A Discourse-Aware Graph Neural Network for Emotion Recognition in Multi-Party Conversation

Yang Sun[1], Nan Yu[1], Guohong Fu[1,2]*

[1]School of Computer Science and Technology, Soochow University, China

[2]Institute of Artificial Intelligence, Soochow University, China

{ysun23,nyu}@stu.suda.edu.cn

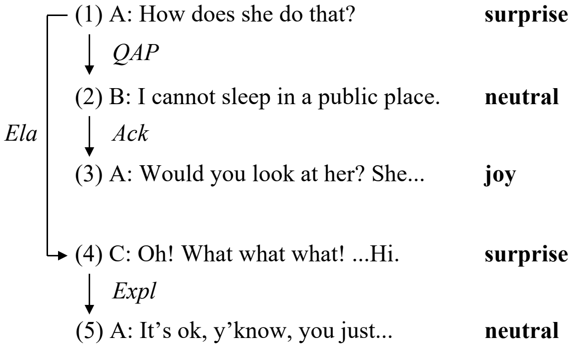
ghfu@suda.edu.cn

Abstract

Emotion recognition in multi-party conversation (ERMC) is becoming increasingly popular as an emerging research topic in natural language processing. Prior research focuses on exploring sequential information but ignores the discourse structures of conversations. In this paper, we investigate the importance of discourse structures in handling informative contextual cues and speaker-specific features for ERMC. To this end, we propose a discourse-aware graph neural network (ERMC-DisGCN) for ERMC. In particular, we design a relational convolution to lever the self-speaker dependency of interlocutors to propagate contextual information. Furthermore, we exploit a gated convolution to select more informative cues for ERMC from dependent utterances. The experimental results show our method outperforms multiple baselines, illustrating that discourse structures are of great value to ERMC.

1 Introduction

In the past few years, emotion recognition in conversation (ERC) has become increasingly popular in natural language processing (NLP) with the proliferation of open conversational data on social media platforms (Poria et al., 2019a). Similar to text sentiment analysis, ERC is a task to determine the emotion of each utterance within a conversation, as shown in Fig. 1, and plays important role in many NLP applications, such as opinion mining in conversation (Cambria et al., 2017), social media analysis (Majumder et al., 2019) and emotion-aware dialogue systems (Ghosal et al., 2019). However, ERC, particularly the emotion recognition in multi-party conversation (ERMC), often exhibits more difficulties than traditional text sentiment analysis due to the emotional dynamics of conversations (Poria et al., 2019b). Consequently, recognizing the emotion of an utterance in a multi-party conversation primarily depends on not only the utterance itself and its context but also the self and interpersonal dependencies and the emotions expressed in the preceding utterances (Poria et al., 2017; Majumder et al., 2019; Jiao et al., 2019; Zhong et al., 2019; Shen et al., 2021).



(1) A: How does she do that? **surprise**

(2) B: I cannot sleep in a public place. **neutral**

(3) A: Would you look at her? She... **joy**

(4) C: Oh! What what what! ...Hi. **surprise**

(5) A: It's ok, y'know, you just... **neutral**

Figure 1: An example of the ERC task, the gold labels are different emotions of the utterances, and the discourse structure is shown on the left. QAP, Ack, Ela, and Expl respectively represent the Question-Answer Pair, Acknowledgment, Elaboration, and Explanation relations.

Many approaches have been proposed for ERC with a focus on conversational context representation and speaker-specific modeling. While earlier works on ERC focus on two-party conversation and exploit recurrent neural networks (RNNs) to capture sequential context features of conversations (Poria et al., 2017; Majumder et al., 2019; Jiao et al., 2019; Ghosal et al., 2019), recent studies exert more efforts on ERMC and explore different techniques such as multi-task learning (Li et al., 2020) and pre-training language modeling (Shen et al., 2021) to capture speaker-specific information. Although these studies have greatly promoted the progress of ERC, most of them ignore the important conversational discourse structures. Therefore, they can only leverage cues in neighboring context of conversations, and are difficult to handle

*Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 2949–2958
November 7–11, 2021. ©2021 Association for Computational Linguistics

informative distant dependencies for ERC.

Actually, conversational discourse structures contain discourse relations or discourse dependencies between utterances and thus provide a straightforward way to capture both adjacent and distant cues for ERMC. Fig. 1 illustrates a multi-party conversation example with its discourse structure obtained from the discourse parser proposed by Shi and Huang (2019). As we can see, although the first and the fourth utterances are distant in position within the conversation, they have an immediate discourse relation and are thus annotated with the same emotion type *surprise*. Therefore, such discourse relations offer important contextual cues for ERMC. On the other hand, discourse structures have proven to be useful for document-level sentiment analysis (Bhatia et al., 2015; Märkle-Huß et al., 2017; Kraus and Feuerriegel, 2019) and we believe that they are also beneficial for ERMC. Moreover, recent progress in conversational discourse parsing (Shi and Huang, 2019; Li et al., 2021) makes it applicable to explore discourse structures to help model conversational contexts and speakers for ERMC.

