Past, Present, and Future: Conversational Emotion Recognition through Structural Modeling of Psychological Knowledge

Jiangnan Li[1,2], **Zheng Lin**[1,2]*, **Peng Fu**[1], **Weiping Wang**[1]
[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2]School of Cyber Security, University of Chinese Academy of Sciences, Beijing, China
{lijiangnan,linzheng,fupeng,wangweiping}@iie.ac.cn

Abstract

Conversational Emotion Recognition (CER) is a task to predict the emotion of an utterance in the context of a conversation. Although modeling the conversational context and interactions between speakers has been studied broadly, it is important to consider the speaker's psychological state, which controls the action and intention of the speaker. The state-of-the-art method introduces CommonSense Knowledge (CSK) to model psychological states in a sequential way (forwards and backwards). However, it ignores the structural psychological interactions between utterances. In this paper, we propose a pSychological-Knowledge-Aware Interaction Graph (SKAIG). In the locally connected graph, the targeted utterance will be enhanced with the information of action inferred from the *past* context and intention implied by the *future* context. The utterance is self-connected to consider the *present* effect from itself. Furthermore, we utilize CSK to enrich edges with knowledge representations and process the SKAIG with a graph transformer. Our method achieves state-of-the-art and competitive performance on four popular CER datasets.

1 Introduction

As one of the most ubiquitous ways of communicating, conversations contain rich information and emotional expressions of the participants. With the explosive growth of conversational data on the Internet, it is of great importance to employ machines to automatically identify the emotions expressed by speakers in the conversation. Therefore, in recent years, Conversational Emotion Recognition (CER) receives increasing attention from the researchers (Poria et al., 2017; Jiao et al., 2019; Shen et al., 2021).

Unlike traditional emotion recognition, CER needs to model not only the semantic information of an utterance, but also the conversational contextual information between utterances (Jiao et al., 2019, 2020; Shen et al., 2021). Additionally, the speaker information attaching to the utterance is thought to facilitate modeling the conversational context. Different speaker modeling schemes and the corresponding solutions are proposed to enhance the interactions between utterances (Majumder et al., 2019; Li et al., 2020b; Ghosal et al., 2019; Li et al., 2020a).

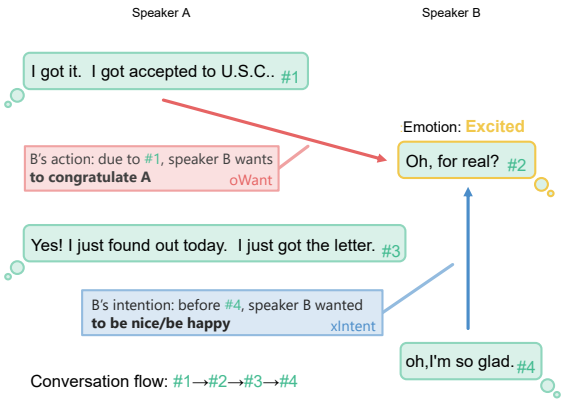


Figure 1: A conversation clip between two speakers. The utterance #1 provides the action of speaker B for #2, and #4 provides the intention. Both give positive and rational hints for #2 to predict the positive emotion `excited`. The descriptions of action, intention are generated by COMET (Bosselut et al., 2019).

Although these works yield significant performance, the modeling of conversational context and speakers does not consider psychological states of speakers. The psychological state will control the speaker's action and intention along the conversation, which can help predict the emotion more reasonably. As the original conversation provides no extra information about psychological states, to guide a model to realize psychological states, CommonSense Knowledge (CSK) can be introduced. From the perspective of CSK proposed by Sap et al.

* Zheng Lin is the corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2021, pages 1204–1214
November 7–11, 2021.

(2019), which is a kind of widely-used socialized CSK (Hwang et al., 2021), action means what the speaker **wants to do in the next step**, which can be triggered by speaker him/herself or other speakers. Intention means what the speaker **wanted to do before this step**, which can only be inferred by speaker him/herself. Therefore, for a targeted utterance, the action can be inferred from its *past* context, the intention from its *future* context. As illustrated in Fig. 1, the targeted utterance #2 can be positively enhanced by the action inferred from #1 of speaker A and the intention from #4. COSMIC (Ghosal et al., 2020) introduces this kind of CSK into CER to model the speaker's psychological state, and then utilizes bidirectional GRUs to model these states in every time step. However, COSMIC ignores the structural psychological influences from contextual utterances to the targeted utterance (i.e. an utterance can directly and explicitly pass psychological messages to other utterances over several time steps, which is more than just sequential and implicit modeling of psychological states over utterances). In addition, modeling all psychological states both forwards and backwards does not conform with the nature of the CSK (Sap et al., 2019) mentioned above (e.g. intention cannot be inferred forwards and should be only inferred backwards as illustrated in Fig. 1).

