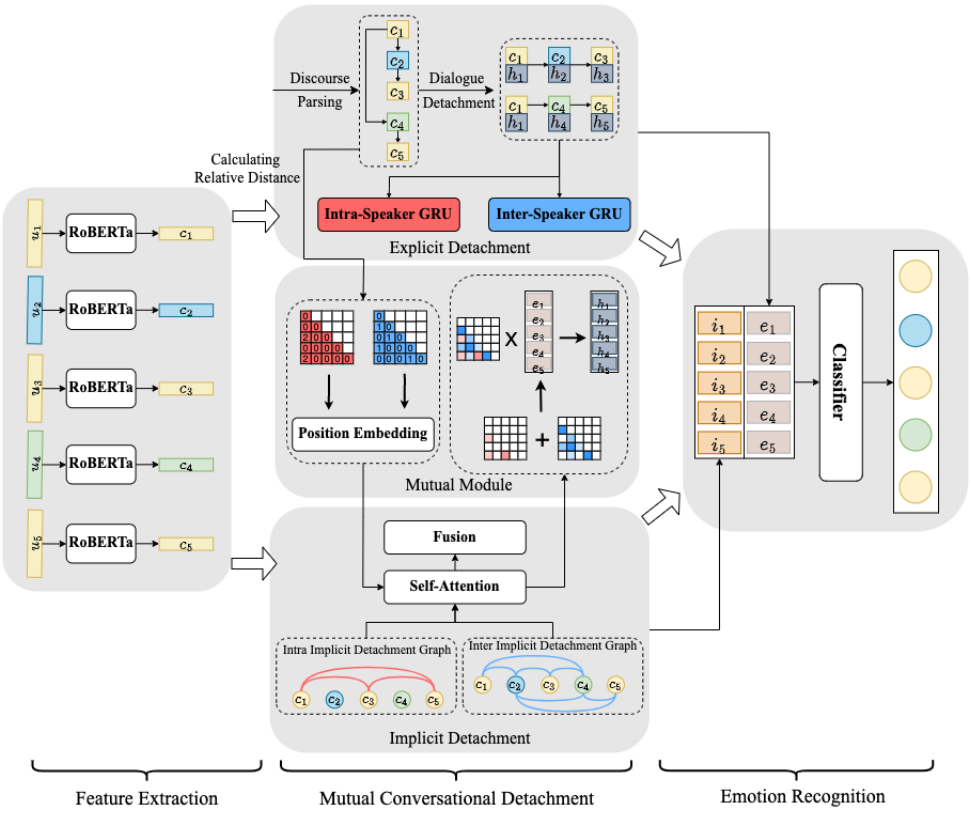

Figure 2: The overall architecture of our proposed model.

2.2 Emotion Recognition in Multi-Party Conversations

For emotion recognition in multi-party conversations, recent studies endeavor to capture contextual clues and speaker-specific information in the complex interactions. Zhang et al. (2019) construct a large graph over the entire corpus and propose a conversational graph-based convolutional neural network to model both context- and speaker-sensitive dependency. Shen et al. (2021a) make the pretrained language model XL-Net (Yang et al., 2019) adaptive to the dialog scenario and design four types of dialog-aware self-attention to model contextual information. Sun et al. (2021) explore the importance of discourse structures in handling informative contextual cues and speaker-specific features and build a graph based on the dialog discourse structure.

However, all the aforementioned methods ignore the notable characteristic of multi-party conversations that multiple conversation threads are entangled in the dialog history. And in order to clearly capture emotional clues for emotion recognition, we propose to perform context and speaker-specific modeling within individual threads detached from multi-party conversations.

3 Methodology

First, we define the problem of the ERMC task. Given a conversation with N consecutive utterances $\{u_1, u_2, \cdots, u_N\}$ and M speakers ($M > 2$) $\{s_1, s_2, \cdots, s_M\}$, ERMC aims to predict the emotion label e_i of each utterance u_i spoken by s_i.