However, two new problems may arise when discourse structures are applied to ERMC. First, previous works have shown that speaker-specific information is very important for ERMC (Zhang et al., 2019; Li et al., 2020). So it becomes a key issue how to incorporate conversational discourse structures into speaker-specific modeling for ERMC. Second, discourse structures involve dependent relations between utterances. However, not all information from dependent utterances is useful for conversational emotion recognition. Therefore, another important problem might be how to select more informative cues for ERMC.

To address the aforementioned issues, we propose a discourse-aware graph neural network for emotion recognition in multi-party conversation, named ERMC-DisGCN. It consists of three main modules: Firstly, a sequential context encoding module exploits Long Short-Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) to capture the sequential features of utterances in a conversation. Then, we exploit discourse dependency links and discourse relations to construct a graph, which contains two main convolution operations, namely a relational convolution and a gated convolution. The relational convolution is used to model the self-speaker dependency based on discourse structures, where individual speakers resist the change of their own emotion against external influence (Ghosal et al., 2019), while the gated convolution adopts a gated mechanism to select informative cues for ERMC from dependent utterances. Similar to (Zhang et al., 2019), we take utterances as nodes of the constructed graph. Finally, a decoding module is applied to predict the emotion label for each utterance. In addition, we employ the deep sequential discourse parser developed by Shi and Huang (2019) to obtain the explicit discourse dependency trees of input conversations.

In summary, we make the following contributions:

- We propose a discourse-aware graph neural network for emotion recognition in multi-party conversation (ERMC).
- We devise a discourse-based relational graph convolution to exploit the self-speaker dependency of interlocutors to propagate contextual information, and further use a gated convolution to select more informative cues for ERMC from dependent utterances.
- We conduct experiments on both multi-party and two-party conversation corpora, and demonstrate that using conversational discourse structures can benefit ERMC.

2 Related work

Recently, ERC has become a new trend due to the emergence of publicly available conversational datasets collected from social media platforms and scripted situations (Busso et al., 2008; Zahiri and Choi, 2018; Poria et al., 2019a). Earlier works focus on capturing sequential context features for emotion recognition in two-party conversation. Poria et al. (2017) propose a LSTM-based network to propagate contextual information within conversations. Majumder et al. (2019) propose a recurrent-based model to track the speaker states and global context during conversations. Jiao et al. (2019) propose a hierarchical Gated Recurrent Unit (GRU) (Chung et al., 2014) structure that trains utterance-level and conversation-level encoders jointly. Ghosal et al. (2019) construct a fully connected graph within a context utterance window to aggregate information. Zhong et al. (2019) incorporate external commonsense knowledge and employ the Transformer encoder (Vaswani et al., 2017) to capture contextual information.

For emotion recognition in multi-party conversation (ERMC), studies exert more effort in handling speaker-specific information. Zhang et al. (2019) represent the entire conversational corpus as a large graph to model speaker-sensitive dependency. Li et al. (2020) use speaker identification as an auxiliary task to capture speaker-specific features. Shen et al. (2021) propose an all-in-one XLNet (Yang et al., 2019) model with dialog-aware self-attention to deal with the multi-party structures. However, these studies neglect the informative discourse structures in multi-party conversations. To the best of our knowledge, we are the first to investigate the importance of discourse structures in handling informative contextual cues and speaker-specific features for ERMC.

Discourse structures have been successfully applied to document-level sentiment analysis (Bhatia et al., 2015; Märkle-Huß et al., 2017; Kraus and Feuerriegel, 2019), where discourse structures are produced by Rhetorical Structure Theory (RST) parser (Li et al., 2014). Recently, Shi and Huang (2019) propose a deep sequential model for conversational discourse parsing and achieve new state-of-the-art (SOTA) results. With this model, Jia et al. (2020) transform dialogue histories into threads for multi-turn response selection. Inspired by (Xia et al., 2019) and (Zhang et al., 2020), we exploit discourse dependency links and discourse relations to construct a graph. Especially, we stack two convolutional layers to aggregate contextual and speaker-specific information of the neighborhood for each utterance in the graph.

3 Methodology

3.1 Problem Definition

Suppose there are N constituent utterances $u_1, u_2, \ldots, u_N$ from a conversation with $X (X \geqslant 2)$ speakers $s_1, s_2, \ldots, s_X$. Utterance u_i is uttered by speaker $S_{m(u_i)}$, where the function m maps an utterance into its corresponding speaker. ERMC is to predict the emotion label for each utterance.

3.2 Pre-processing

Similar to most existing studies, the input of our model is a multi-party conversation consisting of context-independent utterance-level feature vectors. Besides, we need to obtain discourse structures to construct a graph. We complete these works in this pre-processing module.

