Beyond Linguistic Cues: Fine-grained Conversational Emotion Recognition via Belief-Desire Modelling

Bo Xu[1], **Longjiao Li**[1], **Wei Luo**[2], **Mehdi Naseriparsa**[3], **Zhehuan Zhao**[1], **Hongfei Lin**[4], **Feng Xia**[5]
[1]School of Software, Dalian University of Technology, Dalian, China
[2]School of Information Technology, Deakin University, Geelong, Australia
[3]Federation University Australia, Ballarat, Australia
[4]School of Computer Science and Technology, Dalian University of Technology, Dalian, China
[5]School of Computing Technologies, RMIT University, Melbourne, Australia
{boxu, lymc, z.zhao, hflin}@dlut.edu.cn
wei.luo@deakin.edu.au, m.naseriparsa@federation.edu.au, f.xia@ieee.org

Abstract

Emotion recognition in conversation (ERC) is essential for dialogue systems to identify the emotions expressed by speakers. Although previous studies have made significant progress, accurate recognition and interpretation of similar fine-grained emotion properly accounting for individual variability remains a challenge. One particular under-explored area is the role of individual beliefs and desires in modelling emotion. Inspired by the Belief-Desire Theory of Emotion, we propose a novel method for conversational emotion recognition that incorporates both belief and desire to accurately identify emotions. We extract emotion-eliciting events from utterances and construct graphs that represent beliefs and desires in conversations. By applying message passing between nodes, our graph effectively models the utterance context, speaker's global state, and the interaction between emotional beliefs, desires, and utterances. We evaluate our model's performance by conducting extensive experiments on four popular ERC datasets and comparing it with multiple state-of-the-art models. The experimental results demonstrate the superiority of our proposed model and validate the effectiveness of each module in the model.

Keywords: Emotion Recognition in Conversation, The Belief-Desire Theory of Emotion, Fine-grained Emotion Recognition

1. Introduction

As humans' central communication medium, natural language plays a pivotal role in the process of emotion generation. Recently, in order to develop artificial intelligence capable of understanding human emotions, emotion recognition in conversation (ERC) has become a hot research field. ERC aims to identify the emotions of each utterance in a conversation, which contributes to generating emotion-aware dialogues and develops empathic conversation agents or chatbots for psychotherapy (Sharma et al., 2021; Xu and Zhuang, 2022).

The task of conversational emotion recognition is different from the semantic or sentiment analysis tasks of traditional texts such as sentences and documents (Naseriparsa et al., 2019a; Lin and Joe, 2023), which requires not only the semantic information of the utterance itself but also the context modeling of each utterance. Previous studies on conversational emotion recognition adopted sequence-based (Majumder et al., 2019; Ghosal et al., 2020) or graph-based (Ghosal et al., 2019; Ishiwatari et al., 2020) approaches to model the context of utterances and the interaction between speakers as much as possible. Recently, multiple methods (Zhao et al., 2022; Yi et al., 2022; Zhang et al., 2023) advocate introducing commonsense knowledge or utilizing specific prompts into emotion recognition in conversation. They utilize commonsense knowledge to model the speaker's needs and intentions in utterances to assist the judgment of emotion types. However, how to accurately identify similar fine-grained emotions remains a challenge. For example, many previous research methods have difficulty identifying whether the emotion of an utterance in a dialogue is angry or disgusted. In particular, solely relying on the superficial linguistic cues, these methods are unable to capture individual-level cognitive processes underlying emotions.

To address this challenge, we turn to the Belief-Desire Theory of Emotion (BDTE) (Reisenzein, 2022, 2021), which provides a psychological framework to understand emotion via complex interplay between individual beliefs and desires. This means that the same desire can lead to different emotions due to varying cognitive beliefs. For example in Figure 1, George thinks Rowan's team is strong, but Mary thinks Rowan has a tough team to compete against. Moreover, both George and Mary want Rowan to win the soccer game. Although George and Mary have the same desire, they have different beliefs. When they finally learn that Rowan has lost, George and Mary have diverse emotions, George is disappointed, while Mary is sad. Therefore, it

LREC-COLING 2024, pages 2318–2328
20-25 May, 2024.

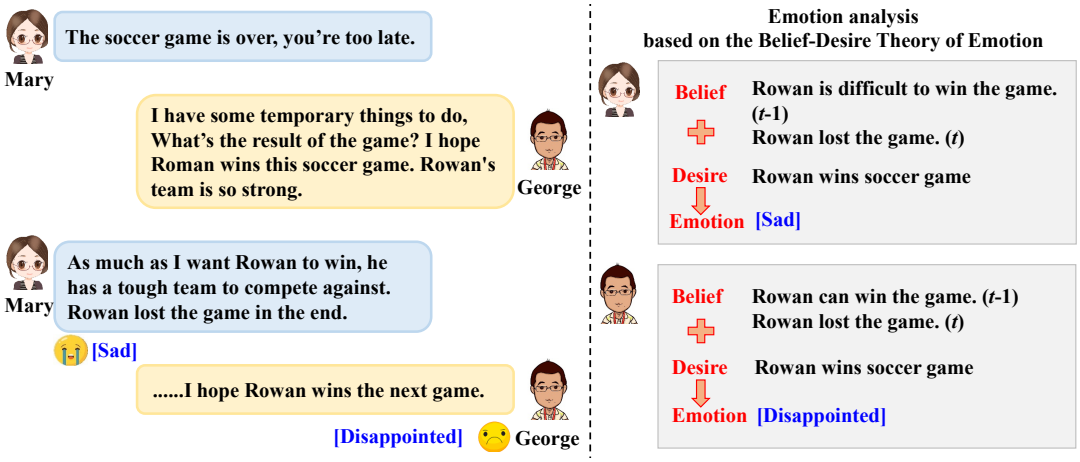

Figure 1: An example showing how emotions are generated according to the Belief-Desire Theory of Emotion.

is difficult to accurately judge emotions solely by considering a person's needs while ignoring their cognitive belief. According to previous research about emotion recognition in conversation, George and Mary will have the same emotion, but the Belief-Desire Theory of Emotion can more accurately deduce that they actually have diverse emotions.