To alleviate these issues, we propose a pSychological-Knowledge-Aware Interaction Graph (SKAIG). Utterances, which are locally connected, act as the nodes in the graph. There are four relations considered in SKAIG: `xWant`, `oWant`, `xIntent`, `xEffect`. For a targeted utterance, `xWant`, `oWant` model the action indicated by utterances in the *past* context with the same speaker (`x`) and other speakers (`o`) respectively. Conversely, `xIntent` models the intention inferred by utterances in the *future* context. And `xEffect` is the self-connected relation to model the influence from the *present* utterance itself. By taking the three sources: *past*, *present*, and *future* into consideration, we believe the graph can more structurally and rationally enhance context modeling. Furthermore, these relations will be assigned to edges between utterances accordingly. Therefore, edges take the role to model the psychological interactions between utterances. To realize this, we enrich edges with their corresponding knowledge representations. These representations are produced by commonsense transformer COMET (Bosselut et al., 2019) which takes utterances and relations as inputs. As edges in SKAIG possess knowledge representations that require to be considered, we therefore utilize the graph transformer (Shi et al., 2021) for message passing. We then use the final outputs for classification.

To evaluate our method, we conduct experiments on four datasets: IEMOCAP, DailyDialog, EmoryNLP, and MELD. Our method achieves state-of-the-art performance on the first three datasets, and competitive performance on MELD. Further experiments also demonstrate the efficacy of our proposed method.

2 Methodology

In this section, we first formalize the CER task, and then elaborate on our proposed model. The framework of the model (illustrated in Fig. 2) consists of three parts: Utterance-level Encoder, Conversation-level Encoder, and Emotion Classifier.

2.1 Task Definition

For the subsequent context, a conversation containing N textual utterances is denoted as $C = [u_1, u_2, ..., u_N]$. In an utterance $u_n = [w_1, w_2, ..., w_{L_n}]$, L_n words are contained. In addition, a conversation involves at least two speakers, and each utterance within is expressed by its corresponding speaker $s \in (S_1, S_2, ..., S_P)$. Therefore, CER task aims to classify all utterances in one conversation to their correct emotion labels which belong to the set (E_0, E_1, ..., E_M).

2.2 Utterance-level Encoder

For each utterance out of the conversational context, it is important to extract the contextual information among its words. We employ the widely-used pretrained model RoBERTa (Liu et al., 2019) to encode the utterance. An utterance $u_n = [w_1, w_2, ..., w_{L_n}]$ is fed into RoBERTa, we obtain the hidden states of the last layer:

$$W = \text{RoBERTa}(w_1, w_2, ..., w_{L_n}) \quad (1)$$

where $W \in \mathbb{R}^{L_n \times d_w}$ and d_w is the dimension of hidden states of words. The goal of the utterance-level encoder is to encode the representation for each utterance. Therefore, we deploy a max-pooling operation and a linear projection following Ishiwatari et al. (2020), Li et al. (2020b):

$$c_n = \text{Linear}(\text{Maxpooling}(W)) \quad (2)$$

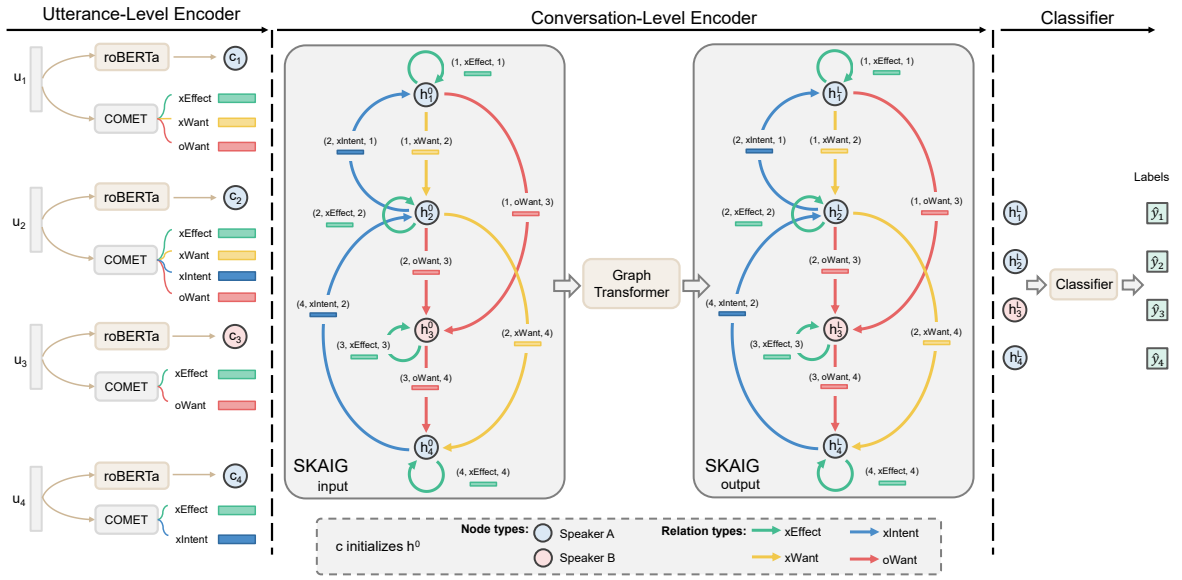


Figure 2: The framework of our model. The utterances are encoded by the utterance-level encoder to produce the utterance representations and the edge representations. The conversation-level encoder processes the SKAIG whose window size is 1. Finally, the classifier predicts the emotion for every utterance. Especially, edges and their representations with different relations are in different colors.