Utterance Encoding: Earlier works adopt the Convolution Neural Network (CNN) (Kim, 2014) to obtain the feature vectors for utterances. To compare with the latest model (Shen et al., 2021) based on XLNet (Yang et al., 2019), we use the BERT model (Devlin et al., 2019) to extract context-independent utterance-level feature vectors for utterances. Let an utterance u consists of a sequence of tokens $x_1, x_2, \ldots, x_N$. First, a special token $[CLS]$ is appended at the beginning of the utterance to create the input sequence for the model: $[CLS], x_1, x_2, \ldots, x_N$. Then, we pass the $[CLS]$ appended utterances to BERT and extract out activations from the final four layers corresponding to the $[CLS]$ token. Finally, these four vectors are averaged to obtain the feature vector with a dimension of 768.

Discourse Parsing: To obtain discourse dependency trees, we utilize the discourse parser proposed by Shi and Huang (2019). It is a deep sequential model that achieves SOTA performance on the STAC corpus (Asher et al., 2016). We feed the conversations into the discourse parser:

$$\{(i, j, r_{ij}, p_{ij}), \ldots\} = \text{Parser}(u_1, \ldots, u_N). \quad (1)$$

The quadri-tuple (i, j, r_{ij}, p_{ij}) are directed edges of a discourse dependency tree with head i and tail j, indicating that u_i has immediate relation r_{ij} with u_j. And p_{ij} is the confidence score of the dependency link. Notice that $i, j = 1, 2, \ldots, N$ and $j > i$.

3.3 Model Overview

As illustrated in Fig. 2, there are three components in our proposed framework: (1) sequential context encoding; (2) discourse graph modeling; (3) emotion recognition. In the following sections, we explain each component in detail.

After the pre-processing, we obtain not only the dependency trees of conversations, but also the context-independent utterance-level feature vectors. Then, we use Bi-directional LSTM to transform these vectors into context-dependent ones. Next, a discourse-based graph stacks two different convolutional layers to aggregate contextual and speaker-specific information. Finally, the output feature vectors from the graph are used to recognize emotions for utterances.

3.4 Sequential Context Encoding

Similar to previous strategies, the sequential context encoder processes the constituent utterances

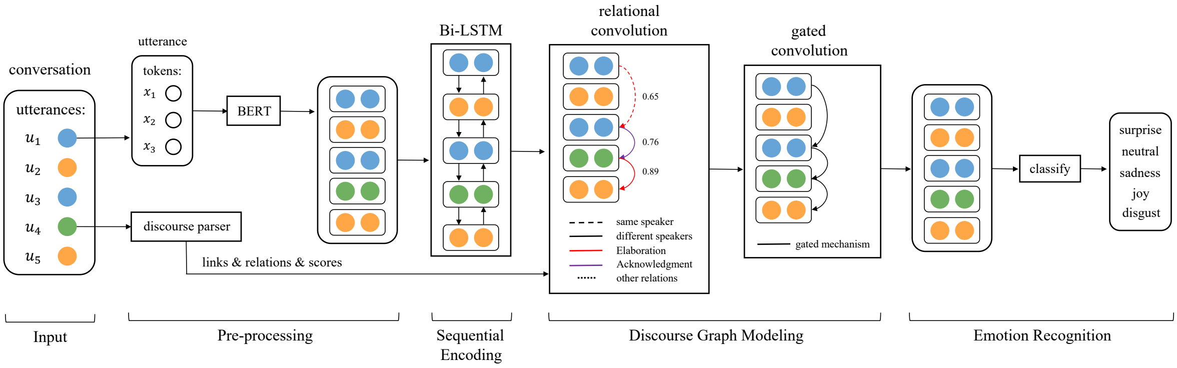


Figure 2: Overview of our proposed model for ERMC, congruent to the illustration in Section III. Different colors of utterances and edges respectively represent different speakers and different discourse relations.

in a conversation as a sequence according to the timeline. Inspired by Poria et al. (2017), we use Bi-directional LSTM to capture sequential context information,

$$g_i = \text{BiLSTM}\left(g_{i(+,-)1}, u_i\right), \tag{2}$$

where, $i = 1, 2, \ldots, N$, u_i and g_i are context-independent and sequential utterance representations, respectively.

3.5 Discourse Graph Modeling

Conversational discourse structures provide a straightforward way to capture both adjacent and distant cues for ERMC. Inspired by (Xia et al., 2019) and (Zhang et al., 2020), we exploit discourse dependency trees to construct graphs to propagate contextual and speaker-specific information. The framework is detailed here.

