PRELIMINARY PREPRINT VERSION: DO NOT CITE
The AAAI Digital Library will contain the published
version some time after the conference.

Is Discourse Role Important for Emotion Recognition in Conversation?

Donovan Ong[1,2,3], Jian Su[1], Bin Chen[1], Anh Tuan Luu[3*], Ashok Narendranath[1], Yue Li[1], Shuqi Sun[4], Yingzhan Lin[4], Haifeng Wang[4]

[1]Institute for Infocomm Research, A*STAR, Singapore
[2]CNRS@CREATE LTD, Singapore
[3]School of Computer Science and Engineering, Nanyang Technological University, Singapore
[4]Baidu Inc., China
{donovan_ong,sujian,bchen,machingaln,li_yue}@i2r.a-star.edu.sg, anhtuan.luu@ntu.edu.sg,
{sunshuqi01,linyingzhan01,wanghaifeng}@baidu.com

Abstract

A conversation is a sequence of utterances, where each utterance plays a specific discourse role while expressing a particular emotion. This paper proposes a novel method to exploit latent discourse role information of an utterance to determine the emotion it conveys in a conversation. Specifically, we use a variant of the Variational-Autoencoder (VAE) to model the context-aware latent discourse roles of each utterance in an unsupervised way. The latent discourse role representation further equips the utterance representation with a salient clue for more accurate emotion recognition. Our experiments show that our proposed method beats the best-reported performances on three public Emotion Recognition in Conversation datasets. This proves that the discourse role information of an utterance plays an important role in the emotion recognition task, which no previous work has studied.

Introduction

With the growing usage of chatbots on e-commerce platforms and increasing online messaging at work due to the ongoing pandemic, a number of potential applications for conversational artificial intelligence (AI) have arisen. One key topic - emotion recognition in conversations (ERC) has also started to gain attention from both the research (Hazarika et al. 2018a,b; Majumder et al. 2019; Zhong, Wang, and Miao 2019; Ghosal et al. 2020; Li et al. 2020) and industrial community. Besides being used to analyse the quality of customer service conversations, automated chatbots use it to detect the emotion of users during ongoing dialogue and engage users with real-time emotion-aware responses.

Previous works on recognising the emotion of an utterance in a conversation mainly consider two factors - (i) context of the conversation (Poria et al. 2017), i.e. what was said previously, and (ii) identity of the speaker (Ghosal et al. 2019; Li et al. 2020). Quite some efforts (Poria et al. 2017; Ghosal et al. 2019; Ishiwatari et al. 2020) focus on how to incorporate information about the conversation context from the surrounding utterances to identify the emotion expressed by the query utterance. Hazarika et al. (2018a) and Majumder et al. (2019) explicitly model the effect of the speaker identity to incorporate the speaker's emotional state. While these factors have proven to be effective in identifying the emotion expressed by an utterance, there are other factors that also influence the emotion. A conversation is a sequence of utterances, where each utterance plays a specific discourse role while expressing a particular emotion. The discourse role played by each utterance, such as asking a question, disagreeing to a previous prompt, and other dialogue acts (Zeng et al. 2019) captures the underlying user intent and information flow of a conversation. Thus, earlier works that ignore the discourse role suffer from the intent-misleading problem, i.e., they cannot capture the utterance's intention correctly, especially when the intent is not explicit, and the utterance has lexical ambiguity.

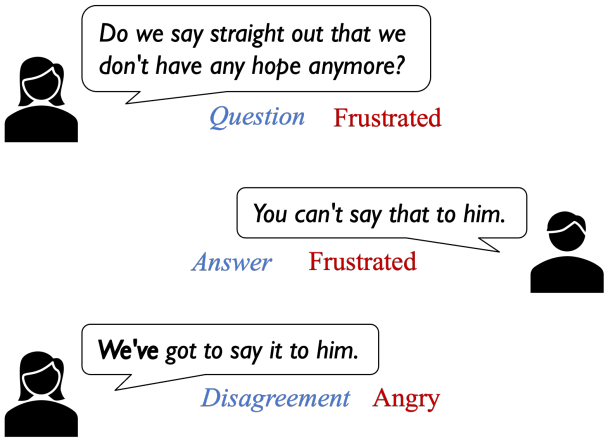

Figure 1: Utterances with their discourse roles and emotions.

For example, in the conversation shown in Figure 1, by recognising that the last utterance *"We've got to say it to him"* plays a *disagreement* role in the dialogue flow, we can readily infer that the utterance conveys an angry emotion. On the other hand, if the utterance is intended to be *sympathy*, then the speaker will be expressing a sad emotion. Consequently, we observe that the discourse role could provide a signal to identify the emotion expressed.

*The author's contribution to this paper was made when he was at Institute for Infocomm Research, A*STAR.