where $c_n \in \mathbb{R}^{d_u}$ is the representation of the utterance u_n and d_u is the dimension of the representation. After all utterances encoded, we obtain the representation of the conversation $C \in \mathbb{R}^{N \times d_u}$.

2.3 Conversation-Level Encoder

Considering each utterance in its conversational context, there is rich contextual information. For an utterance, the action and intention of the speaker and interactions with other utterances in *past*, *present*, and *future* are crucial to model the context more precisely. Therefore, we construct a pSychological-Knowledge-Aware Interaction Graph (SKAIG) of utterances in a conversation, and then utilize the Graph Transformer (Shi et al., 2021) network to process SKAIG.

2.3.1 SKAIG Construction

We construct a directed graph modeling interactions between utterances. We denote the graph as $\mathcal{G} = (\mathcal{V}, \mathcal{E}, \mathcal{R}, \mathcal{A})$. Specifically, $u_n \in \mathcal{V}$ is an utterance node, $r \in \mathcal{R}$ is an edge type, $e_{i,j} = (u_i, r, u_j) \in \mathcal{E}$ is the edge between utterance i and j, and $a_{i,j} \in \mathcal{A}$ is the edge attribute (representation) of $e_{i,j}$.

Vertices: For an utterance u_n acting as a node in the graph, we use the representation $c_n \in \mathbb{R}^{d_u}$ encoded by the utterance-level encoder to initiate the node feature h_n^0. The initial node feature contains no conversational contextual information.

Relations: The interaction between utterances is often indicated by the relations between the speakers. In previous works(Ghosal et al., 2019; Ishiwatari et al., 2020), there are two important speaker relations r considered: self-dependency and inter-speaker dependency. Based on this scheme, we propose more refined types of relations so that the speaker's action and intention in the conversation can be modeled. In our setting, the utterances in the post context can guide the action of the current utterance and those in the future context can predict the intention. Therefore, for two utterances u_i, u_j where u_i appears before u_j, if they share the same speaker, the relation $u_i \rightarrow u_j$ means that u_i passes the guidance of the speaker's action to u_j, and we denote this relation as `xWant`. The relation $u_i \leftarrow u_j$ represents that u_j can predict the intention of the speaker as u_j is in the future context for u_i, and we denote it as `xIntent`. Conversely, if u_i and u_j do not share the speaker, $u_i \rightarrow u_j$ will provide the influence of u_i's speaker on the action of u_j's speaker, and we denote it as `oWant`. As the intention only can be inferred by the speaker him/herself (Sap et al., 2019), no "intent" relation connects two utterances with different speakers. Furthermore, an utterance can be self-connected ($u_i \rightarrow u_i$) and the self-effected relation is denoted as `xEffect`. Therefore, we get four types of edge relations $\mathcal{R}$ = (`xEffect`, `xWant`, `oWant`,

xIntent).

Edges: An edge $e_{i,j} = (u_i, r, u_j)$ between two utterances u_i and u_j models the interactions between these utterances. We think that the influence of an utterance on contextual utterances can be locally effective, so we connect the targeted node with the contextual nodes of every speaker in a window whose size is k. When $k = 1$, the targeted utterance considers one utterance of every speaker in the past and future context respectively, which is exemplified in Fig. 2.

Edge Representations: Different from the previous works (Ghosal et al., 2019; Ishiwatari et al., 2020) that only assign a weight to the edge, we introduce the commonsense knowledge to enrich the edges with different relations. Fortunately, commonsense transformer COMET (Bosselut et al., 2019), which is a GPT (Radford et al., 2018) model, can provide such features for all of our relations. We utilize a COMET model trained on ATOMIC (Sap et al., 2019) which is a knowledge base of *If-Then* reasoning. There are nine relations in ATOMIC, which cover all of the relations we require. Under such circumstances, COMET can generate descriptions of "then" based on the input and the selected relation. For example, if taking u_n and the relation xWant as inputs, COMET will generate a reasoning sequence following "If u_n, then X wants to".

We concatenate u_n and a relation with mask tokens (e.g. u_n [MASK] <xWant>) in the inputting format of COMET, and then COMET processes the input. Following Ghosal et al. (2020), we take the hidden state of the relation token from the last layer of COMET transformer encoder as the relation's representation. For an edge $e_{i,j} = (u_i, \texttt{xWant}, u_j)$, the corresponding representation is $a_{i,j}$, whose dimension is mapped from 768 to d_u with a following linear unit.