3.5.1 Graph Construction

First, we introduce the following notation: a multi-party conversation having N utterances is represented as a directed graph $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R}, \mathcal{W})$, with vertices/nodes $v_i \in \mathcal{V}$, labeled edges (relations) $e_{ij} \in \mathcal{E}$ where $r_{ij} \in \mathcal{R}$ is the relation type of the edge between v_i and v_j, and α_{ij} is the weight of the labeled edge e_{ij}, with $0 \leqslant \alpha_{ij} \leqslant 1$, where $\alpha_{ij} \in \mathcal{W}$ and $i, j = 1, 2, \ldots, N$. The graph is constructed based on discourse dependency trees in the following way,

Vertices: In the graph, each utterance within a multi-party conversation is represented as a vertex $v_i \in \mathcal{V}$. Each vertex v_i is initialized with the corresponding sequentially encoded representation g_i, and $i = 1, 2, \ldots, N$.

Edges: Construction of the edges $\mathcal{E}$ depends on discourse dependency trees. For instance, if there is a quadri-tuple (i, j, r_{ij}, p_{ij}) from a dependency tree, there would be an edge e_{ij} in the graph with head u_i and tail u_j. As the graph is directional, e_{ij} is not equivalent to e_{ji}. In most cases, an utterance only depends on its historical utterances, so the direction of edges is often directed as a topological sort from earlier utterances to later ones.

For speaker-specific information, Ghosal et al. (2019) model the emotional inertia of speakers in two-party conversations, where individual speakers resist the change of their own emotion against external influence. However, it is a challenge to incorporate discourse structures into speaker modeling for ERMC. In our model, we leverage the self-speaker dependency of interlocutors to model the emotional inertia of speakers by directly letting one utterance know whether its dependent utterance belongs to the same speaker. In Fig. 2, we use a dashed line to represent discourse dependencies between utterances from the same speaker and use a solid line to denote discourse dependencies between utterances from different speakers.

Edge Weights: The spatial graph convolutional operation essentially propagates node information along edges (Wu et al., 2020), thus proper edge weights is helpful. In our graph model, we set the edge weights statically,

$$\alpha_{ij} = p_{ij}, \tag{3}$$

where p_{ij} is the confidence score of edge e_{ij} obtained from the discourse parser.

Relations: The relation r_{ij} of an edge e_{ij} is set depending upon two aspects:

Discourse relations — These relations depend on discourse dependency trees. For example, r_{ij} is the discourse relation type of edge e_{ij} which is the dependency link between utterance u_i and u_j. According to (Shi and Huang, 2019), there are

16 types of discourse relations: *Comment, Clarification question, Elaboration, Acknowledgment, Continuation, Explanation, Conditional, Question-Answer pair (QAP), Alternation, Question-Elab(Q-Elab), Result, Background, Narration, Correction, Parallel and Contrast.*

Self-speaker dependency — This relation depends upon speakers. If two utterances are from the same speaker and have discourse relation r^q (r^q is one of the 16 discourse relations), we transform r^q into $r^{q'}$ to model the self-speaker dependency.

3.5.2 Feature Transformation

We now describe the methodology to transform the sequentially encoded feature vectors using the graph network. After a two-step graph convolution process, the vertex representations g_i are transformed into contextual and speaker-specific ones.

In the first step, we consider discourse dependencies as important cues to propagate contextual and speaker-specific information. As there are many types of edges, inspired by Schlichtkrull et al. (2018), the new features h_i^1 of utterance u_i is computed as:

$$h_i^1 = \sigma(W_0^1 g_i + \sum_{r \in \mathcal{R}} \sum_{j \in N_i^r} \frac{\alpha_{ij}}{c_{i,r}} W_r^1 g_j), \quad (4)$$

where, α_{ij} is edge weight, N_i^r represents the neighboring indices of node g_i under relation $r \in \mathcal{R}$. And $c_{i,r}$ is a problem specific normalization constant which is set in advance ($c_{i,r} = |N_i^r|$). σ is an activation function such as ReLU, W_0^1 and W_r^1 are trainable parameters, only edges of the same relation type r are associated with the same projection weight W_r^1.

In the second step, to select more informative cues from dependent utterances, another residual gated convolutional layer (Bresson and Laurent, 2018) is applied over the output of the first step,

$$h_i^2 = \sigma(W_0^2 h_i^1 + \sum_{j \in N_i^r} \eta_{i,j} \odot W_1^2 h_j^1), \quad (5)$$

$$\eta_{i,j} = \text{sigmoid}(W_2^2 h_i^1 + W_3^2 h_j^1), \quad (6)$$

where W_0^2, W_1^2, W_2^2, and W_3^2 are trainable parameters. This stack of graph convolutional layers effectively aggregates normalized contextual and speaker-specific information of the neighborhood for each utterance in the graph.