As BDTE is largely unfamiliar to the ERC community, in this paper, we study how this theory can be applied to comprehensively model inference of beliefs and desires to achieve more accurate fine-grained emotion recognition. According to BDTE, beliefs and desires are semantically related to emotions as they both involve emotion-eliciting events. We propose a heterogeneous conversation graph to encode emotion-belief-desire relationships. Our graph consists of three types of nodes: utterances, speakers, and emotion-eliciting events. To effectively capture the context, speaker's state, and beliefs and desires, we employ Graph Transformer Network to facilitate information transfer between these heterogeneous nodes. Specifically, we model the speaker's desires through the relationship between utterances and emotion-eliciting events, and the speaker's beliefs through the relationship between the emotion-eliciting events. In summary, the main contributions of this paper are as follows:

- We introduce the Belief-Desire theory to the emotion recognition in conversation (ERC) task, for more nuanced and accurate modelling of individual emotion.
- We constructed an emotion-belief-desire conversation graph to support knowledge representation and inference based on key principles in BDTE. The graph captures the utterance context, speaker's global state, and supports the inference of beliefs and desires.
- We extensively evaluated our method on four commonly used ERC datasets, demonstrating its superiority and effectiveness.

2. Related work

In conversation emotion recognition, a particular focus is context modeling of utterances. Most models are either sequence-based or graph-based. Sequence-based approaches(Jiao et al., 2019; Majumder et al., 2019; Hu et al., 2021) take previous utterances into account and utilize sequential information by chronologically or sequentially encoding utterances. These methods tend to update the state of the current utterance by obtaining only relatively limited information from the recent utterance. Graph-based models (Ghosal et al., 2019; Ishiwatari et al., 2020; Shen et al., 2021b) encode utterances as nodes in a graph and use edges to exchange information among the utterance nodes. These methods simultaneously collect surrounding utterances within a specific window for context modeling while they ignore distant utterances and sequence information.

Recently, commonsense knowledge bases (Speer et al., 2017; Sap et al., 2019) have been applied to better understand emotions in conversation. KET (Zhong et al., 2019) is the first model that integrates commonsense knowledge with emotional information from the conversational text. COSMIC (Ghosal et al., 2020) is a commonsense guided framework for emotion recognition in conversation that captures some of the complex interactions between personalities, events, mental states, intentions, and emotions. SKAIG (Li et al., 2021) proposed a knowledge perception interaction graph that contains four relationships to simulate the speaker's mental state. CauAIN (Zhao et al., 2022) focused on retrieving

causal clues from commonsense knowledge to identify emotional causes in conversations. CISPER (Yi et al., 2022) utilized contextual information and commonsense prompts related to the interlocutor's utterances. MPLP (Zhang et al., 2023) used history-oriented and experience-oriented prompts to mimic the thinking process. These models aimed to improve emotion recognition by incorporating commonsense knowledge or specific prompts. However, no existing models explicitly leverage mental state inference for ERC.

3. Belief-Desire Theory of Emotion

The Belief-Desire Theory of Emotion emphasises that emotion is not solely driven by desire/motivation, but also by how one perceives the situation.

3.1. Computational Belief-Desire Theory of Emotion

We present the qualitative belief-desire analysis of emotions in Table 1 by combining the Computational Belief-Desire Theory of Emotion (Reisenzein, 2009) into the conversational emotion recognition task. The following is an illustration of the emotion analysis in Table 1 combined with the example in Figure 1.

If George wants Rowan to win the game and he is certain that Rowan wins, then he would be **happy**. If George wants Rowan to not win but he is certain that Rowan actual wins, then he would be **sad**. If George wants Rowan to win but he is not certain about that, then he would be **excited**. If George wants Rowan to not win the game and he is not sure about that, then he would be **fearful**. If George originally believed that Rowan did not win but is then told otherwise, he would be **surprised**. If George wants Rowan to win and believed so, then upon knowing that Rowan did not win, he would be **disappointed**. However in the previous case, if George wants Rowan not to win, then he would be **peaceful**.

3.2. Belief-Desire Representation Inference in Conversational Emotion Recognition

Psychologist Rainer Reisenzein's research (Reisenzein, 2010; Reisenzein and Junge, 2012; Reisenzein, 2012) on emotion detection proposed that reasoning belief and desire by emotion-eliciting events significantly improve the emotion reasoning ability of AI. Therefore, we utilize the Belief-Desire Theory of Emotion to identify the emotion of utterances in dialogue. In conversational emotion recognition, we first extract the emotion-eliciting events. Then, we analyze the emotion in utterances based on cognitive belief and desire reasoning. We infer the speaker's desire from the relationship between utterance and emotion-eliciting events. Furthermore, we infer the speaker's belief from the relationship between the emotion-eliciting events.