2.3.2 Graph Transformer

We utilize an L-layer graph transformer to propagate the interactive information through the SKAIG. We update the node representation $h_i^{(l)} \in \mathbb{R}^{d_u}$ of each node $u_i \in \mathcal{V}$ by:

$$h_i^{(l+1)} = (1 - \beta_i) \left(\sum_{j \in \mathcal{N}(i)} \alpha_{i,j} m_j \right) + \beta_i \mathrm{W_s} h_i^{(l)} \tag{3}$$

where $\mathcal{N}(i)$ is the set of source nodes connected with the targeted node i, m_j is the message passed by these nodes, $\alpha_{i,j}$ is the attention score, $\beta_i \in \mathbb{R}^1$ is the gate for the residual connections, and $\mathrm{W_s} \in \mathbb{R}^{d_u \times d_u}$ is a mapping weight.

The message passed by neighboring nodes contains two parts of information: the contextual relevance and the psychological information, so the message is computed by:

$$m_j = f_v(h_j^{(l)}) + \mathrm{W_e} a_{j,i} \tag{4}$$

where $\mathrm{W_e} \in \mathbb{R}^{d_{head} \times d_u}$ is a trainable weight and $f_v(x) = \mathrm{W_v} x + \mathrm{b_v}$is a projection, both mapping dimension from $\mathbb{R}^{d_u}$ to the head dimension $\mathbb{R}^{d_{head}}$. Furthermore, the attention score that controls how much information should be gathered from neighbors can be computed by:

$$\alpha_{i,j} = \mathrm{softmax} \left(\frac{f_q(h_i^{(l)})(f_k(h_j^{(l)}) + \mathrm{W_e} a_{j,i})}{\sqrt{d_{head}}} \right) \tag{5}$$

where $f_q(x) = \mathrm{W_q} x + \mathrm{b_q}$; $f_k(x) = \mathrm{W_k} x + \mathrm{b_k}$ are projections. Eq. (3) only considers one attention head, while multiple heads are involved in practice. We concatenate outputs from all heads after message aggregation and denote it as o_i. As for the gate, $\beta_i = \mathrm{sigmoid}(\mathrm{w_g}^T [h_i^{(l)}; o_i; h_i^{(l)} - o_i])$, where [] is the concatenating operation.

In addition, we replace the original operation after the attention in Shi et al. (2021) to a pointwise feed forward network proposed by Vaswani et al. (2017). We denote the final output of the conversation as $H^L \in \mathbb{R}^{N \times d_u}$.

2.4 Emotion Classifier

We utilize a linear unit as the classifier to predict the emotion distributions:

$$\hat{Y} = \mathrm{softmax}(H^L \mathrm{W_c} + \mathrm{b_c}) \tag{6}$$

where $\mathrm{W_c} \in \mathbb{R}^{d_u \times M}$, $\mathrm{b_c} \in \mathbb{R}^M$. The cross-entropy loss utilized to train the model is calculated on a conversation by:

$$L = -\frac{1}{N} \sum_{i=1}^{N} \sum_{e=1}^{M} y_i^e \log(\hat{Y}_i^e) \tag{7}$$

where y_i is the one-hot vector denoting the emotion of utterance i in the conversation, and e is the dimension of each emotion.

Dataset	Num. of dialogue			Num. of utterance		
	train	dev	test	train	dev	test
IEMOCAP	120		31	5810		1623
DailyDialog	11118	1000	1000	87170	8069	7740
EmoryNLP	659	89	79	7551	954	984
MELD	1039	114	280	9989	1109	2610
	Avg. dialogue len.			Avg. utterance len.		
IEMOCAP	48		52	12		13
DialyDialog	8	8	8	12	11	12
EmoryNLP	12	11	13	8	7	8
MELD	10	10	9	8	8	8

Table 1: Statistics of IEMOCAP, DialyDialog, MELD, EmoryNLP.

3 Experimental Setup

3.1 Dataset

We conduct experiments with our model on four datasets: IEMOCAP (Busso et al., 2008), DailyDialog (Li et al., 2017), EmoryNLP (Zahiri and Choi, 2018), and MELD (Poria et al., 2019). Statistics about the datasets are showen in Tab. 1.

IEMOCAP IEMOCAP consists of dyadic covesations between ten speakers. Six emotions are considered in previous works: `neutral`, `happy`, `sad`, `angry`, `excited`, `frustrated`. We split training and validation set following Ghosal et al. (2019).

DailyDialog DailyDialog is a dataset containing two-way dialogues about the daily life. Seven emotions are included: `neutral`, `happiness`, `sadness`, `anger`, `surprise`, `disgust`, `fear`. In DailyDilaog, over 83% of the utterances are labeled with `neutral`.

EmoryNLP EmoryNLP is collected from the TV series *Friends*, which contains multi-speaker conversations. Seven emotions are annotated: `neutral`, `mad`, `sad`, `scared`, `powerful`, `peaceful`, `joyful`.