3.6 Emotion Recognition

After the feature transformation, we consider h_i^2 as the contextual and speaker-specific representations of utterances. Then, we classify each utterance using a fully connected network:

$$\mathcal{P}_i = \text{softmax}(W_{smax} h_i^2 + b_{smax}), \quad (7)$$

$$\hat{y}_i = \underset{k}{\text{argmax}}\left(\mathcal{P}_i[k]\right). \quad (8)$$

To train the model, we choose the cross-entropy loss function:

$$\mathcal{L} = -\sum_{v \in y_{\mathcal{V}}} \sum_{z=1}^{Z} Y_{vz} \ln \mathcal{P}_{vz}, \quad (9)$$

where $y_{\mathcal{V}}$ is the set of node indices that have labels and Y is the label indicator matrix.

Dataset	Conversations			Utterances		
	Train	Val	Test	Train	Val	Test
MELD	1038	114	280	9989	1109	2610
EmoryNLP	713	99	85	9934	1344	1328
IEMOCAP	120		31	5810		1623

Table 1: The statistics of three datasets

4 Experimental Setting

4.1 Datasets

To verify the effectiveness of integrating discourse structures for ERMC, we evaluate our model on both multi-party and two-party conversation corpora. All these datasets contain multimodal information for each utterance within a conversation, while we only focus on the textual information in this work. Table 1 shows the corpora statistics.

MELD (Poria et al., 2019a): A multi-party conversation corpus collected from the TV show *Friends*. Each utterance is annotated as one of the seven emotion classes: *neutral, surprise, fear, sadness, joy, disgust*, and *anger.*

EmoryNLP (Zahiri and Choi, 2018): A multi-party conversation corpus collected from *Friends*, but varies from MELD in the choice of scenes and emotion labels. The emotion labels include *neutral, joyful, peaceful, powerful, scared, mad,* and *sad.*

IEMOCAP (Busso et al., 2008): A two-party conversation corpus. The emotion labels include *neutral, happiness, sadness, anger, frustrated,* and *excited.* Since this dataset has no validation set, we follow (Shen et al., 2021) to use the last 20 dialogues in the training set for validation.

4.2 Implementation Details

We use pre-trained BERT-Base[1] to encode utterances and adopt Adam (Kingma and Ba, 2015) as the optimizer with an initial learning rate of 1e-4 and L2 weight decay of 1e-5 for three datasets. The batch size is set to be {32,32,16} for MELD, EmoryNLP, and IEMOCAP respectively. The dimensions of g_i, h_i^1 and h_i^2 are set to be 100, 64, and 64. The dropout (Srivastava et al., 2014) is set to be 0.5. We train all models for a maximum of 100 epochs and stop training if the validation loss does not decrease for 20 consecutive epochs.

4.3 Baseline Methods

For a comprehensive evaluation of our proposed ERMC-DisGCN, we compare it with the following baseline methods:

cLSTM (Poria et al., 2017): Contextual utterance representations are generated by capturing the content from surrounding utterances using a Bi-directional LSTM network.

DialogueRNN (Majumder et al., 2019): It is a recurrent network that uses three GRUs to track individual speaker states, global context, and emotional state within conversations.

HiGRU (Jiao et al., 2019): It is a hierarchical GRU structure that trains utterance-level and conversation-level encoders jointly.

ConGCN (Zhang et al., 2019): This model represents the entire conversational corpus as a large heterogeneous graph to capture context-sensitive and speaker-sensitive features.

DialogueGCN (Ghosal et al., 2019): This is a graph-based model to encode speaker dependencies and temporal information within a window context.

KET (Zhong et al., 2019): Enriched by the external commonsense knowledge, KET employs the Transformer encoder and decoder (Vaswani et al., 2017) for ERC.

BERT-MTL (Li et al., 2020): It is a multi-task learning framework where features extracted from BERT are used for emotion recognition and speaker identification.

DialogueXL (Shen et al., 2021): An all-in-one XLNet model with dialog-aware self-attention to deal with multi-party structures.

BERT-LSTM: A variation of cLSTM where the CNN-based utterance-level feature vectors are replaced by our BERT-based feature vectors. We consider this model as our strong baseline.

Model	**Multi-party**		**Two-party**
	MELD	EmoryNLP	IEMOCAP
cLSTM	56.44	32.89	54.95
DialogueRNN	57.03	31.27	62.75
HiGRU	56.92	31.88	59.79
ConGCN	57.40	33.52^*	-
DialogueGCN	58.10	33.85^*	64.18
KET	58.18	33.95	59.56
BERT-MTL	61.90	34.85	-
DialogueXL	62.41	34.73	**65.94**
BERT-LSTM	62.34	34.66	63.10
ERMC-GCN	62.71	34.97	63.68
ERMC-DisGCN	**64.22**	**36.38**	64.10

Table 2: Overall performance on both multi-party and two-party conversation corpora, which is statistically significant under the paired t-test (p<0.05). We use the average F1 score to evaluate each model. The scores marked by "*" are based on our re-implementation, because of the differences in datasets between the corresponding work and ours.

ERMC-GCN: A variation of our approach where the graph modeling is based on the timeline of conversations. It means that there are no discourse structures in this model.