Copyright © 2022, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

In this paper, we aim to enhance utterance representations with their discourse role information to determine the emotion being conveyed. Our work focuses only on the textual information in a conversation to detect the underlying emotion. However, it is noticeable that the discourse role annotation based on the conversation text is not available in most public ERC datasets. Only one ERC dataset, DailyDialog (Li et al. 2017) has such annotations. It is tedious and time-consuming to annotate discourse role information for large-scale datasets. To overcome this limitation, and more importantly to facilitate the real-world application in different domains / datasets when discourse roles are not available, we propose a variant of the Variational-Autoencoder (VAE) (Kingma and Welling 2014) to obtain a latent variable for each utterance in an unsupervised manner. Specifically, the VAE learns the latent variable by reconstructing the utterance. The learned latent variable is used to represent the latent discourse role. Additionally, we observe the existence of i) dependency between adjacent discourse roles and ii) dependency between the conversation context and discourse role, e.g. a question followed by an answer in Figure 1. Therefore, we extend the VAE to capture these two dependencies. In particular, we employ a recurrent neural network to model the sequential nature of discourse roles and use a pair of adjacent utterances to construct a context-aware input for the VAE.

The contributions of our paper are three-fold:

- We are the first to study the importance of the discourse role information of utterances in the emotion recognition in conversation task.
- To this end, we introduce a variant of VAE to model the context-aware latent discourse role with a latent variable to overcome the lack of discourse role annotation in ERC datasets.
- We validate the efficacy of our proposed model and achieve state-of-art performance on three publicly available datasets of differing sizes - DailyDialog (Li et al. 2017), MELD (Poria et al. 2019) and IEMOCAP (Busso et al. 2008).

Related Work

Emotion Recognition in Conversations

Earlier studies on detecting emotion of utterances in conversations focus on tracking the speaker's state using recurrent neural networks (Hazarika et al. 2018a,b). Majumder et al. (2019) added two recurrent neural networks to track the conversation's global context and emotional state. In another line of work, Ghosal et al. (2019), Zhang et al. (2019) and Ishiwatari et al. (2020) used graph neural networks to model the self and inter-speaker dependencies in a conversation. The nodes in the graph represent the utterances and edges represent the self and inter-speaker dependencies between utterances. Meanwhile, Zhong, Wang, and Miao (2019), Li et al. (2020) and Zhang et al. (2020) modelled the correlation between intra and inter utterance using self-attention and cross-attention in the transformer (Vaswani et al. 2017) to generate context-aware utterance representations.

In a conversation, humans do not always explicitly express their emotions in the words they say and often rely on common sense knowledge to understand one another. Hence, Zhong, Wang, and Miao (2019), Ghosal et al. (2020) and Zhang et al. (2020) incorporated common sense knowledge to improve emotion detection. Notably, transfer learning has been widely adopted in NLP as it has shown significant improvements on multiple tasks. Jiao, Lyu, and King (2020) explored pretraining an encoder on utterance completion task before fine-tuning on ERC datasets. Naturally, we also observe that earlier emotions influence the emotion in conversations. Lu et al. (2020) model the interaction between emotions with an iterative algorithm.

Latent Variables in Conversation

A number of previous works have employed latent variables from deep generative models to model latent states in conversations. In particular, Zhao, Lee, and Eskenazi (2018), Zeng et al. (2019), Shi, Zhao, and Yu (2019) and Bao et al. (2020) applied Variational-Autoencoders (VAEs) (Kingma and Welling 2014) to learn latent states in conversations. Zhao, Lee, and Eskenazi (2018) and Bao et al. (2020) model the latent states for response generation. Our work is directly related to Zeng et al. (2019) that represents the latent discourse role of each utterance with a latent variable. We enhance the latent discourse role modelling with two context dependencies which shows significant benefit in our ERC experiment. Shi, Zhao, and Yu (2019) model the sequential nature of latent states in a conversation. Our work is in line, but we employ a recurrent neural network to connect all the latent states in a recurrent manner instead of using a fully connected layer between two latent states.

Methodology

In a conversation, the discourse role played by each utterance provides a salient clue to identify the emotion expressed. To overcome the lack of discourse role annotation in public ERC datasets, we propose to represent the discourse role with the latent variable from a VAE for each utterance in an unsupervised manner.

The overall architecture of our model is shown in Figure 2. Our proposed model consists of two key components - a hierarchical conversation encoder and a latent discourse role encoder. The hierarchical conversation encoder generates an utterance representation in a hierarchical manner, it first encodes each utterance independently, then incorporates the conversation context with a recurrent neural network. The latent discourse role encoder is a variant of VAE that generates the latent discourse role for each utterance.

Problem Definition

In each conversation, there is a sequence of K utterances $u_1, u_2, ..., u_K$. Each utterance u_k consists of a sequence of N_k tokens $w_{k,1}, w_{k,2}, ..., w_{k,N_k}$. The objective of ERC is to predict the emotion label $y_k \in E$ of each utterance u_k, where E is the set of emotion labels.