Emotion	if	Belief at t	Desire at t	Belief at $t-1$
Happy/Joyful		Certain(e, t)	Des(e, t)	
Sad		Certain(e, t)	Des($\neg e$, t)	
Excited		Uncertain(e, t)	Des(e, t)	
Feared/Scared		Uncertain(e, t)	Des($\neg e$, t)	
Surprised		Certain(e, t)	–(irrelevant)	Bel($\neg e$, $t-1$)
Disappointed		Certain($\neg e$, t)	Des(e, t)	Bel(e, $t-1$)
Peaceful		Certain($\neg e$, t)	Des($\neg e$, t)	Bel(e, $t-1$)

Table 1: A qualitative formulation of Belief–Desire Theory. See more details in Section 3.1.

4. Methodology

The overall framework of our proposed approach is shown in Figure 2. The framework modules are as follows: 1) Feature Extraction; 2) Conversation Graph Encoder; 3) Emotion Classification.

In ERC, a typical conversation consists of a sequence of utterances $[U_1, U_2, \ldots, U_i, \ldots, U_N]$. An utterance U_i consists of a sequence of words $[w_1, w_2, \ldots, w_{L_i}]$. Each utterance U_i is made by a speaker S. The goal of ERC is to identify the correct emotion label y_i for each utterance U_i, where $y_i \in (Y_1, Y_2, ..., Y_K)$, a predefined set of emotion labels.

4.1. Feature Extraction

In the feature extraction module, we extract the utterance-level features and emotion-eliciting events within the utterances and use pre-training language model to represent the semantic feature representation of emotion-eliciting events.

4.1.1. Utterance-level Feature Extraction

We utilize the pre-trained language model RoBERTa-Large (Liu et al., 2019) for context-independent utterance-level feature extraction. We fine-tune the RoBERTa-Large for the emotion classification task on each ERC dataset and freeze its parameters for subsequent model training. Specifically, for each utterance $U_i = [w_1, w_2, ..., w_{L_i}]$, we append a special token [CLS] to its beginning so that the model input is sequenced as $\{[\text{CLS}], w_1, w_2, ..., w_{L_i}]\}$ and fed into Roberta-Large:

$$h_i^u = RoBerta([CLS], w_1, w_2, ..., w_{L_i}). \quad (1)$$

obtain [CLS]'s pooled embedding of last layer hidden-state h_i^u as context independent utterance-level feature of U_i.

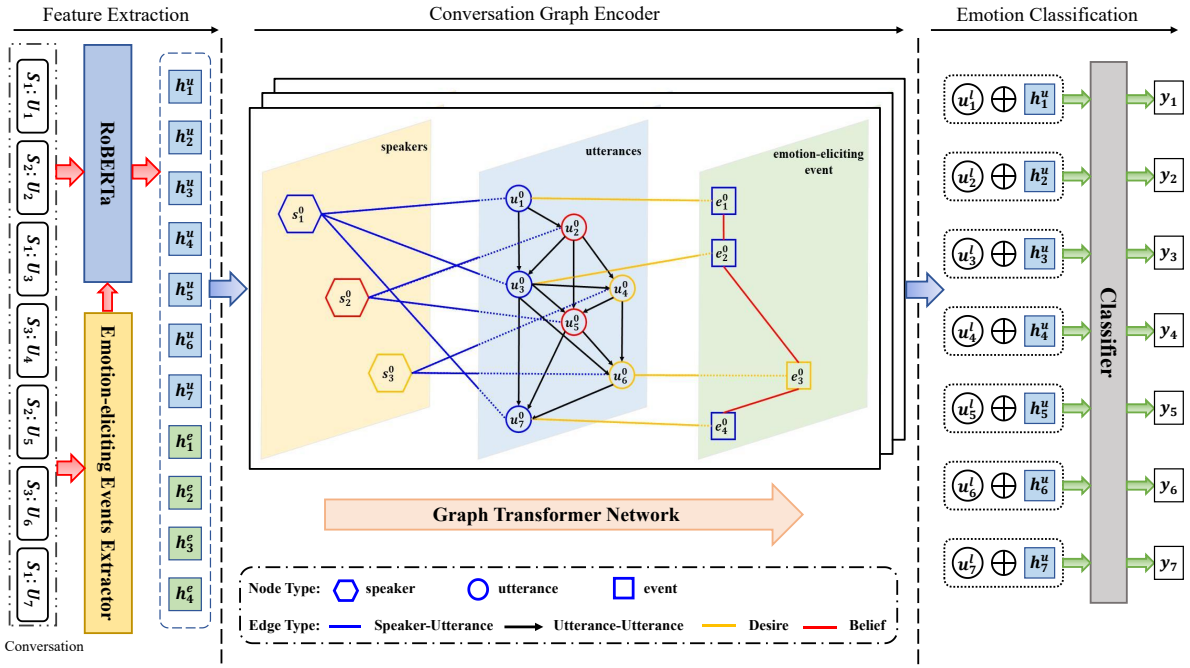

Figure 2: Our framework consists of three components: a feature extractor, a conversation graph encoder, and an emotion classifier.

4.1.2. Emotion-eliciting Event Feature Extraction

As proposed by (Reisenzein and Junge, 2012), verbal communication is the primary source required to calculate the beliefs and desires for specific emotion-eliciting events. Since beliefs and desires are semantically related to emotion-eliciting events, we extract emotion-eliciting events from target utterances to judge the confirmation or disconfirmation of beliefs and to determine the satisfaction or frustration of desires. Referring to eventuality definition and extraction method proposed by ASER (Zhang et al., 2020b, 2022), we define emotion-eliciting event e is a combination of multiple words $v_1, ..., v_N$, where N is the number of words in emotion-eliciting event E, here, $v_1, ..., v_N \in V$ are in the utterance. A pair of words in $e(v_i, v_j)$ may follow a syntactic relation.