Figure 2 gives the overall architecture of our proposed model, which consists of five parts: Utterance-Level Feature Extraction, Explicit Detachment, Implicit Detachment, Mutual Module and Emotion Recognition. After attaining context-independent vector of each utterance, we propose to clearly carry out context and speaker-specific modeling within individual threads detached from multi-party conversations. On the one hand, explicit detachment is performed on the outcome of the dialogue discourse parser and two speaker-aware GRUs are adopted for information propagating across each detached thread. On the other hand, the latent and global interaction is captured by implicit detachment through attention values calculated by self-attention mechanism. Further, mutual module exchanges the interaction information for both explicit detachment and implicit detachment. And the final representation is obtained by concate-

nating outcomes from the two detachment modules. We will elaborate each one of the five modules in the rest of this section.

3.1 Utterance-Level Feature Extraction

We employ the widely-used pretrained language model RoBERTa (Liu et al., 2019) to perform utterance-level feature extraction. Specifically, for each utterance $u_i = \{w_1, w_2, \cdots, w_L\}$, a special token $[CLS]$ is concatenated to the beginning of the utterance. Then we feed the sequence $\{[CLS], w_1, w_2, \cdots, w_L\}$ to fine-tune the pretrained RoBERTa model by an utterance-level emotion classification task and the $[CLS]$ token from the last hidden layer is fed to a pooling layer to attain the result of emotion classification.

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

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

where $c_i \in R^{d_m}$ and d_m is the dimension of hidden states 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. Then each utterance vector c_i is transformed to the dimension of d_h with a linear projection. And the vectorized representation of a conversation C is $\{c_1, c_2, \cdots, c_N\}$.

3.2 Explicit Detachment

To clearly and effectively perform context and speaker-specific modeling under the circumstance of multi-party scenario, we separate a conversation into several individual threads according to the results of its corresponding discourse structure.

Discourse Parsing and Detachment: We utilize the discourse parser proposed by Shi and Huang (2019), which is a deep sequential model and performs well on STAC (Asher et al., 2016) corpus. Given each conversation $\{u_1, u_2, \cdots, u_N\}$, the discourse structure could be obtained by:

$$\{(i, j, e_{ij}), \cdots\} = \text{Parser}(\{u_1, u_2, \cdots, u_N\}) \quad (2)$$

where e_{ij} is the directed edge with head j and tail i and i, j are indices of utterances in the conversation ($i < j$). According to their settings, the output of the parser is a discourse dependency tree where each node (utterance) is connected with a single parent node. And detached threads are naturally formed according to each definite path from the start node to the current one. We use an explicit detachment matrix D to represent the derived threads from discourse tree:

$$D_{i,j} = \begin{cases} 1, & \text{if } e_{ij} \text{ exists in discourse tree} \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

Speaker-Aware Context Modeling: With individual conversation threads detached, we propose to perform clear and effective context and speaker-specific modeling upon them. Specifically, we adopt two speaker-aware GRUs, named intra-speaker GRU and inter-speaker GRU, for contextual information propagating through each detached thread along with connections guided by explicit detachment matrix D:

$$e_i = \begin{cases} \text{GRU}^{intra}(c_i,\ e_p), & \text{if } \phi(u_i) = \phi(u_p) \\ \text{GRU}^{inter}(c_i,\ e_p), & \text{otherwise} \end{cases} \quad (4)$$

where $e_i \in R^{d_h}$ and ϕ maps the utterance into that of the corresponding speaker and u_p is the single precursor of the current utterance u_i which means $D_{pi} = 1$ ($p < i$). GRU^{intra} is utilized to model the intra-speaker dependency from the same speaker, while GRU^{inter} are for interactions from other speakers.