5 Results and Discussions

5.1 Comparison with Baseline Methods

We compare the performance of our proposed ERMC-DisGCN framework with multiple baselines in Table 2. To verify the effectiveness of integrating discourse structures for ERMC, we conduct experiments on both multi-party and two-party conversation datasets.

MELD and EmoryNLP: On these multi-party conversation datasets, we first report our baseline results which achieve comparable performance with the previous systems. Then, our proposed ERMC-DisGCN achieves average F1 scores of 64.22% and 36.38%, which are around 2% better than the strong baseline. Compared to ERMC-GCN, integrating discourse structures leads to F1 improvements of around 1.5% on two datasets. We attribute this gap in performance to the nature of conversations. There are many utterances, like *"yeah"*, *"okay"*, and *"no"*, that can express different emotions depending on the context within conversations. In these cases, discourse structures indicate the most informative historical utterances, which contributes to emotion recognition.

IEMOCAP: On this two-party conversation dataset, we observe the inferior performance of our

[1] `https://github.com/google-research/bert,BERT-Base,Uncased`

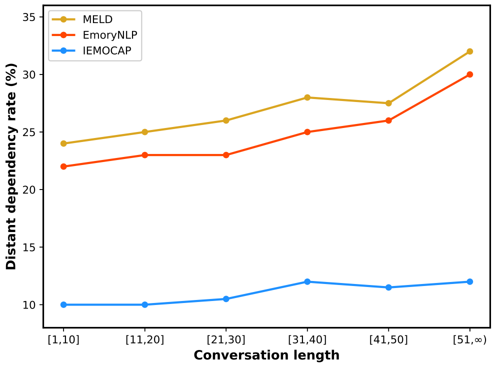


Figure 3: The discourse dependency rate between distant utterances on three datasets.

baseline BERT-LSTM to dialogueXL. The average conversation length is 50 utterances in IEMOCAP which is much longer than MELD and EmoryNLP, so LSTM fails to propagate rich long-term information, while DialogueXL remains the SOTA result with enhanced memory for historical context. And compared to ERMC-GCN, integrating discourse structures only leads to an F1 score increase of 0.42%. In the following section, we explain the reason for different performance of integrating discourse structures in these datasets.

5.2 Multi-Party vs Two-Party

According to those results shown in Table 2, we find that integrating discourse structures in multi-party conversations leads to more significant improvements than in two-party conversations. To explain this difference, it is important to understand the nature of multi-party and two-party conversations. After examining the datasets, we report the distant dependency rate of them in Fig. 3. As we can see, discourse structures in multi-party conversations are much more complex. About 25% utterances have discourse relations with distant ones in multi-party conversations, and this rate rises as conversation length increases. In MELD and EmoryNLP, there are often more than 5 interlocutors within a conversation, thus speakers' turns change quickly and one speaker may respond to another after many turns. However, in two-party conversations, the distant dependency rate is only around 11% and keeps steady when conversation length increases. Since there are only two interlocutors, they tend to speak utterances cyclically, adjacent utterances are more related. From the above discussion, we can conclude that it is more necessary to exploit discourse structures to handle the rich dependencies between distant utterances in multi-party conversations.

Speaker modeling method	Average F1 score	
	MELD	EmoryNLP
ours (based on discourse)	**64.22**	**36.38**
ours (independent of discourse)	63.69	36.02
speaker-specific GRUs	63.74	36.07
speaker role embedding	63.79	35.98

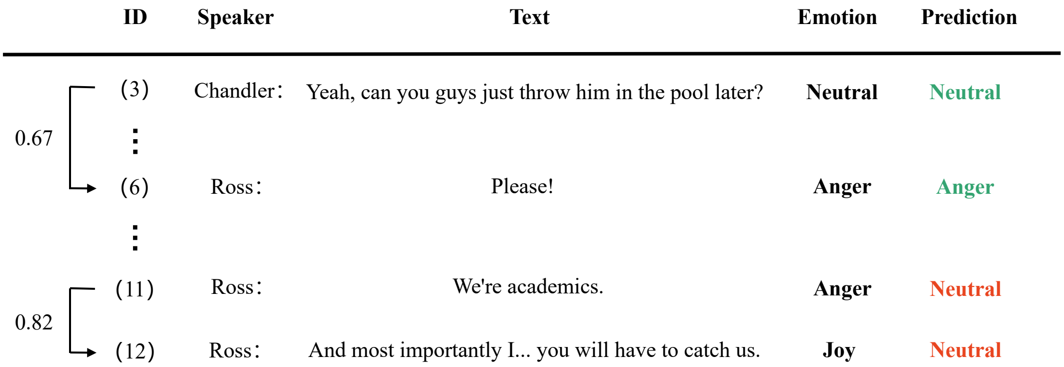
Table 3: Results of comparison between four speaker modeling approaches on the MELD and EmoryNLP datasets.