MELD MELD is also collected from *Friends*. Therefore, it is a dataset with multi-speaker conversations. The emotions are the same as those in DailyDialog.

3.2 Baselines and Compared Methods

We compare our model with the following baselines and state-of-the-art models:

CNN (Kim, 2014) is the widely-used text convolution network. **DialogueRNN** (Majumder et al., 2019) employs GRUs to track speakers' global and emotional states. Ghosal et al. (2020) implement both CNN and RoBERTa based DialogueRNN. **DialogueGCN** (Ghosal et al., 2019) uses graph convolutional networks to process the graph constructed from self-dependency and inter-speaker dependency. **KET** (Zhong et al., 2019) is a hierarchical transformer using their proposed graph attention to extract information from knowledge base. **HiTrans** (Li et al., 2020a) is a hierarchical transformer based on BERT which is augmented with a speaker relation prediction task. **RGAT-POS** (Ishiwatari et al., 2020) is a relation-aware graph attention network utilizing the proposed relational position encoding. The speaker modeling of this model is based on DialogueGCN. **DialogXL** (Shen et al., 2021) is an all-in-one XLNet that processes the conversation in one step. DialogXL also utilizes the speaker modeling of DialogueGCN. **COSMIC** (Ghosal et al., 2020) is a modified DialogueRNN based on RoBERTa-large. COSMIC models more refined states of speakers by utilizing bidirectional GRUs. COSMIC utilizes commonsense knowledge COMET to initialize a part of inputs of the speaker's internal, external, and intent GRUs. **RoBERTa** (Liu et al., 2019) is the utterance-level encoder directly followed by a classifier. **RoBERTa-Transformer** replaces the graph transformer with a transformer, which can be regarded as a locally and fully connected graph without mental relation modeling. We implement RoBERTa and RoBERTa-transformer in the setting of our method. For other models, we refer the performance from the corresponding papers.

3.3 Implementation

For IEMOCAP, we use RoBERTa-base [1] to initialize the utterance-level encoder. For other datasets, RoBERTa-large is selected, which is deployed by *HuggingFace* transformers toolkit (Wolf et al., 2019). RoBERTa is fine-tuned when training. The batch size is set to 1 for IEMOCAP and 8 for other datasets. For graph transformer, the dimension of the utterance is set 200 for MELD and 300 for other datasets; the dimension of the feed forward network is set to 200 for MELD and 600 for other datasets; the head dimension is set to 50 for all datasets; the number of layers is searched from 1 to 6. We train the model with the AdamW optimizer (Loshchilov and Hutter, 2019) whose learning rate is set to 8e-6 for MELD and 1e-5 for other datasets. The training step is set to 10000 with the first 1000 steps for warming up and other steps decaying the

[1] We find that the performance on IEMOCAP of RoBERTa-base and RoBERTa-large is similar. To reduce the computation, we use RoBERTa-base.

Methods	IEMOCAP	DailyDialog		EmoryNLP	MELD
	weighted-F1	micro-F1	macro-F1	weighted-F1	weighted-F1
CNN	52.04	50.32	36.87	32.59	55.02
DialogueRNN	62.57	55.95	41.8	31.7	57.03
DialogueGCN	64.18	-	-	-	58.1
KET	59.56	53.37	-	34.39	58.18
HiTrans	64.5	-	-	36.75	61.94
RGAT-POS	65.22	54.31	-	34.42	60.91
DialogXL	65.94	54.93	-	34.73	62.41
RoBERTa DialogueRNN	64.76	57.32	49.65	37.44	63.61
COSMIC	65.28	58.48	51.05	38.11	**65.21**
RoBERTa	55.67	55.16	48.2	37.0	62.75
RoBERTa Transformer	63.78	58.28	47.0	37.5	64.59
Ours	**66.96**	**59.75**	**51.95**	**38.88**	**65.18**

Table 2: Results of our method and state-of-the-art baselines. The results of RoBERTa on DailyDialog are referred from Ghosal et al. (2020).

learning rate. Early stopping is activated with 10 epochs. [2]

For IEMOCAP, EmoryNLP, and MELD, the weighted F1 score is selected as the evaluating metric. For DailyDialog, following previous works, we report the micro F1 score excluding those utterances labeled with `neutral` and the macro F1 score. All of our results are averaged on 5 runs.

4 Results and Discussions

4.1 Overall Results

Illustrated in Tab. 2, our method achieves state-of-the-art results on IEMOCAP, DailyDialog, EmoryNLP, and competitive results on MELD.