3.3 Implicit Detachment

The reason for the design of implicit detachment is that the explicit detachment mainly focuses on the local interaction, which means only recurrently propagating information in each thread would ignore the global information contained in a multi-party conversation. When a speaker is buried in a conversation, the triggering of his/her target emotion may be more or less from multi sources, not just a single precursor analyzed by the parser. Thus, to capture the global emotional clues and dig the latent emotional dependency among utterances, we attempt to detach the multi-party conversation in an implicit way. Specifically, we construct two partially fully-connected implicit detachment graphs (IDG) with speaker-aware information injected according to whether two utterances are from the same speaker or not. IDG^{intra} and IDG^{inter} are obtained in the form of adjacent matrices by:

$$IDG^{intra}_{i,j} = \begin{cases} 0, & \text{if } j <= i \text{ and } \phi(u_i) = \phi(u_j) \\ -\infty, & \text{otherwise} \end{cases} \quad (5)$$

$$IDG_{i,j}^{inter} = \begin{cases} 0, \text{ if } j < i \text{ and } \phi(u_i) \neq \phi(u_j) \\ -\infty, \text{ otherwise} \end{cases} \tag{6}$$

Then we utilize multi-head self-attention mechanism (MHSA) and add IDG to the result of dot product between query and key to achieve speaker-aware information propagating. We omit the formula for multi-head computation due to the limited space and more details about MHSA could be found in Vaswani et al. (2017):

$$\begin{aligned} G &= \text{MHSA}(C,\ IDG^t), \\ \text{Att}(Q, K, V, IDG^t) &= \text{Softmax}(\frac{QK^T}{\sqrt{d_k}} + IDG^t)V \end{aligned} \tag{7}$$

where $G \in R^{N \times d_h}$ and $t \in \{intra, inter\}$ is the type of IDG. Viewing self-attention weights as edges connecting representations, to what extent we obtain the detached threads is determined by the attention values calculated by the self-attention mechanism. we then fuse the two speaker-aware implicit detachment representations with a gated manner inspired by Liu et al. (2021b):

$$\begin{aligned} F^t &= \text{ReLU}(\text{FC}([C, G^t, C - G^t, C \odot G^t])), \\ g &= \text{Sigmoid}(\text{FC}[F^{intra}, F^{inter}]), \\ I &= g \odot F^{intra} + (1 - g) \odot F^{inter} \end{aligned} \tag{8}$$

where $I \in R^{N \times d_h}$ and FC is the fully-connected layer.

3.4 Mutual Module

This module is designed for mutual interaction between the process of explicit detachment and implicit detachment.

For the interaction from implicit detachment to explicit detachment, it servers as the complement of the global information and latent interactions in the conversations. First, we obtain two attention score matrices from the self-attention layer with the average of each attention head. And the joint score A^{joint} is measured in the same scale by softmax function. Then the i-th global representation complemented for its corresponding utterance in the explicit detached threads is obtained by:

$$h_i = A_{i,<i}^{joint} \times E_{<i} \tag{9}$$

where $h_i \in R^{d_h}$, $\times$ represents the operation of matrix multiply, $A_{i,<i}^{joint}$ is the prior $i-1$ elements in the i-th row of A^{joint} and $E_{<i} = \{e_1, e_2, \cdots, e_{i-1}\}$ stands for the prior $i-1$ vectors calculated by speaker-aware GRU. We concatenate each h_i with its corresponding c_i before thread information propagating and the original computation in Equation (4) is updated:

$$e_i = \begin{cases} \text{GRU}^{intra}([c_i, h_i],\ e_p), & \text{if } \phi(u_i) = \phi(u_p) \\ \text{GRU}^{inter}([c_i, h_i],\ e_p), & \text{otherwise} \end{cases} \tag{10}$$

In addition, explicit detachment would provide exact relative position information based on the discourse tree for the process of implicit detachment. Ishiwatari et al. (2020) argue that human emotions may depend on more immediate utterances in the temporal order. We extend such a temporal distance to the structural form under the circumstance of multi-party scenario and assume that upon the discourse structure, the more immediate utterances may be more relevant to the target one and are more likely to appear in the same thread. Therefore, we calculate the relative distance measured by hops upon the discourse tree and obtain two position matrix P^{intra} and P^{inter} according to the speaker type. And if two nodes are unreachable to each other, we set the position value to zero. Then we adopt a trainable position embedding layer to encode each element in P^{intra} or P^{inter} to a scalar value and add them to the attention scores in the multi-head attention layer.

$$\begin{aligned} Pos^t &= \text{Embedding}(P^t), \\ G &= \text{MHSA}(C,\ IDG^t, Pos^t), \\ \text{Att}(Q, K, V, IDG^t, Pos^t) &= \text{Softmax}(\frac{QK^T}{\sqrt{d_k}} \\ &+ IDG^t + Pos^t)V \end{aligned} \tag{11}$$

Through this, Pos^t could be viewed as an extra weight to guide the fully-connected process of implicit detachment which is considered to be a relatively blind way.