In order to avoid the extracted emotion-eliciting events being too sparse and to ensure that emotion-eliciting events have complete semantics, we extract the necessary words to construct emotion-eliciting events from the utterances according to the eventuality patterns in ASER (Zhang et al., 2020b, 2022). Specifically, we first use the discourse parsing system (Wang and Lan, 2015) to parse the utterances and then use the 18 event modes (such as subject-verb-object) proposed by ASER to match and obtain the emotion-eliciting events, that is

$$E = Matching(Parsing(U_i)). \quad (2)$$

where the function $Parsing()$ returns dependency graph by syntactic parsing, the function $Matching()$ returns matched emotion-eliciting events set.

Then, we adopt RoBERTa to encode the extracted events, obtain the last hidden state, and deploy a max-pooling operation to get the semantic representation of emotion-eliciting events:

$$h_i^e = Maxpooling(RoBERTa(e_i)), e_i \in E. \quad (3)$$

4.2. Conversation Graph Encoder

In the Conversation Graph Encoder module, we first build an Emotion-belief-desire Conversation Graph. Then, we adopt the Graph Transformer network to extract the conversation-level features of the utterances.

4.2.1. Emotion-belief-desire Conversation Graph Construction

Based on the Belief-Desire Theory of Emotion, we construct a heterogeneous graph G. Graph G models context of utterances, speakers-global state, and inference of belief and desire for each utterance U_i in the dialogue, where V denotes the set of nodes and E denotes the set of edges.

Our graph G contains three types of nodes:

Utterance node: We consider each utterance in the conversation as a node u, whose features are initialized with its utterance-level feature. $u_i^0 = h_i^u$.

Speaker node: We treat each speaker in the dialogue as a node s. We initialize the speaker node features by the average of the semantic features

in all utterances expressed by this speaker in the dialogue. $s_j^0 = avg(h_i^u)$, $\forall U_i$ spoken by S_j.

Emotion-eliciting event node: We treat the emotion-eliciting events extracted from the utterances as emotion-eliciting event nodes and initialize the node features by semantic representations of the events, $e_k^0 = h_k^e$.

The set of nodes can be represented as:

$$V = V_u \cup V_s \cup V_e. \tag{4}$$

where utterance node set $V_u = \{u_i\}$, speaker node set $V_s = Unique(\{s_j\})$, emotion-eliciting event node set $V_e = \{e_k\}$, the function $Unique()$ returns all unique elements of the set.

Our graph G contains four types of edges:

Utterance-utterance edge: We connect the current utterance to the last utterances spoken by all speakers before it, which point to the current target utterance. We believe that the last utterance of each speaker before the target utterance (including the speaker of the target utterance) has the most significant impact on the context of the target utterance, and the rest have a less significant impact. In addition, it is worth noting that the edges between utterances are unidirectional. The edges between utterances model the impact of past utterances on the current utterance, namely context modeling. These edges are denoted as $E_{uu} = \{(u_i, u_t)\}, t > i$.

Speaker-utterance edge: We connect each target utterance to the corresponding speaker. The edges between the utterances and the speaker model the speaker's global state on the utterances, and are denoted as $E_{su} = \{(s_j, u_i)\} \cup \{(u_i, s_j)\}$.

Desire edge: We connect each target utterance to the emotion-eliciting event extracted from the utterance. The edges between utterances and emotion-eliciting events are denoted as $E_{eu} = \{(e_k, u_i)\} \cup \{(u_i, e_k)\}$. Desire edges are connected with emotion-eliciting events nodes via the utterance nodes to capture the speaker-specific desires and their influence on emotion-eliciting events.

Belief edge: According to the order in which the utterances and emotion-eliciting events are extracted, we successively connect the emotion-eliciting event nodes extracted from the utterances. Belief edges are denoted as $E_{ee} = \{(e_{k-1}, e_k)\} \cup \{(e_k, e_{k-1})\}, k > 1$. Belief edges capture the temporal changes of cognitive understanding as the conversation progresses.

The set of edges can be given by:

$$E = E_{uu} \cup E_{su} \cup E_{eu} \cup E_{ee}. \tag{5}$$

After constructing the heterogeneous dialogue graph, we obtain a feature matrix X, which represents the input features of each node and a set of adjacency matrices $\{A_k\}$ to represent the edge connection between nodes, where $A_k \in R^{N \times N}$, k denotes the type of edge and $A_k[i, j]$ is non-zero when there is a k-th type edge from i to j.

4.2.2. Graph Transformer Network

After building and initializing the emotion-belief-desire conversation graph, we learn the node representation of the graph by using the graph transformer network (Yun et al., 2019). Specifically, we set up C output channels for convolution to consider multiple metapaths simultaneously. Then, we apply the l-layer graph transformer (GT) to softly select the adjacency matrix (edge type) from the adjacency matrix A of heterogeneous graph G. After stacking l GT layers, we perform Graph Convolutional Network (GCN) (Kipf and Welling, 2016) on each metapath graph structure for graph convolution. Finally, we concatenate multiple node representations from the same GCNs on multiple metapath graphs:

$$H = \Big\Vert_{i=1}^{C} \sigma(\widetilde{D}_i^{-1} \widetilde{A}_i^{(l)} XW). \tag{6}$$

$$u_i^l = H_{u_i}. \tag{7}$$

where H denotes the new nodes set feature representation after message passing between nodes, u_i^l denotes feature of the l layer of utterance u_i, $\|$ denotes the concatenation operator, C denotes the number of channels, X is the node's feature matrix, $\widetilde{A}_i^{(l)} = A_i^{(l)} + I$ is the adjacency matrix from the ith channel of $A^{(l)}$, $\widetilde{D}_i$ is the degree matrix of $\widetilde{A}_i^{(l)}$, and W is a trainable weight matrix shared across channels.