5.3 Different Speaker Modeling Methods

Previous studies have proven that capturing speaker-specific features benefits emotion recognition in conversation. In this section, we conduct experiments to answer the following two questions: (1) Is it helpful to propagate speaker information based on discourse structures? (2) Which speaker modeling method contributes most to our approach?

We replace our self-speaker dependency modeling method with the following three methods. The first one is a variation of ours that the self-speaker dependency is modeled independently of discourse structures by directly letting one utterance know whether the adjacent one is from the same speaker. The second method is to use speaker-specific GRUs (Hazarika et al., 2018) to process the histories of each speaker which represent the individual states of speakers. The third one is speaker role embedding, which maps each interlocutor to a trainable vector (Zhang et al., 2019). These methods are all independent of discourse structures but capture different speaker-specific features.

The results of different speaker modeling methods are shown in Table 3. We observe that the discourse-based self-speaker modeling method performs better than the independent method. This gap supports our hypothesis that the discourse dependencies between distant utterances offer informative cues for capturing speaker-specific features. So, it is necessary to integrate discourse structures into speaker modeling. Besides, although the other two methods capture different kinds of speaker-specific features, they have similar performance with our independent model.



	ID	Speaker	Text	Emotion	Prediction
0.67	(3)	Chandler:	Yeah, can you guys just throw him in the pool later?	Neutral	Neutral
	⋮				
	(6)	Ross:	Please!	Anger	Anger
	⋮				
0.82	(11)	Ross:	We're academics.	Anger	Neutral
	(12)	Ross:	And most importantly I... you will have to catch us.	Joy	Neutral

Figure 4: Results of case study, where two utterances from a conversation are provided, along with their dependent historical utterances. We use green and red to highlight right and wrong predictions. The confidence scores of two dependency links are shown in the left.

Method	Average F1 score	
	MELD	EmoryNLP
ERMC-DisGCN	**64.22**	**36.38**
- self-speaker dependency	63.45(↓ 0.77)	35.88(↓ 0.50)
- gated convolution	63.67(↓ 0.55)	35.89(↓ 0.49)
- relational convolution	63.01(↓ **1.21**)	35.41(↓ **0.97**)

Table 4: Results of ablation study on MELD and EmoryNLP.

5.4 Ablation Study

We perform an ablation study for three components of our model by removing them one by one at a time. Experimental results are shown in Table 4. First, we find that the self-speaker dependency is of significance in our model. This phenomenon is in tune with previous works that capturing speaker-specific features benefits emotion recognition in multi-party conversation, where there are often more than 5 interlocutors. By eliminating the gated convolutional layer in the graph, our model falls by 0.55% on MELD and 0.49% on EmoryNLP. Discourse structures only offer contextual cues, not all information from dependent utterances helps emotion recognition. Therefore, this gated convolutional layer is necessary to select informative cues in our graph modeling. Further, the relational convolutional layer successfully aggregates contextual and speaker-specific information from the neighborhood of each utterance according to edge types and makes the most contribution to our approach.

5.5 Case Study

For a comprehensive understanding of our proposed method, we visualize its performance by a case study, which is selected from the MELD test dataset. As illustrated in Fig. 4, utterance **(6)** is too short to carry rich semantic features for emotion recognition. However, its dependent utterance **(3)** offers an informative cue and helps make the right prediction. From the ablation study, we draw the conclusion that modeling the self-speaker dependency benefits ERMC, but it is not always the case. For instance, we observe two wrong predictions for the adjacent utterances **(11)** and **(12)**, which are from the same speaker and have a discourse relation. Modeling the self-speaker dependency is hard to deal with the emotional shifts (i.e., the emotion labels of two consecutive utterances from the same speaker are different) (Poria et al., 2019a; Shen et al., 2021). Roughly, our model commits mistakes for 40% of similar cases, which calls for further investigations.

6 Conclusion

In this paper, we investigate the importance of discourse structures in handling informative contextual cues and speaker-specific features for ERMC. We propose a discourse-aware graph neural network and devise two graph convolutional layers to aggregate normalized contextual and speaker-specific information for each utterance in the graph. Experimental results show that our proposed model outperforms all the baselines on all multi-party conversation datasets. Furthermore, we apply extensive analyses for the proposed model and have the following findings. First, discourse structures are more helpful for emotion recognition in multi-party conversation than in two-party conversation. Second, it is important to integrate discourse structures into speaker modeling. Third, the gated mechanism helps select more informative cues from dependent utterances for ERMC.

In our future work, we would like to capture various speaker-specific features to deal with the emotional shifts. Since our method focuses on using explicit discourse structures, we also plan to employ implicit methods to avoid error propagation and address consequent issues.

Acknowledgments

We thank all reviewers for their valuable comments. This work was supported by National Natural Science Foundation of China under grants 62076173, 61672211 and U1836222.