3.5 Emotion Recognition

Finally, taking the the concatenation of C, E and I as input, an emotion classifier is applied to predict the emotion of the utterance.

$$\hat{y} = \text{Softmax}(W_e[C, E, I] + b_e) \tag{12}$$

where W_e and b_e are trainable parameters.

Dataset	Dialogues			Utterances		
	Train	Val	Test	Train	Val	Test
EmoryNLP	713	99	85	9,934	1,344	1,328
MELD	1,039	114	280	9,989	1,109	2,610

Table 1: Dataset statistics

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

$$L = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{Emo}\hat{y}_i^j \cdot log(y_i^j) \quad (13)$$

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

4 Experiments

4.1 Dataset

We conduct experiments on two ERMC datasets. The statistics of them are shown in Table 1.

EmoryNLP (Zahiri and Choi, 2018): TV show scripts collected from *Friends* and the emotion classes include *neutral*, *sad*, *mad*, *scared*, *powerful*, *peaceful* and *joyful*.

MELD (Poria et al., 2019a): A multimodal dataset with multi-speaker conversations. It is also collected from the TV show *Friends*, but varies from EmoryNLP in the choice of scenes and emotion labels. And the emotion labels belong to *anger*, *disgust*, *fear*, *joy*, *neutral*, *sadness* and *surprise*.

We utilize only the textual modality for the experiments. Following previous works, the weighted F1 score is chosen as evaluation metrics.

4.2 Compared Models

We compare our proposed model with the following methods. CSK is short for the commonsense knowledge:

Methods for ERMC:

ConGCN (Zhang et al., 2019) constructs a large graph over the entire corpus to model both context- and speaker-sensitive dependency.

DialogXL (Shen et al., 2021a) devises four types of dialog-aware self-attention to make the model aware of interactions in multi-party conversations.

ERMC-DisGCN (Sun et al., 2021) builds a graph based on the dialog discourse structure to explore context and speaker-specific features.

Methods for ERC with CSK:

KET (Zhong et al., 2019) leverages commonsense knowledge to enrich context modeling and dynamically retrieves context-aware and emotion-related knowledge.

KAITML (Zhang et al., 2020) applies incremental transformer to encode multi-turn contextual utterances with commonsense knowledge incorporated and introduces multi-task learning to this task.

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

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.

COSMIC (Ghosal et al., 2020) explores more types of commonsense knowledge about several factors of conversations to understand emotional dynamics better.

Methods for ERC without CSK:

DialogueRNN (Majumder et al., 2019) devises three states including global state, party state and emotion state with GRUs.

DialogueGCN (Ghosal et al., 2019) uses graph convolutional network to encode context and speaker dependencies.

IEIN (Lu et al., 2020) focuses on explicit interactions among emotion of utterances and iteratively predicts emotion labels based on previous ones.

RGAT (Ishiwatari et al., 2020) enhances relation-aware graph attention network with the proposed relational position encoding.

DialogueCRN (Hu et al., 2021) designs reasoning modules from a cognitive perspective to fully integrate emotional clues.

DAG-ERC (Shen et al., 2021b) presents a idea of modeling conversation context with a directed acyclic graph and proposes a directed acyclic graph neural network.

4.3 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. And for the training of MuCDN on emotion recognition, the batch size is set to be {16, 8} for EmoryNLP and MELD respectively. The dimension of hidden representation d_h in the rest of our model is set to 300 and the number of attention head is 6. We train the model with Adam optimizer in a learning rate of 1e-4. All of our results are averaged on 5 runs.