4.3. Emotion Classification

Finally, we connect the utterance-level embedding representation of an utterance with the final node-embedding representation of an utterance node and feed it to a feedforward neural network for emotion classification:

$$z_i = h_i^u \| u_i^l. \tag{8}$$

$$p_{x,i} = Softmax(W_z z_i + b_z). \tag{9}$$

$$y_{x,i} = Argmax(p_{x,i}). \tag{10}$$

where $\|$ denotes the concatenation operator, h_i^u is an utterance-level feature representation of the utterance u_i, and u_i^l is the conversation-level feature representation of the utterance u_i. For training, we use the standard cross entropy loss as our training objective:

$$L = -\sum_{x=1}^{M} \sum_{i=1}^{N_x} y_{x,i} \log p_{x,i}. \tag{11}$$

	Dataset	IEMOCAP	DailyDialog	MELD	EmoryNLP
#Dia	Train	100	11118	1038	713
	Val	20	1000	114	99
	Test	31	1000	280	85
#Utt	Train	5236	87170	9989	9934
	Val	574	8069	1109	1344
	Test	1623	7740	2610	1328

Table 2: Statistics of four ERC datasets

where M is the total number of training conversations, N_x is the number of utterances in the x-th conversation, $p_{x,i}$ is the predicted probability distribution of emotion labels, and $y_{x,i}$ is the truth label for utterance i of the dialogue x.

5. Experimental Settings

5.1. Datasets and Evaluation

To verify the validity and reliability of our model, we conducted extensive experiments on four well-established ERC datasets. Table 2 presents the statistics of the datasets.

IEMOCAP (Busso et al., 2008) is a two-party dialogue multi-modal ERC dataset. Each conversational utterance was annotated with one of the following six emotions: neutral, happy, sad, angry, frustrated, and excited. We follow (Ghosal et al., 2020) to generate a validation set, where the training set dialogues come from the first eight speakers, and the test set dialogues come from the last two.

MELD (Poria et al., 2019) is a multi-modal dataset of multi-party conversations collected from the TV series *Friends*, and is extended from EmotionLines dataset (Chen et al., 2018). The dataset includes seven emotion labels: neutral, happy, surprised, sad, angry, disgusted, feared.

DailyDialog (Li et al., 2017) is a dataset of high-quality two-part plain text conversations. The seven categories of emotion labels are neutral, happy, surprised, sad, angry, disgusted, and feared. Since there is no fixed speaker in this dataset, we regard utterance turns as speaker turns in accordance with Shen et al.'s approach(Shen et al., 2021b).

EmoryNLP (Zahiri and Choi, 2018) is a multi-party dialogue dataset derived from *Friends*. The utterances in this dataset were annotated with seven emotional categories: neutral, joyful, peaceful, powerful, sad, mad, and scared.

Considering the textual modes, we adopted weighted average F1 for IEMOCAP, MELD and EmoryNLP. Following (Shen et al., 2021b), we chose micro average F1 for DailyDialog and exclude the utterances with majority class labels (neutral).

5.2. Implementation Details

We utilized RoBERTa-large and deployed the HuggingFace transformer toolkit (Wolf et al., 2019) for utterance-level feature extraction, which was fine-tuned during training. The final extracted feature size was 1024. By using the validation set for hold-out validation and AdamW optimizer to train the model, we selected hyperparameters such as learning rate, GT layer number, channel number, and weight decay on the graph transformer network of each dataset. Also, we set the embedded dimension of the graph transformer network of the four datasets to 64. We set the learning rate to 0.001 for IEMOCAP and MELD, 0.01 for DailyDailog, and 0.02 for EmoryNLP. In addition to setting the GT layer number of EmoryNLP to 4, the other datasets are set to 1. For EmoryNLP, the number of channels is set to 2, the number of channels for IEMOCAP is set to 6, and other datasets are set to 3. Each training and testing process is run on a single RTX 4090 Ti GPU. Each training process contains 400 epochs. We reported the results of the models based on the average score of five runs in the test set.

5.3. Baselines

We compare our proposed model with the following baselines and state-of-the-art models, CNN(Kim, 2014), KET (Zhong et al., 2019), DialogueRNN (Majumder et al., 2019), DialogueGCN (Ghosal et al., 2019), RoBERTa (Liu et al., 2019), HiTrans (Li et al., 2020), RGAT-POS (Ishiwatari et al., 2020), COSMIC (Ghosal et al., 2020), DialogXL (Shen et al., 2021a), DialogueCRN (Hu et al., 2021), SKAIG-ERC (Li et al., 2021), CauAIN (Zhao et al., 2022), CISPER (Yi et al., 2022), MPLP (Zhang et al., 2023).

6. Results and Analysis

6.1. Overall Results

As shown in Table 3, our proposed method achieves state-of-the-art results on IEMOCAP, DailyDialog and EmoryNLP.