References

Nicholas Asher, Julie Hunter, Mathieu Morey, Farah Benamara, and Stergos Afantenos. 2016. Discourse structure and dialogue acts in multiparty dialogue: the stac corpus. In *10th International Conference on Language Resources and Evaluation (LREC 2016)*, pages 2721–2727.

Parminder Bhatia, Yangfeng Ji, and Jacob Eisenstein. 2015. Better document-level sentiment analysis from RST discourse parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 2212–2218, Lisbon, Portugal. Association for Computational Linguistics.

Xavier Bresson and Thomas Laurent. 2018. Residual gated graph convnets.

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

Erik Cambria, Soujanya Poria, Alexander Gelbukh, and Mike Thelwall. 2017. Sentiment analysis is a big suitcase. *IEEE Intelligent Systems*, 32(6):74–80.

Junyoung Chung, Caglar Gulcehre, Kyunghyun Cho, and Yoshua Bengio. 2014. Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Deepanway Ghosal, Navonil Majumder, Soujanya Poria, Niyati Chhaya, and Alexander 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)*, pages 154–164, Hong Kong, China. Association for Computational Linguistics.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018. ICON: Interactive conversational memory network for multimodal emotion detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2594–2604, Brussels, Belgium. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Qi Jia, Yizhu Liu, Siyu Ren, Kenny Zhu, and Haifeng Tang. 2020. Multi-turn response selection using dialogue dependency relations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1911–1920, Online. Association for Computational Linguistics.

Wenxiang Jiao, Haiqin Yang, Irwin King, and Michael R Lyu. 2019. Higru: Hierarchical gated recurrent units for utterance-level emotion recognition. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 397–406.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1746–1751, Doha, Qatar. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *International Conference on Learning Representations*.

Mathias Kraus and Stefan Feuerriegel. 2019. Sentiment analysis based on rhetorical structure theory: Learning deep neural networks from discourse trees. *Expert Systems with Applications*, 118:65–79.

Jiaqi Li, Ming Liu, Zihao Zheng, Heng Zhang, Bing Qin, Min-Yen Kan, and Ting Liu. 2021. Dadgraph: A discourse-aware dialogue graph neural network for multiparty dialogue machine reading comprehension. *arXiv preprint arXiv:2104.12377*.

Jingye Li, Meishan Zhang, Donghong Ji, and Yijiang Liu. 2020. Multi-task learning with auxiliary speaker identification for conversational emotion recognition. *arXiv e-prints*, pages arXiv–2003.

Jiwei Li, Rumeng Li, and Eduard Hovy. 2014. Recursive deep models for discourse parsing. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2061–2069, Doha, Qatar. Association for Computational Linguistics.

Navonil Majumder, Soujanya Poria, Devamanyu Hazarika, Rada Mihalcea, Alexander Gelbukh, and Erik Cambria. 2019. Dialoguernn: An attentive rnn for emotion detection in conversations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 6818–6825.

Joscha Märkle-Huß, Stefan Feuerriegel, and Helmut Prendinger. 2017. Improving sentiment analysis with document-level semantic relationships from rhetoric discourse structures. In *Proceedings of the 50th Hawaii International Conference on System Sciences*.

Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. 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)*, pages 873–883, Vancouver, Canada. Association for Computational Linguistics.

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 Annual Meeting of the Association for Computational Linguistics*, pages 527–536, Florence, Italy. Association for Computational Linguistics.

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

Michael Schlichtkrull, Thomas N Kipf, Peter Bloem, Rianne Van Den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *European Semantic Web Conference*, pages 593–607. Springer.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021. Dialogxl: All-in-one xlnet for multiparty conversation emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Zhouxing Shi and Minlie Huang. 2019. A deep sequential model for discourse parsing on multi-party dialogues. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7007–7014.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *The journal of machine learning research*, 15(1):1929–1958.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Zonghan Wu, Shirui Pan, Fengwen Chen, Guodong Long, Chengqi Zhang, and S Yu Philip. 2020. A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*.

Qingrong Xia, Zhenghua Li, Min Zhang, Meishan Zhang, Guohong Fu, Rui Wang, and Luo Si. 2019. Syntax-aware neural semantic role labeling. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 7305–7313.

Zhilin Yang, Zihang Dai, Yiming Yang, Jaime Carbonell, Russ R Salakhutdinov, and Quoc V Le. 2019. Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in Neural Information Processing Systems*, 32:5753–5763.

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*, pages 44–52.

Bo Zhang, Yue Zhang, Rui Wang, Zhenghua Li, and Min Zhang. 2020. Syntax-aware opinion role labeling with dependency graph convolutional networks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3249–3258, Online. Association for Computational Linguistics.

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 *IJCAI*, pages 5415–5421.

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)*, pages 165–176, Hong Kong, China. Association for Computational Linguistics.