Model	EmoryNLP	MELD
ERMC Methods		
ConGCN	-	57.40
DialogXL	34.73	62.41
ERMC-DisGCN	36.38	64.22
ERC Methods with CSK		
KET	34.39	58.18
KAITML	35.59	58.97
KI-Net	-	63.24
SKAIG	38.88	65.18
COSMIC	38.11	65.21
COSMIC w/o CSK	37.10	64.28
ERC Methods without CSK		
DialogueRNN	31.7	57.03
DialogueGCN	-	58.1
IEIN	-	60.72
RGAT	34.42	60.91
DialogueCRN	-	58.39
DAG-ERC	39.02	63.65
MuCDN (Ours)	**40.09**	**65.37**

Table 2: Comparison of our model against state-of-the-art baselines. CSK represents the commonsense knowledge utilized in COSMIC. Weighted F1 score is adopted as evaluation metrics.

5 Results and Analysis

5.1 Overall Results

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

EmoryNLP. Since there are often more than 5 participants within a conversation in EmoryNLP, which leads to the results that average individual conversation threads derived from Explicit Detachment module are 4.77/4.66/4.91 (Train/Valid/Test), the complicated emotional interactions pose great challenge for capturing contextual clues and speaker features. Benefiting from a clear and effective context and speaker-specific modeling by separating multi-party conversations into individual threads, MuCDN achieves new state-of-the-art weighted F1 scores of 40.09. Compared with DAG-ERC which proposes a directed acyclic graph to link utterances in a locally fully-connected way, distinct detached threads are more suitable and effective for conversation context modeling in multi-party scenario. In addition, the improvement of performance over ERMC-DisGCN demonstrates it may not be effective enough to simply propagate contextual information upon the discourse graph. To a certain extent, such improvement verify the effectiveness of our proposed discourse-based detachment, which is a better way of utilizing discourse structure to clearly perform conversation modeling. And it also suggests the effectiveness of the complementary information provided by implicit detachment and mutual module.

Model	EmoryNLP	MELD
MuCDN	**40.09**	**65.37**
w/o explicit detachment	38.45	64.45
w/o implicit detachment	38.84	64.47
w/o E2I interaction	39.28	64.61
w/o I2E interaction	39.54	64.56

Table 3: Results of ablation study on the two ERMC datasets. E2I interaction is the relative position embedding provided by explicit detachment, while I2E interaction is the complementary global information from implicit detachment.

MELD. Advantages brought by clear conversation modeling could also be observed on MELD where MuCDN performs better than all the baselines. Here, the results of the average detached conversation threads are 3.92/3.96/3.85 (Train/Valid/Test). However, it is worth noting that our model slightly outperforms those baseline models with commonsense knowledge incorporated such as COSMIC. To make a clear comparison regarding the conversation modeling ability of the model, we also compare our model with COSMIC without commonsense knowledge implemented by Shen et al. (2021b). And under the same circumstance of no commonsense knowledge incorporated, the performance advantage of MuCDN comes from the clear and effective capture of emotional clues in individual conversation thread. Meanwhile, it reminds us to enrich contextual representations with commonsense knowledge within each detached thread for further improvement.

5.2 Ablation Study

We conduct ablation studies to verify the effectiveness of different modules proposed in our model. Results are shown in Table 3.

Effect of Conversation Detachment

To investigate the impact of two detachment modules, we remove either the explicit detachment module or the implicit detachment module, and the corresponding mutual interaction is also discarded at the same time. The performance has a certain degree of decline on both datasets and results are displayed in the second row and the third row in Table 3. This manifests that both ways of

Model	EmoryNLP	MELD
MuCDN	**40.09**	**65.37**
sequence	39.05	64.51
randomness	38.72	64.71

Table 4: Results of our model replaced with different types of dependency structure connecting utterances in Explicit Detachment module.

Model	EmoryNLP	MELD
MuCDN	**40.09**	**65.37**
w/o intra and inter GRU	39.42	64.49
w/o intra and inter graph	38.91	64.46

Table 5: Results of our model without speaker-specific modeling.

explicit detachment and implicit detachment play an important role in simplifying and clarifying the complex interactions within multi-party conversations and lay the foundation for effective context and speaker-specific modeling.