We primarily compared our model with classical sequence-based and graph-based models in conversational emotion recognition, as well as recent models that utilize commonsense knowledge to model mental states. On the IEMOCAP dataset, we achieved a weighted F1 score of 68.22, which is 0.61 higher than CauAIN, and 1.26 and 1.57 higher than SKAIG-ERC and MPLP, respectively. For DailyDialog, we obtained a micro F1 score of 60.22, which is 0.47 and 0.3 higher than SKAIG-ERC and MPLP, respectively. On the EmoryNLP dataset, we obtained a weighted F1 score of 40.62, which is 1.74 and 0.76 higher than SKAIG-ERC and CISPER, respectively. Our performance on the IEMOCAP, DailyDialog, and EmoryNLP datasets outperformed the latest methods that utilize common-

Model	IEMOCAP	DailyDialog	MELD	EmoryNLP
Metric	weighted-F1	micro-F1	weighted-F1	weighted-F1
CNN(Kim, 2014)	52.04	50.32	55.02	32.59
KET(Zhong et al., 2019)	59.56	53.37	58.18	34.39
DialogueRNN(Majumder et al., 2019)	62.57	57.03	57.03	31.70
DialogueGCN(Ghosal et al., 2019)	64.18	-	58.10	-
RGAT-POS(Ishiwatari et al., 2020)	65.22	54.31	60.91	34.42
DialogXL(Shen et al., 2021a)	65.94	54.93	-	34.73
DialogueCRN (Hu et al., 2021)	66.20	-	58.39	-
Cosmic (Ghosal et al., 2020)	65.28	58.48	65.21	38.11
SKAIG-ERC (Li et al., 2021)	66.96	59.75	65.18	38.88
CauAIN(Zhao et al., 2022)	67.61	58.21	65.46	-
CISPER (Yi et al., 2022)	-	-	66.10	39.86
MPLP (Zhang et al., 2023)	66.65	59.92	**66.51**	-
RoBERTa (Liu et al., 2019)	63.38	58.08	62.88	37.78
Ours	**68.22**	**60.22**	64.27	**40.62**

Table 3: The performance comparison of our model and other Emotion Recognition in Conversation models.

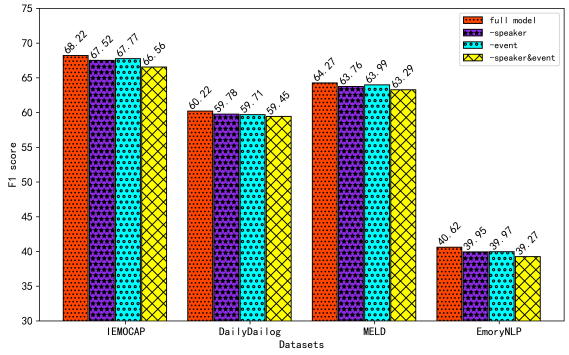

Figure 3: Effect of speakers, belief and desire.

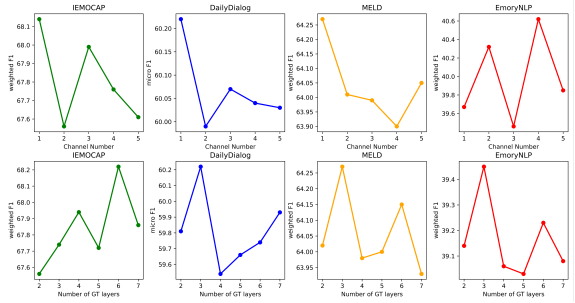

Figure 4: Effect of number of GT layers and channel number.

sense knowledge for mental state modeling. We attribute this to our effective modeling of emotion-based beliefs and desires. However, our performance on the MELD dataset did not show significant improvement. Upon analyzing the MELD dataset, we found that some utterances in conversations may have been deleted during its construction, affecting the context modeling between utterances and the modeling between emotion-eliciting events.

In addition, we found that sequence-based and graph-based methods achieve better performance through modeling context to some extent. When we adopt the better feature extractor RoBERTa, the performances of sequence-based and graph-based models improve considerably. While the feature extraction methods are the same, the graph-based method generally outperforms the sequence-based model on ERC datasets. That is because the graph-based method models the context more effectively. Moreover, combining commonsense knowledge by modeling mental states achieves better performance. Our proposed method models the context based on a heterogeneous conversation graph and introduces emotion-eliciting event nodes to model the interaction between emotional beliefs, desires, and utterances. The experimental result demonstrates the ability of our proposed method to better model the emotional and mental representation without using external commonsense knowledge.

6.2. Ablation Study

In this section, we evaluate the effects of the speaker and emotion-eliciting events on the model. We remove speakers and emotion-eliciting event nodes and edges associated with them from the conversation graph. In this process, utterance nodes and the edges between utterances are not affected, and the conversation context is always modeled. We conducted experiments on four ERC datasets, and their corresponding results are shown in Figure 3.

Removing different kinds of nodes and edges from our heterogeneous conversation graph can cause a decline in performance. When we remove the speaker nodes and the edges between the speakers and the utterances, the interaction between the speaker information and the utter-