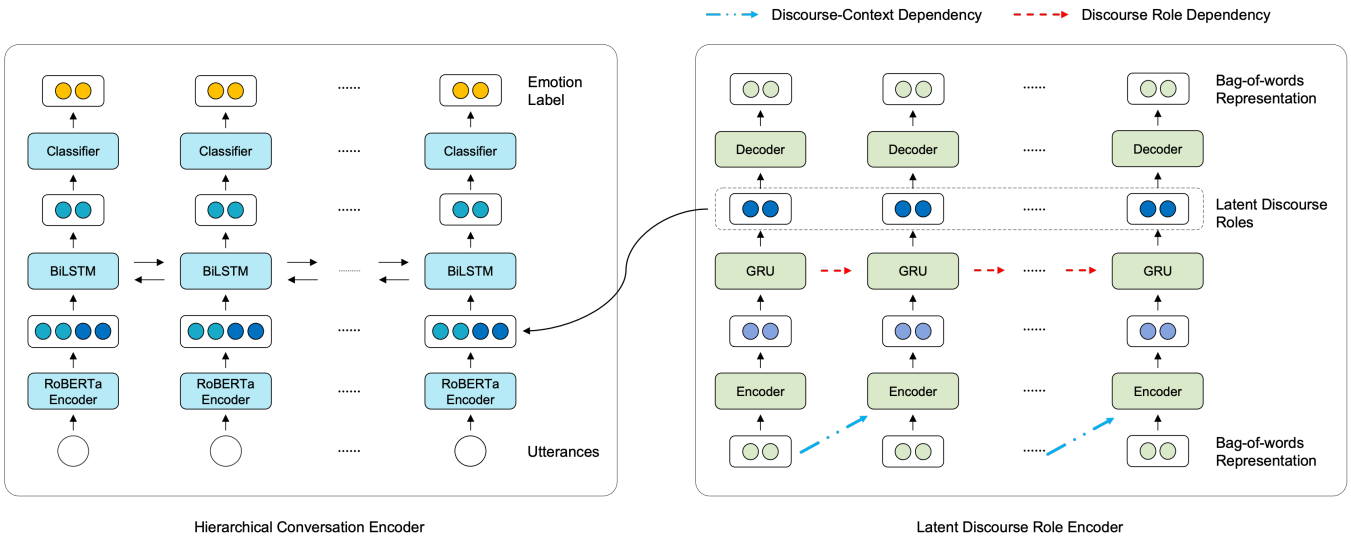

Figure 2: Illustration of our proposed model, which has two key components - a hierarchical conversation encoder and a latent discourse role encoder.

Hierarchical Conversation Encoder

To model the relationship between utterances in a conversation, we use a hierarchical structure. We first encode each utterance independently, then feed the sequence of encoded utterances into a recurrent neural network to produce context-aware utterance representations.

Utterance-level Encoding We employ a pre-trained language model to encode each utterance independently, as the encoded representation by pre-trained language models had been shown to enhance performance for multiple NLP tasks. We leverage RoBERTa (Liu et al. 2019) as our utterance encoder, which is based on the popular BERT (Devlin et al. 2019) with an enhanced training regime. RoBERTa has been shown to have better performance than BERT. Each utterance is lowercased and tokenized by the RoBERTa tokenizer, and at the start of the sequence of tokens, a special *[CLS]* token is inserted The tokenized sequence $[[CLS], w_{k,1}, w_{k,2}, ..., w_{k,N_k}]$ is fed into a transformer encoder (Vaswani et al. 2017), which is initialized with the RoBERTa pretrained weights. We then take the encoded *[CLS]* vector from the transformer encoder output as the utterance representation, u_k.

$$u_k = Transformer([CLS], w_{k,1}, w_{k,2}, ..., w_{k,N_k}) \quad (1)$$

As the latent discourse role can provide salient information to determine the emotion conveyed by the utterance, we concatenate the utterance representation u_k with the latent discourse role representation d_k from the latent discourse role encoder to form a discourse-aware utterance representation $\hat{u}_k$.

$$\hat{u}_k = u_k \oplus d_k \quad (2)$$

We describe how we generate the latent discourse role representation d_k in the next section.

Conversation-level Encoding Subsequently, we use the sequence of discourse-role aware utterance representations $\hat{u}_1, \hat{u}_2, ..., \hat{u}_K$ in a conversation as the input of the conversation-level bidirectional LSTM. Given the k-th discourse-aware utterance representation $\hat{u}_k$, we update h_k as follows:

$$h_k = BiLSTM(\hat{u}_k, h_{k-1}) \quad (3)$$

where $h_k \in \mathbb{R}^H$ is the hidden state of the LSTM for the discourse-aware utterance representation $\hat{u}_k$. The bidirectional LSTM models the sequential nature of a conversation so that each utterance is updated with the information from their predecessors and successors. As such, the output of the conversation level layer is a sequence of context-aware discourse-role aware utterance representations, $h_1, h_2, ..., h_K$.

Emotion Classification Finally, to predict the emotion label y_k for each utterance, a linear layer is used.

$$y_k = \mathbf{W^e} h_k + \mathbf{b^e} \quad (4)$$

where $\mathbf{W^e} \in \mathbb{R}^{E \times H}$ and $\mathbf{b^e} \in \mathbb{R}^E$ are trainable parameters, and E is the number of emotion classes.

We compute the classification loss using cross-entropy loss:

$$L_{cls} = -\frac{1}{\sum_{i=1}^{C} K_i} \sum_{i=1}^{C} \sum_{j=1}^{K_i} log P_{i,j}[y_{i,j}] \quad (5)$$

where C is the number of conversations, K_i is the number of utterances in conversation i, $P_{i,j}$ is the probability distribution of emotion labels for utterance j of conversation i and $y_{i,j}$ is the true emotion label.