For IEMOCAP, RoBERTa performs poorly comparing to models with conversational context modeling, which indicates IEMOCAP contains rich information of conversational context. COSMIC achieves limited improvement against RoBERTa-DialogueRNN, while our method outperforms RoBERTa-Transformer. We think the reason is that our method can benefit from the structural modeling of psychological knowledge in IEMOCAP. Conversely, COSMIC only models psychological states by updating step by step. However, the interactions between utterances in several steps play an important role in IEMOCAP, which can be elucidated in Fig. 3. To this end, our method models better conversational context and outperforms COSMIC by 1.68 weighted-F1. For models based on pretrained models, the performance is similar. Our method performing better indicates the importance of CSK to enhance psychological states.

For DailyDialog, our method exceeds COSMIC by 1.27 micro-F1 and RoBERTa-Transformer by 1.47 mirco-F1. In this case, RoBERTa-Transformer is competitive in micro-F1 but the performance on macro-F1 is poor. Conversely, our method achieves the best macro-F1, which demonstrates the introduction of SKAIG can partly defend the influence of data imbalance on transformer.

For EmoryNLP, the contextual information provided by conversations is limited as RoBERTa achieves similar performance as Transformer and DialogueRNN. In such case, our method still exceed COSMIC by 0.77 weighted-F1. For MELD, our method achieves competitive performance against COSMIC . We think the reason maybe that MELD contains short conversations but involves multiple speakers, which leads to limited interactive influence from psychological state. Therefore, our method does not show advantages on MELD. The error analysis on MELD is present in section 4.5.

4.2 Effect of Relations

We evaluate the effect of different relations to our model. We take one relation off our proposed SKAIG, where the edge will not be eliminated to keep the modeling of conversational context. To achieve this, we only remove the edge representations of the selected relation. In addition, we train

[2]The code is available at https://github.com/LeqsNaN/SKAIG-ERC.

	IEMOCAP	DailyDialog
Method Type	weighted-F1	micro-F1
full model	66.96	59.75
-xWant	64.33	59.42
-oWant	65.03	59.09
-xIntent	64.7	59.46
-xEffect	65.29	58.95
trainable	64.28	58.86

Table 3: The weighted-F1 scores on IEMOCAP and the micro-F1 scores on DailyDialog of our full model taking off different relations and model variants. "-relation" denotes taking off the edge representation of the "relation". "trainable" denotes replacing edge representations with trainable relation embeddings.

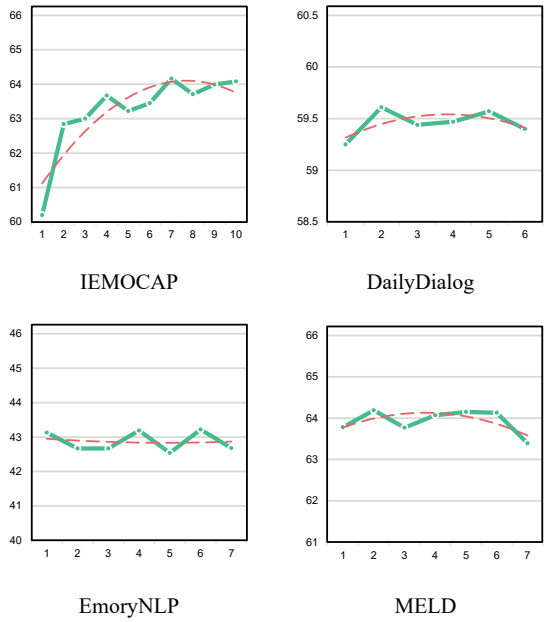


Figure 3: The effect of the window size to our model on different datasets. X-axis denotes the window size. Y-axis denotes the micro-F1 for DailyDialog and weighted-F1 for other datasets on the validation set. The dotted line denotes the trend line of second-order polynomial.

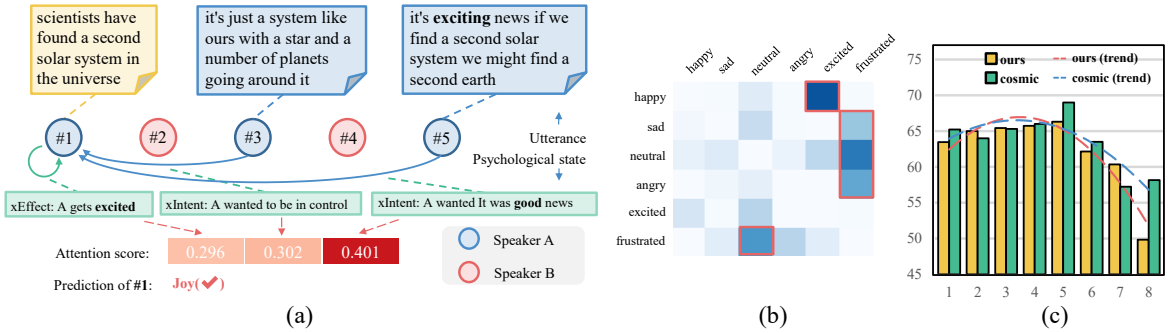
a model with four trainable relation embeddings, where the embeddings model the four relations in SKAIG. This model variant does not introduce any CSK. We conduct the experiments on IEMOCAP and DailyDialog, and the results are illustrated in Tab. 3.