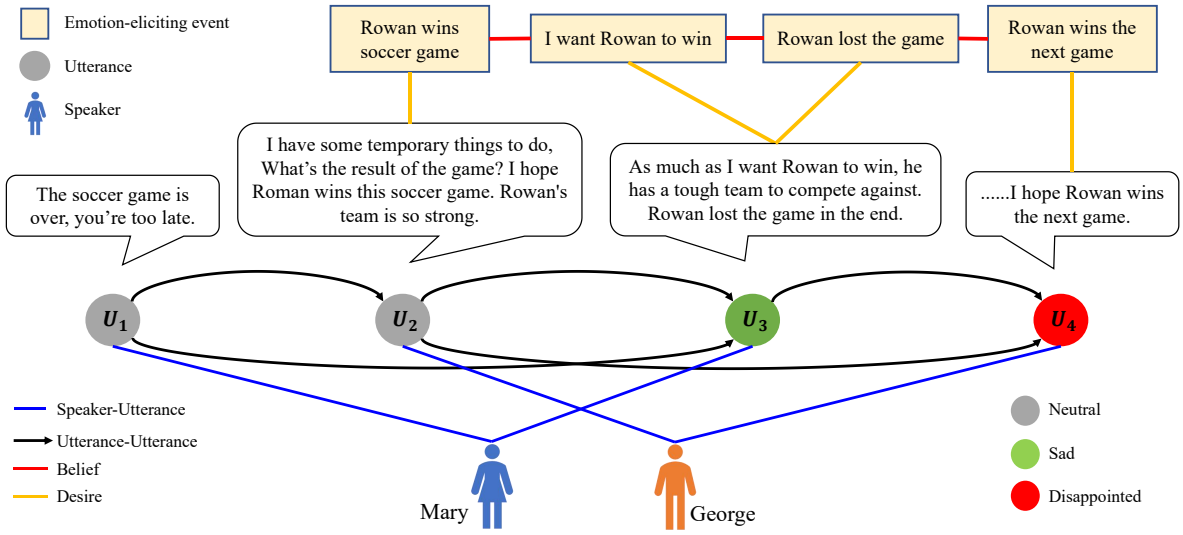

Figure 5: Case study of a conversation instance.

ance is eliminated, and performance degrades. This observation demonstrates the importance of modeling speaker information between speaker and utterance. By removing the emotion-eliciting events nodes and the edges between emotion-eliciting event nodes and utterance nodes, the information contained in emotion-eliciting events cannot be modeled, as well as the interaction between emotion-eliciting events based on beliefs and desires and utterances; thus, the performance degrades. It indicates that extracting emotion-eliciting events and modeling the interaction between events and utterances is crucial for emotion recognition in conversation. In addition, when we remove both kinds of heterogeneous nodes and heterogeneous edges, the performance degrades more severely.

6.3. Effect of GT Layer and Channel Numbers

In this section, we evaluate the influence of the GT layer number and channel number in the Graph Transformer Network on the final emotion classification results. Figure 4 presents the performance of the four datasets.

We tested the channel numbers from 2 to 8 on four datasets. From Figure 4, only IEMOCAP achieved the best performance when the channel number was 6, while other datasets achieved the best performance when the channel number was 3. We analyze that the reason may be that IEMOCAP contains more context information. In general, the influence of channel number changes on the final emotion classification performance, which first increased and then decreased. In addition, we changed the number of GT levels from 1 to 5. Except for EmoryNLP, we found that stacking too many GT layers resulted in smooth transitions and performance degradation. That is because the context information provided by the conversations in EmoryNLP is limited. Generally, with the change of GT layers, the model performance fluctuates in a range.

6.4. A Case Study

We illustrate a case study of a conversation instance in Figure 5. In this dialogue case, the previous models reasoned that Mary and George's intention was that Rowan would win the soccer game, thus, these models inferred that they had the same sad emotion. By extracting the eliciting events that involve emotion in utterances and constructing a heterogeneous graph, our model reasonably models the reasoning of desire and belief through the interaction between utterances and events, between events and events. Therefore, our model deduced that Mary and George had different cognition. George thought that Rowan's team was strong. At the previous moment, he thought that Rowan had won the game and hoped that Rowan would win the game. Mary thought that Rowan's rival was strong but hoped that Rowan would win the game. At last, George and Mary learned that Rowan had not won the game. At this moment, Mary was sad, and George was disappointed.

7. Conclusion

We proposed an approach based on the Belief-Desire Theory of Emotion for conversational emotion recognition. We construct an emotion-belief-desire heterogeneous conversation graph to model the context modeling, speakers' global state, and

belief and desire inference between emotion-eliciting events and utterances. We conducted extensive experiments, the results verify that our proposed method for ERC based on belief and desire inference achieves superior performance than multiple state-of-the-art methods. In addition, the ablation study confirms the rationality and superiority of the proposed heterogeneous conversation graph structure. Finally, how to better model the belief-desire representation of emotions is worth further research.

8. Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62072073, Grant 61906028, Grant 62076046, 62106034 and in part by the Dalian Innovation Fund 2021JJ12GX016.

9. Bibliographical References

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. Comet: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779.

Deepanway Ghosal, Navonil Majumder, Alexander 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*, pages 2470–2481.

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.

Devamanyu Hazarika, Soujanya Poria, Rada Mihalcea, Erik Cambria, and Roger Zimmermann. 2018a. 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.

Devamanyu Hazarika, Soujanya Poria, Amir Zadeh, Erik Cambria, Louis-Philippe Morency, and Roger Zimmermann. 2018b. Conversational memory network for emotion recognition in dyadic dialogue videos. In *Proceedings of the conference. Association for Computational Linguistics. North American Chapter. Meeting*, volume 2018, page 2122. NIH Public Access.

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 (Volume 1: Long Papers)*, pages 7042–7052.

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)*, pages 7360–7370.

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)*. Association for Computational Linguistics.

Thomas N Kipf and Max Welling. 2016. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*.

J. Li, D. Ji, F. Li, M. Zhang, and Y. Liu. 2020. Hitrans: A transformer-based context- and speaker-sensitive model for emotion detection in conversations. In *Proceedings of the 28th International Conference on 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*, pages 1204–1214.

Te Lin and Inwhee Joe. 2023. An adaptive masked attention mechanism to act on the local text in a global context for aspect-based sentiment analysis. *IEEE Access*, 11:43055–43066.

Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach.

Fei Long, Kai Zhou, and Weihua Ou. 2019. Sentiment analysis of text based on bidirectional LSTM with multi-head attention. *IEEE Access*, 7:141960–141969.

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.