Effect of Mutual Interaction

To verify the effectiveness of the mutual interaction that explicit detachment and implicit detachment provide for each other, the relative position embedding and the complementary global information are removed individually. The dropped results in the second-to-last row on both ERMC datasets demonstrates that the extra relative position features derived from the dialogue discourse structure provide implicit detachment with explicit guidance to capture the latent emotional interactions better. Besides, results in the last row suggests that the global information from implicit detachment complement multi-source emotional interactions for the context modeling of each thread.

5.3 Variants of Dependency Structure in Explicit Detachment

In this section, we investigate how explicit detachment benefits from dialogue discourse structure for a clear and effective conversation modeling. Two additional kinds of dependency structure of utterances are devised: (1) sequence, in which utterances are connected one by one following the temporal order; (2) randomness, in which each utterance link any one of the previous utterances with a random selection. To achieve this, the detachment matrix D is substituted and the position matrix P in mutual module is also changed to be consistent with such two types of dependency structure. We keep all the rest parts same with our complete model.

The test performance is shown in Table 4 and we make two instructive observations from the experimental results. First, the performance of random structure is not as terrible as we expected, even a little better than that of sequential structure on MELD dataset. On the one hand, it manifests that the effectiveness of the complementary global information provided by implicit detachment and latent emotional interactions among utterances are captured. On the other hand, the little improvement over the structure of sequence suggests that it may not be suitable enough to plainly model the conversational context without any explicit detachment for the entangled multi-party conversations. Second, both results obtained upon the variants of dependency structure do not decrease to a large margin. It reminds us that the results of the dialogue discourse parser are not reliable enough and the problem of error propagation may have influence on our model. And effective techniques to prune the dialogue discourse tree should be explored for future work.

5.4 Analysis of Speaker-Specific Modeling

As a key factor to provide emotional clues for emotion recognition, especially under the circumstance of multi-party setting, modeling speaker-specific information has proven to be beneficial. We investigate the impact of speaker-aware modules in our proposed model. First, instead of utilizing intra-speaker GRU and inter-speaker GRU in explicit detachment, we only adopt a single GRU without the speaker-aware identification for contextual information propagating through each detached thread. Also, only one implicit detachment graph is constructed without speaker information injected to identify connections among utterances.

According to results shown in Table 5, the drop of performance implies that it is effective to take speaker-specific information into account when carrying out conversational context modeling. And it also suggests the importance of figuring out the intra and inter-speaker dependency to capture emotional clues. But such a relatively slight drop inspires us that the identification of whether two utterances have the same speaker may not be sufficient in multi-party scenario. Richer speaker-specific information contained in the contextual context should be excavated further.

6 Conclusion

In this paper, in order to capture emotional clues for emotion dynamics in a clear and effective way under the circumstance of multi-party scenario, we propose Mutual Conversational Detachment Network (MuCDN) for emotion recognition in multi-party conversations. Joint context and speaker-specific modeling is performed within individual detached threads by two detachment ways and a mutual module. Experimental results on two ERMC datasets demonstrate the superiority of our proposed MuCDN. Also, by conducting comprehensive evaluations and ablation study, we confirmed the effectiveness of our MuCDN and the impact of its components.

For future work, we would like to explore effective techniques to prune the dialogue discourse tree to alleviate error propagation. Moreover, for a more comprehensive speaker-specific modeling, richer speaker information in the conversational context should be excavated further.

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

Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos D. Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the STAC corpus. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016*. European Language Resources Association (ELRA).

Deepanway Ghosal, Navonil Majumder, Alexander F. Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: commonsense knowledge for emotion identification in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020, Online Event, 16-20 November 2020*, volume EMNLP 2020 of *Findings of ACL*, pages 2470–2481. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander F. Gelbukh. 2019. 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, Hong Kong, China, November 3-7, 2019*, pages 154–164. Association for Computational Linguistics.

Dou Hu, Lingwei Wei, and Xiaoyong Huai. 2021. Dialoguecrn: Contextual reasoning networks for emotion recognition in conversations. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 7042–7052. Association for Computational Linguistics.