Latent Discourse Role Encoder

In this section, we describe how our latent discourse role encoder generates context-aware latent discourse roles for each utterance in a conversation. We assume that there are D latent discourse roles at the corpus level, and each latent discourse role is captured by a multinomial word distribution over the vocabulary size V similar to (Zeng et al. 2019). Our latent discourse role encoder employs a variant of VAE (Kingma and Welling 2014), comprising two elements - an encoder and a decoder. The VAE generates a D-dimensional latent variable for each utterance that represents their latent discourse role distribution. Specifically, the encoder first encodes the input with a non-linear function and parameterises a prior distribution from which the latent variable $\hat{z}$ is sampled. Next, we generate the latent discourse role distribution via a softmax construction conditioned on $\hat{z}$. The decoder then learns to reconstruct the input from the latent discourse role distribution and the multinomial word distribution.

Unlike the input for the hierarchical conversation encoder, we process each utterance as a bag-of-words vector, $u_k^{bow} \in \mathbb{R}^V$, before feeding the vector into the latent discourse role encoder.

Discourse-Context Dependency One key characteristic of the discourse role is its dependency on the previous conversation context. For example, an utterance can only play the "answer" role if there is a "question" being asked earlier in the conversation. In order to generate context-aware latent discourse roles, we combine the preceding utterance with the current utterance by summing the two bag-of-words vectors to incorporate the conversational context.

$$\bar{u}_k^{bow} = u_{k-1}^{bow} + u_k^{bow} \tag{6}$$

While the combined pair of utterances is used as the input to the encoder, we use the current utterance u_k^{bow} as the reconstruction target so that the model will learn the latent discourse role of the utterance u_k.

Encoder The combined bag-of-words vector, $\bar{u}_k^{bow}$ is fed into two feed-forward encoders f^{μ} and f^{σ^2} to generate mean $\mu_k \in \mathbb{R}^D$ and variance $\sigma_k^2 \in \mathbb{R}^D$ to parameterize $q(z_k|\bar{u}_k^{bow}) = \mathcal{N}(\mu_k, \sigma_k^2)$, where $z_k \in \mathbb{R}^D$ is a latent variable.

$$\mu_k = f^{\mu}(\bar{u}_k^{bow}) \tag{7}$$

$$\sigma_k^2 = f^{\sigma^2}(\bar{u}_k^{bow}) \tag{8}$$

Then, we sample $\hat{z}_k$ from $q(z_k|\bar{u}_k^{bow})$ using a reparametrization trick as described in (Kingma and Welling 2014)

$$\hat{z}_k = \mu_k + \epsilon \cdot \sigma_k^2 \tag{9}$$

where ϵ is sampled from $\mathcal{N}(0, I^2)$ and I is the identity matrix.

Discourse Roles Dependency Notably, the latent variable for each utterance is sampled independently. Unlike previous works that apply a feed-forward layer on the sampled latent variable $\hat{z}_k$ before decoding, we model the sequential dependency between the sampled latent variables with a recurrent neural network. The recurrent neural network model the sequential information flow across all latent variables. Specifically, we feed the sequence of independently sampled latent variables $\hat{z}_0, \hat{z}_1, ..., \hat{z}_K$ into a unidirectional GRU.

$$\overline{z}_k = GRU(\hat{z}_k, \overline{z}_{k-1}) \tag{10}$$

where $\overline{z}_k \in \mathbb{R}^D$ is the hidden state of the GRU for the sampled latent variable $\hat{z}_k$.

Finally, we obtain the context-aware discourse-aware latent discourse roles distribution $\theta_k \in \mathbb{R}^D$ from the hidden states of the GRU:

$$\theta_k = softmax(\overline{z}_k) \tag{11}$$

Decoder In the decoder, discourse role embeddings $E_D \in \mathbb{R}^{D \times M}$ and word embeddings $E_W \in \mathbb{R}^{V \times M}$ are randomly initialized, where D is the number of latent discourse roles, V is the vocabulary size and M is the embedding dimension. The two embeddings are used to construct the discourse role-words distribution $\beta \in \mathbb{R}^{D \times V}$ as follows:

$$\beta = softmax(\frac{E_W^{\top} \cdot E_D}{\sqrt{M}}) \tag{12}$$

The utterance u_k^{bow} is then reconstructed as $\hat{u}_k^{bow}$ using the context-aware latent discourse role distribution θ_k and the discourse role-words distribution β as follows:

$$\hat{u}_k^{bow} = \theta_k \beta \tag{13}$$

To enhance our utterance representation with the latent discourse role information, we obtain the latent discourse role representation $d_k \in \mathbb{R}^M$ from the context-aware latent discourse roles distribution θ_k and discourse role embeddings E_D as follows:

$$d_k = \theta_k E_D \tag{14}$$

The latent discourse role representation, which can be seen as the weighted discourse role representation, is then concatenated with the independently encoded utterance representation as shown in Equation 2.

The overall loss for the latent discourse role encoder is defined as:

$$\begin{aligned} L_{dis} = & KL[q(z_k|u_k^{bow})||p(z_k)] \\ & - \mathbb{E}_{q(z_k|u_k^{bow})}[logp(u_k^{bow}|z_k)] \end{aligned} \tag{15}$$

where the first term (Kullback-Leibler divergence) ensures that the approximated posteriors are close to the true prior distribution, and the second term ensures that the generated latent discourse roles can reconstruct the current utterance.