Mehdi Naseriparsa, Md. Saiful Islam, Chengfei Liu, and Lu Chen. 2019a. Xsnippets: Exploring semi-structured data via snippets. *Data Knowl. Eng.*, 124.

Mehdi Naseriparsa, Md. Saiful Islam, Chengfei Liu, and Irene Moser. 2018. No-but-semantic-match: computing semantically matched xml keyword search results. *World Wide Web*, 21(5):1223–1257.

Mehdi Naseriparsa, Chengfei Liu, Md. Saiful Islam, and Rui Zhou. 2019b. Xplorerank: exploring XML data via you may also like queries. *World Wide Web*, 22(4):1727–1750.

Rainer Reisenzein. 2009. Emotional experience in the computational belief–desire theory of emotion. *Emotion Review*, 1(3):214–222.

Rainer Reisenzein. 2010. Broadening the scope of affect detection research. *IEEE Transactions on Affective Computing*, 1(1):42–45.

Rainer Reisenzein. 2012. What is an emotion in the belief-desire theory of emotion?

Rainer Reisenzein. 2021. Beliefs, desires, and emotions: A theory of emotions and some implications for the understanding of viewer reactions to tv serials. In *Cognition, Emotion, and Aesthetics in Contemporary Serial Television*, pages 117–138. Routledge.

Rainer Reisenzein. 2022. Emotions as affective position-takings and as nonconceptual metarepresentations: A comparison. *Emotion Review*, 14(4):273–278.

Rainer Reisenzein and Martin Junge. 2012. Language and emotion from the perspective of the computational belief-desire theory of emotion. *Dynamicity in emotion concepts*, 27:37–59.

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A Smith, and Yejin Choi. 2019. Atomic: An atlas of machine commonsense for if-then reasoning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 33, pages 3027–3035.

Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. 2021. Towards facilitating empathic conversations in online mental health support: A reinforcement learning approach. In *Proceedings of the Web Conference 2021*, pages 194–205.

Weizhou Shen, Junqing Chen, Xiaojun Quan, and Zhixian Xie. 2021a. Dialogxl: All-in-one xlnet for multi-party conversation emotion recognition. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13789–13797.

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 (Volume 1: Long Papers)*, pages 1551–1560.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31.

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*, pages 2949–2958.

Jianxiang Wang and Man Lan. 2015. A refined end-to-end discourse parser. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning-Shared Task*, pages 17–24.

Yan Wang, Jiayu Zhang, Jun Ma, Shaojun Wang, and Jing Xiao. 2020. Contextualized emotion recognition in conversation as sequence tagging. In *Proceedings of the 21th annual meeting of the special interest group on discourse and dialogue*, pages 186–195.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Bei Xu and Ziyuan Zhuang. 2022. Survey on psychotherapy chatbots. *Concurrency and Computation: Practice and Experience*, 34(7):e6170.

Jingjie Yi, Deqing Yang, Siyu Yuan, Kaiyan Cao, Zhiyao Zhang, and Yanghua Xiao. 2022. Contextual information and commonsense based prompt for emotion recognition in conversation. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 707–723. Springer.

Seongjun Yun, Minbyul Jeong, Raehyun Kim, Jaewoo Kang, and Hyunwoo J Kim. 2019. Graph transformer networks. *Advances in neural information processing systems*, 32.

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.

Duzhen Zhang, Xiuyi Chen, Shuang Xu, and Bo Xu. 2020a. Knowledge aware emotion recognition in textual conversations via multi-task incremental transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4429–4440.

Hongming Zhang, Xin Liu, Haojie Pan, Haowen Ke, Jiefu Ou, Tianqing Fang, and Yangqiu Song. 2022. Aser: Towards large-scale commonsense knowledge acquisition via higher-order selectional preference over eventualities. *Artificial Intelligence*, 309:103740.

Hongming Zhang, Xin Liu, Haojie Pan, Yangqiu Song, and Cane Wing-Ki Leung. 2020b. Aser: A large-scale eventuality knowledge graph. In *Proceedings of the web conference 2020*, pages 201–211.

Ting Zhang, Zhuang Chen, Ming Zhong, and Tieyun Qian. 2023. Mimicking the thinking process for emotion recognition in conversation with prompts and paraphrasing. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 6299–6307.

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*, pages 4524–4530.

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.

Lixing Zhu, Gabriele Pergola, Lin Gui, Deyu Zhou, and Yulan He. 2021. Topic-driven and knowledge-aware transformer for dialogue emotion detection. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1571–1582.

10. Language Resource References

Busso, Carlos and Bulut, Murtaza and Lee, Chi-Chun and Kazemzadeh, Abe and Mower, Emily and Kim, Samuel and Chang, Jeannette N and Lee, Sungbok and Narayanan, Shrikanth S. 2008. *IEMOCAP: Interactive emotional dyadic motion capture database*. Springer.

Chen, Sheng-Yeh and Hsu, Chao-Chun and Kuo, Chuan-Chun and Ku, Lun-Wei and others. 2018. *Emotionlines: An emotion corpus of multi-party conversations*.

Li, Yanran and Su, Hui and Shen, Xiaoyu and Li, Wenjie and Cao, Ziqiang and Niu, Shuzi. 2017. *Dailydialog: A manually labelled multi-turn dialogue dataset*.

Poria, Soujanya and Hazarika, Devamanyu and Majumder, Navonil and Naik, Gautam and Cambria, Erik and Mihalcea, Rada. 2019. *MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations*.

Zahiri, Sayyed M and Choi, Jinho D. 2018. *Emotion detection on tv show transcripts with sequence-based convolutional neural networks*.