Taichi Ishiwatari, Yuki Yasuda, Taro Miyazaki, and Jun Goto. 2020. Relation-aware graph attention networks with relational position encodings for emotion recognition in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 7360–7370. Association for Computational Linguistics.

Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang. 2021. Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 1204–1214. Association for Computational Linguistics.

Hui Liu, Zhan Shi, and Xiaodan Zhu. 2021a. Unsupervised conversation disentanglement through co-training. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2345–2356, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Longxiang Liu, Zhuosheng Zhang, Hai Zhao, Xi Zhou, and Xiang Zhou. 2021b. Filling the gap of utterance-aware and speaker-aware representation for multi-turn dialogue. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13406–13414. AAAI Press.

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

Xin Lu, Yanyan Zhao, Yang Wu, Yijian Tian, Huipeng Chen, and Bing Qin. 2020. An iterative emotion interaction network for emotion recognition in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13,*

2020, pages 4078–4088. International Committee on Computational Linguistics.

Xinbei Ma, Zhuosheng Zhang, and Hai Zhao. 2021. Structural modeling for dialogue disentanglement. *CoRR*, abs/2110.08018.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander F. Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive RNN for emotion detection in conversations. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6818–6825. AAAI Press.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019a. MELD: A multimodal multi-party dataset for emotion recognition in conversations. In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 527–536. Association for Computational Linguistics.

Soujanya Poria, Navonil Majumder, Rada Mihalcea, and Eduard H. Hovy. 2019b. Emotion recognition in conversation: Research challenges, datasets, and recent advances. *IEEE Access*, 7:100943–100953.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021a. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13789–13797. AAAI Press.

Weizhou Shen, Siyue Wu, Yunyi Yang, and Xiaojun Quan. 2021b. Directed acyclic graph network for conversational emotion recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 1: Long Papers), Virtual Event, August 1-6, 2021*, pages 1551–1560. Association for Computational Linguistics.

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 7007–7014. AAAI Press.

Yang Sun, Nan Yu, and Guohong Fu. 2021. A discourse-aware graph neural network for emotion recognition in multi-party conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2949–2958. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Yunhe Xie, Kailai Yang, Chengjie Sun, Bingquan Liu, and Zhenzhou Ji. 2021. Knowledge-interactive network with sentiment polarity intensity-aware multitask learning for emotion recognition in conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 2879–2889. Association for Computational Linguistics.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime G. Carbonell, Ruslan Salakhutdinov, and Quoc V. Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 5754–5764.

Sayyed M. Zahiri and Jinho D. Choi. 2018. Emotion detection on TV show transcripts with sequence-based convolutional neural networks. In *The Workshops of the The Thirty-Second AAAI Conference on Artificial Intelligence, New Orleans, Louisiana, USA, February 2-7, 2018*, volume WS-18 of *AAAI Technical Report*, pages 44–52. AAAI Press.

Dong Zhang, Liangqing Wu, Changlong Sun, Shoushan Li, Qiaoming Zhu, and Guodong Zhou. 2019. Modeling both context- and speaker-sensitive dependence for emotion detection in multi-speaker conversations. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 5415–5421. ijcai.org.

Duzhen Zhang, Xiuyi Chen, Shuang Xu, and Bo Xu. 2020. Knowledge aware emotion recognition in textual conversations via multi-task incremental transformer. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4429–4440. International Committee on Computational Linguistics.

Duzhen Zhang, Zhen Yang, Fandong Meng, Xiuyi Chen, and Jie Zhou. 2022. TSAM: A two-stream attention model for causal emotion entailment. *CoRR*, abs/2203.00819.

Weixiang Zhao, Yanyan Zhao, and Xin Lu. 2022. Cauain: Causal aware interaction network for emotion recognition in conversations. In *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI 2022, Vienna, Austria, 23-29 July 2022*, pages 4524–4530. ijcai.org.

Peixiang Zhong, Di Wang, and Chunyan Miao. 2019. Knowledge-enriched transformer for emotion detection in textual conversations. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019*, pages 165–176. Association for Computational Linguistics.