We train both hierarchical conversation encoder and latent discourse roles encoder together in an end-to-end manner with the loss function as defined:

$$L = L_{cls} + \lambda_{dis} L_{dis} \tag{16}$$

where λ_{dis} is the weight of loss for the latent discourse role encoder determined based on experiments.

Dataset	# dialogues / # utterances			# labels	Evaluation Metric
	train	val	test		
DailyDialog	11,118 / 87,832	1,000 / 7,912	1,000 / 7863	7*	Micro-F1
MELD	1,039 / 9,989	114 / 1,109	280 / 2,610	7	Weighted Avg. F1
IEMOCAP	108 / 5,236	12 / 574	31 / 1,623	6	Weighted Avg. F1

Table 1: Dataset description and Evaluation Metric. **Neutral* labels are excluded when calculating the Micro F1 score for the DailyDialog dataset.

Hyperparameters	DailyDialog	MELD	IEMOCAP
C	150	300	200
D	19	7	19
M	50	50	50
$\lambda_{discourse}$	1e-3	1e-1	1e-5
learning rate	5e-4	2e-5	5e-3

Table 2: Hyperparameter settings. C is the hidden size of the conversation-level bidirectional GRU, D is the number of latent discourse roles, M is the discourse role and word embedding dimension.

Experimental Settings

In this section, we present the experimental settings used to validate the effectiveness of our proposed methods. The experimental settings include the datasets, evaluation metrics and implementation details. We also briefly describe the models that we compare against.

Datasets and Evaluation

We evaluate our model on three publicly available datasets, differing in magnitudes of size. We present the summary of statistics for the datasets used in our experiments in Table 1.

- **DialyDialog** (Li et al. 2017) is a dyadic text-based dialog dataset based on daily written communications. Each utterance in every dialogue is annotated as one of the seven emotion classes: happiness, surprise, sadness, anger, disgust, fear or no emotion.
- **MELD** (Poria et al. 2019) is a multiparty multi-modal dialog dataset from the Friends TV series. We only used the text features. Each utterance in every dialogue is annotated as one of the seven emotion classes: anger, disgust, sadness, joy, surprise, fear or neutral.
- **IEMOCAP** (Busso et al. 2008) is dyadic multi-modal dialog dataset based on videos of two-way conversations. Like MELD, we only used the text features. Each utterance in every dialogue is annotated as one of the six emotion classes: happy, sad, neutral, angry, excited, and frustrated.

For evaluation, we follow the settings from (Ghosal et al. 2020). We use the Micro-F1 score excluding the neutral (no emotion) label for DailyDialog. The neutral label accounts for more than 80% of the labels. For MELD and IEMOCAP, we use the weighted-average F1 score for all labels.

Implementation Details

We preprocess the utterances by lower-casing and tokenizing using RoBERTa and Spacy[1] tokenizers for the input to the hierarchical conversation encoder and latent discourse role encoder, respectively. Utterances fed into the hierarchical conversation encoder are truncated at 128 tokens. All deep learning models are implemented using PyTorch (Paszke et al. 2019).

We use the publicly available RoBERTa$_{BASE}$ weights to initialize the transformer for utterance-level encoding. We optimize the model using the AdamW (Devlin et al. 2019) optimizer with a linear warm-up schedule for every dataset. We employed two different learning rates - one for the utterance-level transformer encoder and one for the rest of the model.

Each model is fine-tuned on the validation dataset and early stopped. We performed grid search for the following hyperparameters: hidden size of the conversation level BiGRU amongst $\{100, 200, 300\}$, number of latent discourse roles amongst $\{4 - 20\}$, dimension for the discourse role and word embedding amongst $\{50, 100, 150, 200\}$, weight of the latent discourse encoder loss amongst $\{1e^{-1}, 1e^{-2}, 1e^{-3}, 1e^{-4}, 1e^{-5}\}$ and learning rate for rest of the model amongst $\{2e^{-3}, 5e^{-3}, 2e^{-4}, 5e^{-4}, 2e^{-5}, 5e^{-5}\}$ for all datasets. We used the same learning rate $2e^{-5}$ for the utterance-level transformer encoder. We evaluate the best fine-tuned model on the test data. Table 2 contains the hyperparameters settings for each dataset which were selected using the validation set.

Compared Methods

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

- **CNN+cLSTM** (Poria et al. 2017) first encode utterance using a CNN and then feed into a conversation-level LSTM to model the sequential nature of a conversation.
- **ICON** (Hazarika et al. 2018a) employ one GRU for each speaker to model speaker-specific utterance representation and another conversation-level GRU.
- **DialogueRNN** (Majumder et al. 2019) employ three separate GRU networks for speaker, context and emotion states.
- **DialogueGCN** (Ghosal et al. 2019) model self-dependency and inter-speaker dependency by using two-layer graph neural networks.