Taking off different relations in SKAIG leads to different degree of performance drop. By taking off the self-connected relation xEffect, the performance drops. This observation indicates the importance of modeling self-effect in the *current* state. Furthermore, by taking off xWant or oWant, where the two relations model the action information provided by the *past* context from different speakers, the performance drops. This demonstrates that the information about action can enhance interactions between utterances. On the other hand, xIntent also affects the performance of our model, which indicates the necessity of considering the intent information from the *future* context. The trainable model variant performs poorly as it achieves the lowest F1 scores. We deduce the reason may be that no CSK is provided to inform what kind of the relation is modeled between two utterances. This emphasizes the importance of CSK to guide the model to learn more rational information about the speaker's psychological states.

4.3 Effect of Window Size

In this section, we evaluate the effect of the window size to our method. The performance on the validation set is illustrated in Fig. 3. Except IEMOCAP, the upper window size for others is 6 or 7, because these datasets contain relatively short conversations as shown in Tab. 1. The increasing rate of the number of edges in the graph becomes slow when the window size exceeds 6 or 7.

From the illustration, only IEMOCAP shows significant improvement with the window widening, while other datasets show flat trends. The reason maybe that IEMOCAP contains more contextual information and obvious interactions of utterances in the conversation (as elucidated in section 4.1). Inferred from trend lines, whose changing ranges are different in different datasets though, the performance basically increases first and then drops as the window becomes large except EmoryNLP. This observation accords with our claim that the psychological interactions between utterances are locally effective. On the other hand, the reason for our method not sensitive to the window size on EmoryNLP may be that the contextual information provided by conversations in this dataset is limited. This can be inferred from the similar performance of RoBERTa and RoBERTa-Transformer (-DialogueRNN) in Tab. 2.



(a) (b) (c)

Figure 4: (a) A case of the targeted utterance #1 getting clues from #3 and #5 after 2 steps. (b) The confusion matrix excluding the diagonal on IEMOCAP. (c) F1 scores of conversations with different number of speakers on MELD achieved by our method and COSMIC. X-axis denotes F1 score; Y-axis denotes the number of speakers in a conversation. The dotted line denotes the trend line of second-order polynomial.

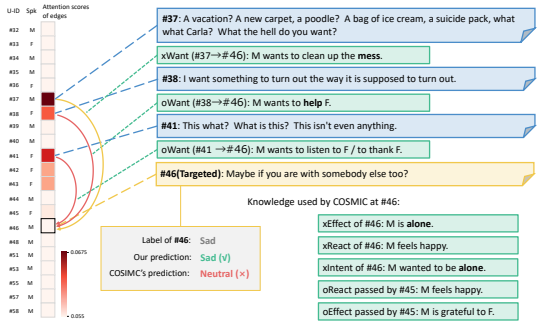


Figure 5: A case that our method gives the correct prediction while COSMIC fails. The attention map depicts the importance of all edges towards #46. For our method, three edges with the highest attention scores are illustrated. For COSMIC, the used knowledge is illustrated.

4.4 Case Study

In Fig. 4 (a), we exemplify a simple case that a targeted utterance gets messages of intent from future utterances. Specifically, the attention scores are averaged from the attention heads in the top layer of graph transformer. For `xEffect` of #1, CSK can provide a positive indication of the speaker's self-effect state, where #1 is likely to be predicted as `neutral` by models without CSK. As for #5 that is two steps from #1, the `xIntent` provided by it can positively enhance #1. In addition, the attention score of the edge (5, `xIntent`, 1) is the highest among all the in-degree edges of #1. This coincides our claim that an utterance can directly and explicitly pass psychological messages to other utterances over several steps, and indicates the necessity of modeling structural interactions.

In Fig. 5, we illustrate a case that our method gives the correct prediction while COSMIC fails. In this case, messages of action from history utterances contribute the most while the self-effect (#46 → #46) and intent (#(> 46) → #46) have lower importance. `xWant` directly passed by #37 and `oWant` from #38 can provide positive guidance to #46 and they are both several steps away, which further demonstrates the importance of direct structural psychological interactions. Conversely, COSMIC considers intent, effect, reaction from #46 itself and effect, reaction from neighboring #45 due to the sequential modeling. Although the knowledge can provide useful clues like "alone", COSMIC fails to make the correct prediction. This indicates that COSMIC is hindered by the implicit and limited psychological interactions with contextual utterances, even though contextual utterances can provide more effective psychological information.

4.5 Error Analysis

In Fig. 4 (b), we illustrate the confusion matrix of predictions on IEMOCAP. To study the condition that our model fails in, the diagonal in the confusion matrix is eliminated to zero. The deeper color denotes that more samples are misclassified. From the heatmap, `happy` samples are likely predicted as `excited`, and other negative emotions like `sad` are more confused with `frustrated`. These observations indicate that the difficulty of discriminating similar emotions in emotion recognition still disturbs our method.