[1] https://spacy.io/

Model	DailyDialog	MELD	IEMOCAP
CNN+cLSTM	50.24	56.87	54.95
ICON	-	-	58.54
DialogueRNN	-	57.03	62.75
DialogueGCN	-	58.10	64.18
KET	53.37	58.18	59.56
KAITML	54.71	58.97	61.43
IEIN	-	60.72	64.37
RGAT	54.31	60.91	65.22
HiTrans	-	61.94	64.50
DialogXL	54.93	62.41	65.94
COSMIC	58.48	65.21	65.28
Ours	**60.95**	**65.34**	**68.23**

Table 3: Experimental results on emotion recognition in conversation on DailyDialog, MELD and IEMOCAP datasets. Evaluation measure is micro-average F1 for DailyDialog and weighted-average F1 for MELD and IEMOCAP.

- **KET** (Zhong, Wang, and Miao 2019) uses hierarchical self-attention and cross-attention to capture intra-utterance and inter-utterance correlations and a context-aware graph attention mechanism.
- **KAITML** (Zhang et al. 2020) augments utterance representation with commonsense knowledge, and apply Incremental Transformer to capture intra-utterance and inter-utterance correlations.
- **IEIN** (Lu et al. 2020) explicitly model the emotion interaction between utterances using an iterative improvement algorithm.
- **RGAT** (Ishiwatari et al. 2020) introduce relation positional encoding to provided graph neural networks with sequential information reflecting relation types.
- **HiTrans** (Li et al. 2020) uses two hierarchical transformers and train on an auxiliary task to predict if a pair of utterances belong to the same speaker.
- **DialogXL** (Shen et al. 2021) extends pretrained language model XLNet (Yang et al. 2019) with a novel memory component to store historical context and dialog-aware self-attention .
- **COSMIC** (Ghosal et al. 2020) introduce commonsense knowledge and uses five bidirectional GRUs to model 5 different states in a conversation. The five states are context, internal, external, intent, and emotion.

Results and Analysis

Comparison with Baselines and State-of-the-Art Methods

Table 3 reports a performance comparison of all published models on the DailyDialog, MELD and IEMCOAP datasets. While earlier models like DialogueRNN, DialogueGCN and KET encode utterances with GloVe embeddings, we observed that using pretrained language models to generate utterance representations resulted in a significant improvement in performance. This is evidenced by recent publications using RGAT, HiTrans, DialogXL and COSMIC which use pretrained language models like BERT, RoBERTa and XLNet. Following this trend, our proposed model encoding utterances with a pre-trained RoBERTa model enhanced with latent discourse role information achieves state-of-the-art results across all three public ERC datasets.

On the DailyDialog dataset, our model achieves 60.95% in micro F1, a 2.47% improvement compared against the best-published model. Similarly, our model significantly outperforms the state-of-the-art by 2.29% weighted F1 on IEMOCAP. Finally, on MELD, our model improves on the state-of-the-art and achieves very competitive performance. The performance improvement gained on MELD is not as significant as the other two datasets. We hypothesize that this may be due to the difference in information flow in a multi-party conversation dataset such as MELD versus dyadic dialogue in DailyDialog and IEMOCAP. In the future, we plan to identify better way to model the structure of latent discourse roles in a multi-party conversation.

It is to be noted that the existing published models are orthogonal to our work. These models focus on modelling the conversation context, speaker identity and incorporating common-sense knowledge. However, none of them considers the discourse role of each utterance. Our proposed model is able to improve emotion recognition performance by modelling only the sequential dependency between utterances and latent discourse role information, without modelling any of the features used by other models. Thus, it may potentially reinforce the effectiveness of discourse role information for detecting emotions in conversation over existing models as well.

Method	DailyDialog	MELD	IEMOCAP
Our Method	60.95	65.34	68.23
w/o Discourse-Context Dependency	59.89	63.57	65.96
w/o Discourse Role Dependency	59.69	63.20	65.08
w/o Both Dependencies	58.44	63.03	64.11

Table 4: Ablation results w.r.t the latent discourse role dependencies on DailyDialog, MELD and IEMOCAP datasets.

Importance of Discourse Roles Dependency

We study the importance of i) incorporating the dependency between latent discourse role and conversation context and ii) the transition dependency between adjacent latent discourse roles. The results of the ablation study are presented in Table 4. To study the effect of conversation context on the generated discourse roles, we use only the target utterance as input to the latent discourse role encoder instead of the combined pair of utterances. Removing the contextual information in the latent discourse role results in a drop in performance. For the discourse role transition dependency ablation, we discard the GRU used to model the sequential nature of the latent discourse roles. It is observed that in all three datasets, the performance drop is higher than in the discourse-context ablation.

The ablation study confirms that both the conversation context and discourse role transitions are intrinsic to enhancing the utterance representation in our proposed model. The

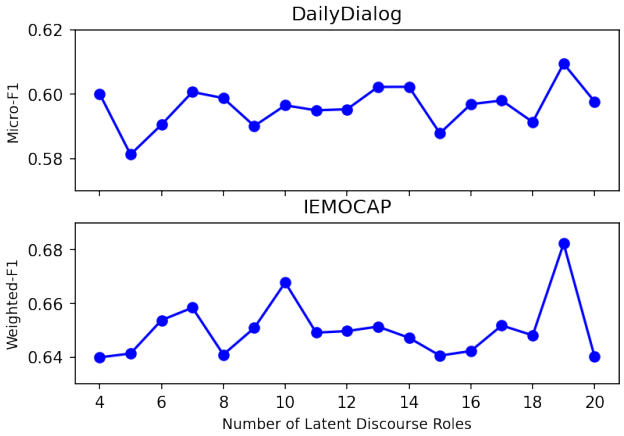

Figure 3: Performance on DailyDialog and IEMOCAP with different number of latent discourse roles.

Hierarchical Conversation Encoder	Latent Discourse Role Encoder	F1
✓	✓	68.23
✓	✗	66.06
✗	✓	44.87

Table 5: Ablation study on IEMOCAP. ✗ means the corresponding encoder is removed from the model.

appreciable performance improvement by incorporating the dependency between latent discourse role and conversation context suggests that the preceding utterance provides valuable information about the context and flow of the conversation. This was expected to be the case because a conversation is interactive, and we react and reply to what was being said earlier in a conversation.

Number of Latent Discourse Roles

To study the effect of varying the number of latent discourse roles on model performance, we report model performance on DailyDialog and IEMOCAP over a range of number of latent discourse roles, shown in Fig 3. The number of latent discourse roles determines how the intent and information flow that each utterance entails are differentiated.

From the visualization, it is clear that the model performance does not increase monotonically with the number of latent discourse roles. The observation aligns with our expectations because it is challenging to explicitly define the underlying discourse roles in a conversation for any corpus. With a smaller set of discourse roles, the utterances are less differentiated, leading to the signals provided by the discourse roles to be weaker. While on the opposite side of the spectrum, having too many discourse roles would result in a noisier signal, making it more difficult to determine the emotion conveyed by the utterance.

Utterances	Gold	Pred
Speaker 1: No, it doesn't pay the bills, but it would pay something. And it would help you get somewhere else	Neutral	Neutral
Speaker 2: I still can't live on in six seven and five. It's not possible in Los Angeles. Housing is too expensive	Anger	Frustrated

Table 6: An example of misclassification between similar emotion labels "Anger" and "Frustrated" in IEMOCAP.

Ablation study

We perform an ablation study on the two encoders of our proposed model, namely Hierarchical Conversation Encoder and Latent Discourse Role Encoder, on the IEMOCAP dataset in Table 5. We feed the latent discourse role representation to the classification layer when we remove the Hierarchical Conversation Encoder.

The Hierarchical Conversation Encoder, which encodes utterance with a pretrained language model independently and performs contextual modelling over the conversation, produces rich representations for emotion recognition in conversations. The Latent Discourse Role Encoder represents each utterance within a fixed set of latent discourse roles, which limits its expressivity for emotion recognition. However, enhancing the representations from the Hierarchical Conversation Encoder with the latent discourse role representations improves the model performance, validating the importance of latent discourse role in emotion recognition.

Error Analysis

Our model performs exceptionally in all three public ERC datasets, showing a significant performance improvement in two datasets and achieving state-of-the-art performance on the third one. Upon analyzing the misclassifications between emotion labels, we observe that errors were most common between emotion pairs that are closely related. In certain cases, it is observed that the model has a tendency to misclassify 'happiness' - 'excited', 'anger' - 'frustrated' in IEMOCAP and 'happiness' - 'surprise' in DailyDialog. Table 6 present a case of misclassification between 'anger' and 'frustrated' in IEMOCAP dataset. It is difficult to ascertain whether Speaker 2 is angry or frustrated when she/he is complaining about expensive housing. This phenomenon of misclassification between similar emotions has been observed in previous studies (Zhong, Wang, and Miao 2019; Ghosal et al. 2019). Meanwhile, we also observe that the model performance decreases when the conversation length increases. Longer conversations might involve a change in topic and not all utterances and their discourse roles are relevant. Future investigations can look into segmentation of long conversation into a coherent topic and reduce noise from unrelated utterances.

Conclusion

In this paper, we study the importance of discourse role in the emotion recognition in conversation task. We propose a variant of VAE to model the context-aware latent discourse role with latent variables in an unsupervised manner without discourse role annotations. Our VAE models the relationship between conversation context and discourse role and the sequential nature of discourse roles. We demonstrated the empirical effectiveness of our method on datasets of different magnitudes in size. The results show that enhancing utterance representation with discourse role beats the best-reported performances on three public datasets on ERC.

Acknowledgments

This research is partially supported by the programme DesCartes funded by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

References

Bao, S.; He, H.; Wang, F.; Wu, H.; and Wang, H. 2020. PLATO: Pre-trained Dialogue Generation Model with Discrete Latent Variable. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 85–96. Online: Association for Computational Linguistics.

Busso, C.; Bulut, M.; Lee, C.; Kazemzadeh, A.; Mower, E.; Kim, S.; Chang, J. N.; Lee, S.; and Narayanan, S. S. 2008. IEMOCAP: interactive emotional dyadic motion capture database. *Lang. Resour. Evaluation*, 42(4): 335–359.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 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)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Ghosal, D.; Majumder, N.; Gelbukh, A.; Mihalcea, R.; and Poria, S. 2020. COSMIC: COmmonSense knowledge for eMotion Identification in Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2470–2481. Online: Association for Computational Linguistics.