In Fig. 4 (c), we illustrate the effect of increasing speakers in a conversation to our method and COSMIC on MELD. At first, the performance of ours and COSMIC increases, which is different from that of HiTrans (Li et al., 2020a) that constantly decays. Compared with COSMIC, our method

can achieve competitive performance. This appearance demonstrates that our method can handle the condition involving a small amount of speakers. However, when the number keeps increasing, our method show the same dropping trend of performance as HiTrans and COSMIC do, but the trend is sharper than that of COSMIC. This indicates that it becomes hard for our method when a conversation involves a large scale of speakers. In the future work, we will endeavor to explore more effective schemes of speaker modeling to deal with the condition that involves multiple speakers.

5 Related Work

Conversational Emotion Recognition is a hot-spot task in recent years. Unlike traditional Emotion Recognition, CER involves conversational context. Hazarika et al. (2018b,a); Jiao et al. (2020) utilize memory network to model such context. To consider the speaker and listener in the conversation, Majumder et al. (2019) propose DialogueRNN, which utilizes GRUs to update speakers' states and the global line of the conversation. DialogueGCN (Ghosal et al., 2019) models two relations between speakers: self and inter-speaker dependencies, and utilizes graph networks(Schlichtkrull et al., 2018; Kipf and Welling, 2017) to model the graph constructed by these relations. Zhong et al. (2019) propose a graph attention to extract information from external knowledge base and utilize Transformer (Vaswani et al., 2017) to model conversations.

Furthermore, with the spreading of pretrained models, new works are based on these high-performance and large-scale models. Ishiwatari et al. (2020) propose a relation-aware position encoding based on DialogueGCN and utilize BERT (Devlin et al., 2019) to encode utterances. Li et al. (2020a) utilize BERT and propose a speaker relation prediction task to augment CER. Shen et al. (2021) utilize XLNet (Yang et al., 2019) and model the whole conversation in one step. By introducing commonsense knowledge to CER, Ghosal et al. (2020) propose COMSIC which is based on DialogueRNN equipped with RoBERTa (Liu et al., 2019) to model the speakers' internal, external, intent states.

6 Conclusion

In this paper, we study conversational emotion recognition. The SOTA method ignores the psychological interactions between utterances over several time steps and does not conform with the nature of psychological states. We therefore propose a pSychological-Knowledge-Aware Interaction Graph (SKAIG). The graph contains four relations to model psychological states of speakers. Enhanced by commonsense knowledge and the deployment of the graph transformer, our method yields SOTA or competitive performance on benchmark datasets.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No. 61976207, No. 61906187).

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 Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 4762–4779.

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. *Lang. Resour. Evaluation*.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Deepanway Ghosal, Navonil Majumder, Alexander F. Gelbukh, Rada Mihalcea, and Soujanya Poria. 2020. COSMIC: commonsense knowledge for emotion identification in conversations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, EMNLP 2020, Online Event, 16-20 November 2020*, pages 2470–2481.

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.

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, Brussels, Belgium, October 31 - November 4, 2018*, 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 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2018, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 1 (Long Papers)*, pages 2122–2132.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (comet-) atomic 2020: On symbolic and neural commonsense knowledge graphs. 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 6384–6392.

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.

Wenxiang Jiao, Michael R. Lyu, and Irwin King. 2020. Real-time emotion recognition via attention gated hierarchical memory network. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 8002–8009.

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, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, 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 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1746–1751.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*.

Jingye Li, Donghong Ji, Fei Li, Meishan Zhang, and Yijiang Liu. 2020a. Hitrans: A transformer-based context- and speaker-sensitive model for emotion detection in conversations. In *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4190–4200.

Qingbiao Li, Chunhua Wu, Kangfeng Zheng, and Zhe Wang. 2020b. Hierarchical transformer network for utterance-level emotion recognition. *CoRR*, abs/2002.07551.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing, IJCNLP 2017, Taipei, Taiwan, November 27 - December 1, 2017 - Volume 1: Long Papers*, pages 986–995.

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.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*.

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.

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, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 873–883.

Soujanya Poria, Devamanyu Hazarika, Navonil Majumder, Gautam Naik, Erik Cambria, and Rada Mihalcea. 2019. 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.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

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 *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 3027–3035.

Michael Sejr Schlichtkrull, Thomas N. Kipf, Peter Bloem, Rianne van den Berg, Ivan Titov, and Max Welling. 2018. Modeling relational data with graph convolutional networks. In *The Semantic Web - 15th International Conference, ESWC 2018, Heraklion, Crete, Greece, June 3-7, 2018, Proceedings*, pages 593–607.

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

Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjing Wang, and Yu Sun. 2021. Masked label prediction: Unified message passing model for semi-supervised classification. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 1548–1554.

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.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, and Jamie Brew. 2019. Huggingface's transformers: State-of-the-art natural language processing. *CoRR*, abs/1910.03771.

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*, pages 44–52.

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.