Ghosal, D.; Majumder, N.; Poria, S.; Chhaya, N.; and Gelbukh, A. 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)*, 154–164. Hong Kong, China: Association for Computational Linguistics.

Hazarika, D.; Poria, S.; Mihalcea, R.; Cambria, E.; and Zimmermann, R. 2018a. ICON: Interactive Conversational Memory Network for Multimodal Emotion Detection. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 2594–2604. Brussels, Belgium: Association for Computational Linguistics.

Hazarika, D.; Poria, S.; Zadeh, A.; Cambria, E.; Morency, L.-P.; and Zimmermann, R. 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, Volume 1 (Long Papers)*, 2122–2132. New Orleans, Louisiana: Association for Computational Linguistics.

Ishiwatari, T.; Yasuda, Y.; Miyazaki, T.; and Goto, J. 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)*, 7360–7370. Online: Association for Computational Linguistics.

Jiao, W.; Lyu, M.; and King, I. 2020. Exploiting Unsupervised Data for Emotion Recognition in Conversations. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, 4839–4846. Online: Association for Computational Linguistics.

Kingma, D. P.; and Welling, M. 2014. Auto-Encoding Variational Bayes. In Bengio, Y.; and LeCun, Y., eds., *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Li, J.; Ji, D.; Li, F.; Zhang, M.; and Liu, Y. 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*, 4190–4200. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Li, Y.; Su, H.; Shen, X.; Li, W.; Cao, Z.; and Niu, S. 2017. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 986–995. Taipei, Taiwan: Asian Federation of Natural Language Processing.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *CoRR*, abs/1907.11692.

Lu, X.; Zhao, Y.; Wu, Y.; Tian, Y.; Chen, H.; and Qin, B. 2020. An Iterative Emotion Interaction Network for Emotion Recognition in Conversations. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4078–4088. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Majumder, N.; Poria, S.; Hazarika, D.; Mihalcea, R.; Gelbukh, A. F.; and Cambria, E. 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*, 6818–6825. AAAI Press.

Paszke, A.; Gross, S.; Massa, F.; Lerer, A.; Bradbury, J.; Chanan, G.; Killeen, T.; Lin, Z.; Gimelshein, N.; Antiga, L.; Desmaison, A.; Köpf, A.; Yang, E.; DeVito, Z.; Raison, M.; Tejani, A.; Chilamkurthy, S.; Steiner, B.; Fang, L.; Bai, J.; and Chintala, S. 2019. PyTorch: An Imperative Style, High-Performance Deep Learning Library. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 8024–8035.

Poria, S.; Cambria, E.; Hazarika, D.; Majumder, N.; Zadeh, A.; and Morency, L.-P. 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)*, 873–883. Vancouver, Canada: Association for Computational Linguistics.

Poria, S.; Hazarika, D.; Majumder, N.; Naik, G.; Cambria, E.; and Mihalcea, R. 2019. MELD: A Multimodal Multi-Party Dataset for Emotion Recognition in Conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 527–536. Florence, Italy: Association for Computational Linguistics.

Shen, W.; Chen, J.; Quan, X.; and Xie, Z. 2021. DialogXL: All-in-One XLNet for Multi-Party Conversation Emotion Recognition. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, 13789–13797. AAAI Press.

Shi, W.; Zhao, T.; and Yu, Z. 2019. Unsupervised Dialog Structure Learning. 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)*, 1797–1807. Minneapolis, Minnesota: Association for Computational Linguistics.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, L.; and Polosukhin, I. 2017. Attention is All you Need. In Guyon, I.; von Luxburg, U.; Bengio, S.; Wallach, H. M.; Fergus, R.; Vishwanathan, S. V. N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, 5998–6008.

Yang, Z.; Dai, Z.; Yang, Y.; Carbonell, J. G.; Salakhutdinov, R.; and Le, Q. V. 2019. XLNet: Generalized Autoregressive Pretraining for Language Understanding. In Wallach, H. M.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E. B.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, 5754–5764.

Zeng, J.; Li, J.; He, Y.; Gao, C.; Lyu, M. R.; and King, I. 2019. What You Say and How You Say it: Joint Modeling of Topics and Discourse in Microblog Conversations. *Transactions of the Association for Computational Linguistics*, 7: 267–281.

Zhang, D.; Chen, X.; Xu, S.; and Xu, B. 2020. Knowledge Aware Emotion Recognition in Textual Conversations via Multi-Task Incremental Transformer. In *Proceedings of the 28th International Conference on Computational Linguistics*, 4429–4440. Barcelona, Spain (Online): International Committee on Computational Linguistics.

Zhang, D.; Wu, L.; Sun, C.; Li, S.; Zhu, Q.; and Zhou, G. 2019. Modeling both Context- and Speaker-Sensitive Dependence for Emotion Detection in Multi-speaker Conversations. In Kraus, S., ed., *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, 5415–5421. ijcai.org.

Zhao, T.; Lee, K.; and Eskenazi, M. 2018. Unsupervised Discrete Sentence Representation Learning for Interpretable Neural Dialog Generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1098–1107. Melbourne, Australia: Association for Computational Linguistics.

Zhong, P.; Wang, D.; and Miao, C. 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)*, 165–176. Hong Kong, China: Association for Computational Linguistics.